A Posteriori Error Analysis of the Discontinuous Galerkin Method for Linear Hyperbolic Systems of Conservation Laws

Thomas Weinhart

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Mathematics

Slimane Adjerid, Chair
Christopher Beattie
Tao Lin
Robert Rogers

March 19, 2009
Blacksburg, Virginia

Keywords: Discontinuous Galerkin method, hyperbolic systems of conservation laws, *a posteriori* error estimation, superconvergence, adaptivity

A Posteriori Error Analysis of the Discontinuous Galerkin Method for Linear Hyperbolic Systems of Conservation Laws

Thomas Weinhart

ABSTRACT

In this dissertation we present an analysis for the discontinuous Galerkin (DG) discretization error of multi-dimensional first-order linear symmetric and symmetrizable hyperbolic systems of conservation laws. We explicitly write the leading term of the local DG error, which is spanned by Legendre polynomials of degree p and $p+1$ when p^{th}-degree polynomial spaces are used for the solution. For special hyperbolic systems, where the coefficient matrices are nonsingular, we show that the leading term of the error is spanned by $(p+1)^{th}$-degree Radau polynomials. We apply these asymptotic results to observe that projections of the error are pointwise $\mathcal{O}(h^{p+2})$-superconvergent in some cases and establish superconvergence results for some integrals of the error. We develop an efficient implicit residual-based *a posteriori* error estimation scheme by solving local finite element problems to compute estimates of the leading term of the discretization error. For smooth solutions we obtain error estimates that converge to the true error under mesh refinement. We first show these results for linear symmetric systems that satisfy certain assumptions, then for general linear symmetric systems. We further generalize these results to linear symmetrizable systems by considering an equivalent symmetric formulation, which requires us to make small modifications in the error estimation procedure. We also investigate the behavior of the discretization error when the Lax-Friedrichs numerical flux is used, and we construct asymptotically exact *a posteriori* error estimates. While no superconvergence results can be obtained for this flux, the error estimation results can be recovered in most cases. These error estimates are used to drive h- and p-adaptive algorithms and assess the numerical accuracy of the solution. We present computational results for different fluxes and several linear and nonlinear hyperbolic systems in one, two and three dimensions to validate our theory. Examples include the wave equation, Maxwell's equations, and the acoustic equation.

This research was partially supported by the National Science Foundation (Grant Number DMS 0511806).

Acknowledgments

I want to express my deep-felt thanks to my advisor, Professor Slimane Adjerid, for his warm encouragement and thoughtful guidance. I also thank the other members of my dissertation committee, Professor Robert Rogers, Professor Tao Lin and Professor Christopher Beattie, for their helpful suggestions and the many things they taught me.

I would like to express my gratitude to my parents, Siegfried and Irmgard Weinhart, who encouraged me to study here at Virginia Tech and supported me during all this time. I am grateful for the support I received from my beloved Raquel and for all my friends, who were there when I needed them and gave me advice while I was writing this dissertation.

Finally, I thank the National Science Foundation for their generous support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation presents a new finite element approach to the numerical solution of hyperbolic systems of conservation laws that allows for the efficient computation of local *a posteriori* error estimates.

Systems of first-order partial differential equations in divergence form, also called systems of conservation laws, arise in many areas of continuum physics when fundamental balance laws are formulated (such as the conservation of mass, momentum, or energy) and if other small-scale, dissipative mechanisms can be neglected (such as viscosity, capillarity, heat conduction, Hall effect). Problems of practical interest arise in many fields such as gas and fluid dynamics, acoustics, electromagnetism and aerodynamic or geophysical flow. Solutions to conservation laws exhibit singularities (shock waves), which can appear in finite time even if the initial conditions are smooth. Thus, the system must be interpreted in the sense of distributions and does not have unique solutions unless some entropy condition is imposed, that models a physical process in the limit as dissipation tends to zero, see *e.g.* [32, 50].

In this dissertation, a particular class of systems of conservation laws is considered, namely symmetric or symmetrizable systems. Symmetry is often a consequence of conservation principles in physics, and we will consider important applications that fall into this class, such as Euler's equation of gas dynamics, Maxwell's equations of electromagnetism, and the acoustic wave equation. Symmetric or symmetrizable systems are hyperbolic, which plays a role in the well-definedness of a system.

Thus, when constructing numerical methods to solve systems of conservation laws, the approximate solution needs to capture the physically relevant discontinuities. Also, this method must remain sufficiently accurate near discontinuities in order to capture the possibly complex structure of the exact solution. These difficulties were successfully resolved by the development of finite difference (FD) and finite volume (FV) schemes for such systems. However, numerical methods should also be able to approximate solutions in smooth regions with a high order of accuracy. Since FD and FV methods both use large stencils for

high-order approximations, developing stable algorithms for complex geometries in multiple dimensions is difficult, while finite element (FE) methods can easily handle these difficulties. Also, for conservation laws, in which information flows in specific directions, the continuity requirement can cause stability problems, and spurious oscillations occur near discontinuities. While these difficulties can be resolved, the problem is more easily addressed in FD and FV methods by the use of upwinding and slope limiters.

Discontinuous Galerkin (DG) methods are a higher-order generalization of FV methods and are advantageous compared to FD and FV schemes due to their FE nature: *i*) They can handle complex geometries, even for high-order approximations and *ii*) the treatment of the boundary conditions is much simpler than in stencil-based methods. Additionally, *iii*) since DG methods do not require continuity on element boundaries, the test functions are defined locally. Thus, the mass matrix is block diagonal and can easily be inverted once and for all, resulting in an explicit semi-discrete form. The explicitness and the simple treatment of the boundary conditions makes DG methods easily parallelizable. Also, *iv*) DG methods can easily handle adaptive strategies, since refinement and coarsening is restricted neither by the continuity requirement of conforming finite element methods, nor by the geometric inflexibility of stencil-based methods. Adaptivity is particularly important in systems of conservation laws given the complexity of solutions near discontinuities.

The numerical approximation of a continuum model of any physical process always involves discretization error caused by the discretization of the continuum model into an algorithm that can be fed into computers. *A posteriori* error estimation is used to assess the quality of numerical solutions and guide adaptive algorithms, which aim to yield a solution, whose error in some norm is below a given tolerance, in an effective manner. After estimating the error, elements having high errors are enriched by h-refinement and/or p-refinement, while elements with small errors are $h-$ and/or $p-$coarsened.

1.1 Review of Past Work and Historical Perspective

1.1.1 Discontinuous Galerkin Methods

The discontinuous Galerkin finite element method was introduced in 1973 by Reed and Hill [57] to solve the neutron transport equation,

$$\operatorname{div}(\mathbf{a}u) + \sigma u = f, \tag{1.1.1}$$

where σ is a real number and $\mathbf{a}$ is constant. Because of the linear nature of this equation, the approximate solution can be computed element by element, when the elements are ordered according to the characteristic direction. LeSaint and Raviart [53] first analyzed the method by reducing it to an ordinary differential equation (ODE) and showing a rate of convergence of $\mathcal{O}(h^p)$ for general triangulations and of $\mathcal{O}(h^{p+1})$ for Cartesian grids, if the exact solution

is smooth. In 1986, Johnson and Pitkaränta [46] proved a rate of convergence of $\mathcal{O}(h)^{p+1/2}$ for general triangulations and Peterson [56] confirmed these rates to be optimal. In 1988, Richter [58] proved the optimal rate of $\mathcal{O}(h^{p+1})$ for some structured two-dimensional non-Cartesian grids. For solutions with discontinuities, Lin and Zhou [54] proved convergence of the method. Further studies of initial-value problems for ordinary differential equations include [4, 17, 53, 62, 63].

Later, the DG method was successfully applied to nonlinear hyperbolic conservation laws

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{i=1}^{d} \frac{\partial}{\partial x_i} \mathbf{f}_i(\mathbf{u}) = \mathbf{0}, \tag{1.1.2}$$

equipped with suitable initial and boundary conditions. For the one-dimensional scalar case, Chavent and Salzano [22] constructed an explicit DG Method. They discretized in space using the DG method with piecewise linear elements, yielding an explicit semi-discrete scheme. Then they solved in time using a simple Euler forward method. To improve the stability of the scheme, Chavent and Cockburn [21] modified the scheme by introducing a *slope limiter*, introduced by van Leer [70]. The Runge-Kutta Discontinuous Galerkin (RKDG) method was introduced by Cockburn and Shu [29], where they used a piecewise linear DG method for the space discretization, a special explicit TVD second-order Runge-Kutta time discretization, and a modified slope limiter to maintain formal accuracy of the scheme extrema. In [28], Cockburn and Shu generalized the approach to develop high-order accurate RKDG methods for scalar conservation laws. The RKDG method was extended to one-dimensional systems by Cockburn, Lin, and Shu in [27] and then to multi-dimensional scalar equations by Cockburn, Hou and Shu in [25]. The extension to multi-dimensional systems was done by Cockburn and Shu in [29]. They also developed the Local Discontinuous Galerkin method for convection-diffusion problems [30]. Consult [26] and the references therein for more information on DG methods.

1.1.2 *A posteriori* Error Estimation

A posteriori error estimation is used to verify the error of a numerical solution u_h with respect to the real solution u, and can be used for mesh adaptivity. A posteriori error estimates make use of the numerical solution u_h to a particular problem to obtain an estimate, in contrast to *a priori* error estimates. Thus, they are often much more accurate than *a priori* estimates, which provide only a rough error bound by exploiting properties of the governing equations and assumptions on properties of the real solution u.

According to Adjerid et al. [5], an ideal *a posteriori* error estimation technique should:

i) be *asymptotically correct* in the sense that the error estimate in a particular norm approach zero under enrichment at the same rate as the actual error;

ii) be *computationally simple* by requiring a small fraction of the solution cost;

iii) be *robust* by furnishing accurate estimates for a wide range of meshes and method orders;

iv) provide relatively tight *upper and lower bounds* of the true error in a particular norm; and

v) supply local error indicators that provide global error estimates in *several norms*.

The development of *a posteriori* error estimates mainly focuses on approximations of the error that can be obtained locally. Such estimates can be used to guide adaptive meshing procedures, where elements with high errors are enriched, while elements with small errors are coarsened. Adaptivity processes can also be based on *a priori* interpolation estimates, which depend on estimates of the unknown function u and can therefore provide only crude but effective indications of features of error, such as discontinuities, see *e.g.* Demkovicz et al. [34] and Peraire et al. [55]. However, when more complex features of the solution are present, such as boundary layers or shock-boundary layer interaction, *a priori* error estimates can be inaccurate.

A posteriori error estimation techniques were first developed for finite element methods for elliptic boundary value problems in 1978 by Babuška and Rheinboldt [10]. They developed a technique that delivered approximations η_K of the error in energy norm on each element K. In the following years, Babuška and Rheinboldt obtained a number of explicit error estimation techniques, see *e.g.* [11, 12]. The *element residual method* was developed by Demkovicz et al [35, 36] in 1984, who applied it to a variety of problems in mechanics and physics. Bank and Weiser [14, 15] applied it to scalar elliptic problems and provided a mathematical analysis of the method. Zienkiewicz and Zhu [72] developed a simple *recovery-based method*, where gradients of solutions are smoothed and then compared with the gradient of the original solution. Later, they developed the *super-convergent patch recovery* method. Other methods include using *equilibrated boundary data* [51] and *extrapolation methods* [67].

While *a posteriori* error estimation has attained a certain level of maturity for diffusive problems [9, 71], developing accurate and robust *a posteriori* error estimates for hyperbolic problems remains a challenge. An *a posteriori* error analysis for convection and convection-dominated convection-diffusion problems was presented by Johnson and his collaborators [44, 45, 47]. More recently, Süli and his collaborators [42, 65, 66] investigated local and transmitted errors as well as goal-oriented estimates for several numerical methods applied to hyperbolic problems.

Several *a posteriori* DG error estimates are known for hyperbolic [23, 24, 40, 52] and diffusive [43, 59] problems. The first asymptotically correct *a posteriori* error estimates for hyperbolic problems were developed by Adjerid *et al.* [4] who constructed the first superconvergence-based *a posteriori* DG error estimates for one-dimensional linear and nonlinear hyperbolic

problems. Later, Adjerid and Massey [7, 8] showed how to construct accurate error estimates for multi-dimensional scalar problems on rectangular meshes. They showed that the leading term of error is spanned by two $(p+1)$-degree Radau polynomials in the x and y directions, respectively. Krivodonova and Flaherty [49] showed that the leading term of the local discretization error on triangles having one *outflow* edge is spanned by a suboptimal set of orthogonal polynomials of degree p and $p+1$. They computed DG error estimates by solving local problems involving numerical fluxes, thus requiring information from neighboring *inflow* elements. Adjerid and Baccouch [2, 3] investigated DG methods on structured and unstructured triangular meshes with several finite element spaces to discover new superconvergence properties and compute accurate error estimates. LDG methods for diffusion problems were investigated by Adjerid and Klauser [6] who constructed efficient and accurate *a posteriori* error estimates.

Superconvergence properties for DG methods have been studied in [4, 33, 48, 53] for first-order ordinary differential equations, in [2, 3, 4, 7, 8] for hyperbolic problems and [6, 19, 20] for diffusion and convection-diffusion problems.

1.2 Basic Linear Algebra and Notations

In this section we will include basic linear algebra results that will be needed in this dissertation.

1.2.1 Matrix Theory

We denote vectors in $\mathbb{R}^k$ as

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_k \end{pmatrix}, \tag{1.2.1}$$

with the canonical Euclidean norm

$$\|\mathbf{v}\| = \sqrt{\sum_{i=1}^{k} |v_i|^2}. \tag{1.2.2}$$

The transpose of $\mathbf{v}$ is denoted by $\mathbf{v}^t$ and the orthogonal complement of any set S of vectors in $\mathbb{R}^k$ is defined as

$$S^\perp = \{\mathbf{w} \in \mathbb{R}^k : \ \mathbf{w}^t\mathbf{v} = 0, \ \forall\ \mathbf{v} \in S\}, \tag{1.2.3}$$

while the direct sum of two sets is given by

$$S \oplus T = \{\mathbf{v} \in \mathbb{R}^k : \ \mathbf{v} = \mathbf{s} + \mathbf{t} \text{ for some } \mathbf{s} \in S, \ \mathbf{t} \in t\}. \tag{1.2.4}$$

Lemma 1.2.1. *For any vector space S, T in $\mathbb{R}^k$ we have*

$$S^\perp \cap T^\perp = (S \oplus T)^\perp. \tag{1.2.5}$$

Proof. Let $\mathbf{v} \in S^\perp \cap T^\perp$. For every $\mathbf{w} \in (S \oplus T)$ there are $\mathbf{s} \in S$, $\mathbf{t} \in T$ such that $\mathbf{w} = \mathbf{s} + \mathbf{t}$. Then

$$\mathbf{w}^t\mathbf{v} = \mathbf{s}^t\mathbf{v} + \mathbf{t}^t\mathbf{v} = 0. \tag{1.2.6}$$

Thus, $\mathbf{w}^t\mathbf{v} = 0,\ \forall\ \mathbf{w} \in (S \oplus T)$, so $\mathbf{v} \in (S \oplus T)^\perp$.

Now let $\mathbf{v} \in (S \oplus T)^\perp$. Since S, T are vector spaces, $\mathbf{0} \in S$ and $\mathbf{0} \in T$, thus $\mathbf{s} = \mathbf{s} + \mathbf{0} \in S \oplus T$ and $\mathbf{t} = \mathbf{0} + \mathbf{t} \in S \oplus T$, which yields

$$\mathbf{s}^t\mathbf{v} = 0,\ \forall\ \mathbf{s} \in S, \text{ and } \mathbf{t}^t\mathbf{v} = 0,\ \forall\ \ t \in T, \tag{1.2.7}$$

which yields $\mathbf{v} \in S^\perp \cap T^\perp$. Thus, we have proven (1.2.5). □

Let us denote matrices in $\mathbb{R}^{k\times l}$ by

$$\mathbf{M} = \begin{pmatrix} M_{11} & \dots & M_{1l} \\ \vdots & \ddots & \vdots \\ M_{k1} & \dots & M_{kl} \end{pmatrix}, \tag{1.2.8}$$

and equip them with the norm

$$\|\mathbf{M}\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{M}\mathbf{x}\|}{\|\mathbf{x}\|}. \tag{1.2.9}$$

The transpose of $\mathbf{M}$ is denoted by $\mathbf{M}^t$, its range or column space by

$$\mathcal{R}(\mathbf{M}) = \{\mathbf{v} \in \mathbb{R}^k :\ \mathbf{M}\mathbf{w} = \mathbf{v} \text{ for some vector } \mathbf{w}\}, \tag{1.2.10}$$

and its null space by

$$\mathcal{N}(\mathbf{M}) = \{\mathbf{v} \in \mathbb{R}^l :\ \mathbf{M}\mathbf{v} = \mathbf{0}\}. \tag{1.2.11}$$

Lemma 1.2.2. *For any matrix $\mathbf{M} \in \mathbb{R}^{k\times l}$, we have*

$$\mathcal{R}(\mathbf{M}^t) = \mathcal{N}(\mathbf{M})^\perp. \tag{1.2.12}$$

Proof. If $\mathbf{y} = \mathbf{M}\mathbf{x}$ and $\mathbf{M}^t\mathbf{y}' = \mathbf{0}$ then

$$\langle \mathbf{y}, \mathbf{y}' \rangle = \langle \mathbf{M}\mathbf{x}, \mathbf{y}' \rangle = \langle \mathbf{x}, \mathbf{M}^t\mathbf{y}' \rangle = \langle \mathbf{x}, \mathbf{0} \rangle = 0. \tag{1.2.13}$$

Thus, all vectors from the range of $\mathbf{M}^t$ are orthogonal to all vectors from the null space of $\mathbf{M}$. The rest follows from dimension counting. □

In §4.3.5 we need the following property:

Lemma 1.2.3. *For two matrices* $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{k\times l}$*, we have*

$$\mathcal{N}(\mathbf{M}) \cap \mathcal{N}(\mathbf{N}) = (\mathcal{R}(\mathbf{M}^t) \oplus \mathcal{R}(\mathbf{N}^t))^\perp. \tag{1.2.14}$$

Proof. It is well-known that $(S^\perp)^\perp = S$ for any subspace of a finite dimensional vector space V. Thus, $(\mathcal{N}(\mathbf{M})^\perp)^\perp = \mathcal{N}(\mathbf{M})$, which, together with (1.2.12), yields

$$\mathcal{R}(\mathbf{M}^t)^\perp = (\mathcal{N}(\mathbf{M})^\perp)^\perp = \mathcal{N}(\mathbf{M}), \quad \mathcal{R}(\mathbf{N}^t)^\perp = (\mathcal{N}(\mathbf{N})^\perp)^\perp = \mathcal{N}(\mathbf{N}). \tag{1.2.15}$$

Combining (1.2.15) with Lemma 1.2.1 infers

$$\mathcal{N}(\mathbf{M}) \cap \mathcal{N}(\mathbf{N}) = \mathcal{R}(\mathbf{M}^t)^\perp \cap \mathcal{R}(\mathbf{M}^t)^\perp = (\mathcal{R}(\mathbf{M}^t) \oplus \mathcal{R}(\mathbf{N}^t))^\perp, \tag{1.2.16}$$

which proves the Lemma. □

Definition 1.2.4. *A matrix* $\mathbf{M} \in \mathbb{R}^{k\times k}$ *is* diagonalizable over the reals*, or simply* real diagonalizable*, if and only if there is an invertible eigenvector matrix* $\mathbf{P} \in \mathbb{R}^{k\times k}$ *and a diagonal matrix* $\mathbf{\Lambda} \in \mathbb{R}^{k\times k}$ *such that* $\mathbf{M} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$.

Lemma 1.2.5. *Let* $\mathbf{M} \in \mathbb{R}^{k\times k}$ *be diagonalizable as*

$$\mathbf{M} = \mathbf{P}\ \mathit{diag}(\lambda_1, \ldots, \lambda_k)\mathbf{P}^{-1}. \tag{1.2.17}$$

If $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_k)$*, we have*

$$\mathcal{R}(\mathbf{M}) = \mathit{span}\{\mathbf{p}_i : \lambda_i \neq 0,\ 1 \leq i \leq k\}, \tag{1.2.18}$$

and

$$\mathcal{N}(\mathbf{M}) = \mathit{span}\{\mathbf{p}_i : \lambda_i = 0,\ 1 \leq i \leq k\}. \tag{1.2.19}$$

Proof. Note that $(\lambda_i, \mathbf{p}_i)$, $1 \leq i \leq k$, denote the eigenpairs of $\mathbf{M}$, *i.e.*

$$\mathbf{M}\mathbf{p}_i = \lambda_i\mathbf{p}_i,\ 1 \leq i \leq k. \tag{1.2.20}$$

Since $\mathbf{P}$ is invertible, there is a $\boldsymbol{\alpha} = \mathbf{P}^{-1}\mathbf{v}$ for every $\mathbf{v} \in \mathbb{R}^k$ such that

$$\mathbf{v} = \mathbf{P}\boldsymbol{\alpha} = \sum_{i=1}^{k} \alpha_i\mathbf{p}_i, \tag{1.2.21}$$

which yields

$$\begin{aligned}
\mathcal{R}(\mathbf{M}) &= \{\mathbf{v} \in \mathbb{R}^k :\ \mathbf{M}\mathbf{w} = \mathbf{v} \text{ for some vector } \mathbf{w}\} \\
&= \{\mathbf{v} \in \mathbb{R}^k :\ \mathbf{M}\mathbf{P}\boldsymbol{\alpha} = \sum_{i=1}^{k} \alpha_i\lambda_i\mathbf{p}_i \text{ for some vector } \mathbf{w} = \mathbf{P}\boldsymbol{\alpha}\} \\
&= \operatorname{span}\{\mathbf{p}_i : \lambda_i \neq 0, 1 \leq i \leq k\},
\end{aligned} \tag{1.2.22}$$

and

$$\begin{aligned}\mathcal{N}(\mathbf{M}) &= \{\mathbf{v} \in \mathbb{R}^k : \ \mathbf{M}\mathbf{v} = \mathbf{0}\} \\ &= \{\mathbf{v} = \sum_{i=1}^{d} \alpha_i \mathbf{p}_i : \ \sum_{i=1}^{d} \alpha_i \lambda_i \mathbf{p}_i = \mathbf{0}\} \\ &= \{\mathbf{v} = \sum_{i=1}^{d} \alpha_i \mathbf{p}_i : \ \alpha_i = 0 \text{ or } \lambda_i = 0 \ \forall \ 1 \leq i \leq k\} \\ &= \text{span}\{\mathbf{p}_i : \lambda_i = 0, \ 1 \leq i \leq k\}. \end{aligned} \tag{1.2.23}$$

This proves the lemma. □

Definition 1.2.6. *The square matrix pair* $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{k \times k}$ *is* commuting *if and only if* $\mathbf{MN} = \mathbf{NM}$.

Definition 1.2.7. *The square matrix pair* $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{k \times k}$ *is* simultaneously diagonalizable*, if and only if there is an invertible matrix* $\mathbf{P}$ *and diagonal matrices* $\mathbf{\Lambda}_1, \ \mathbf{\Lambda}_2$ *such that* $\mathbf{M} = \mathbf{P}\mathbf{\Lambda}_1\mathbf{P}^{-1}$ *and* $\mathbf{N} = \mathbf{P}\mathbf{\Lambda}_2\mathbf{P}^{-1}$.

The following result, needed in §1.3, is shown in [61], page 82.

Lemma 1.2.8. *Two diagonalizable matrices are simultaneously diagonalizable if and only if they commute.*

The following matrices will be needed to define the numerical boundary flux in §1.3.

Definition 1.2.9. *For every matrix* $\mathbf{M} \in \mathbb{R}^{k \times k}$ *that is real diagonalizable, such that*

$$\mathbf{M} = \mathbf{P} \ \mathit{diag}(\lambda_1, \ldots, \lambda_m)\mathbf{P}^{-1}, \tag{1.2.24a}$$

we define

$$\mathbf{M}^+ = \mathbf{P} \ \mathit{diag}(\mathit{max}(\lambda_1, 0), \ldots, \mathit{max}(\lambda_m, 0))\mathbf{P}^{-1}, \tag{1.2.24b}$$

$$\mathbf{M}^- = \mathbf{P} \ \mathit{diag}(\mathit{min}(\lambda_1, 0), \ldots, \mathit{min}(\lambda_m, 0))\mathbf{P}^{-1}, \tag{1.2.24c}$$

and

$$\mathit{sgn}(\mathbf{M}) = \mathbf{P} \ \mathit{diag}(\mathit{sgn}(\lambda_1), \ldots, \mathit{sgn}(\lambda_1))\mathbf{P}^{-1}, \tag{1.2.24d}$$

where

$$\mathit{sgn}(x) = \begin{cases} 1, & \mathit{if}\ x > 0, \\ 0, & \mathit{if}\ x = 0, \\ -1, & \mathit{otherwise}. \end{cases} \tag{1.2.24e}$$

Lemma 1.2.10. *Let* $\mathbf{M} \in \mathbb{R}^{k\times k}$ *be real diagonalizable and let* $\mathbf{M}^+$, $\mathbf{M}^-$ *and* $sgn(\mathbf{M})$ *as defined in Definition 1.2.9. Then*

$$\mathbf{M} = \mathbf{M}^+ + \mathbf{M}^-, \tag{1.2.25a}$$

$$\mathcal{N}(\mathbf{M}^+ - \mathbf{M}^-) = \mathcal{N}(sgn(\mathbf{M})) = \mathcal{N}(\mathbf{M}) \subseteq \mathcal{N}(\mathbf{M}^s), s = +, -, \tag{1.2.25b}$$

$$\mathcal{R}(\mathbf{M}) = \mathcal{R}(sgn(\mathbf{M})), \tag{1.2.25c}$$

and

$$\mathbf{M}^+ sgn(\mathbf{M}) = \mathbf{M}^+, \quad \mathbf{M}^- sgn(\mathbf{M}) = -\mathbf{M}^-. \tag{1.2.25d}$$

If, additionally, $\mathbf{M}$ *is symmetric, then*

$$sgn(\mathbf{M}) \text{ is symmetric, and} \tag{1.2.25e}$$

$$\mathbf{M}^+ \text{ and } -\mathbf{M}^- \text{ are symmetric positive semi-definite.} \tag{1.2.25f}$$

Proof. Equation (1.2.25a) follows directly from Definition 1.2.9.

In order to proof (1.2.25b) and (1.2.25c), note that $\mathbf{M}$ is real diagonalizable. Thus, we can apply Lemma 1.2.5 to $\mathbf{M}$, $\mathbf{M}^+ - \mathbf{M}^-$, $\operatorname{sgn}(\mathbf{M})$, $\mathbf{M}^+$ and $\mathbf{M}^-$ to obtain

$$\mathcal{N}(\mathbf{M}) = \operatorname{span}\{\mathbf{p}_i : \lambda_i = 0,\ 1 \le i \le k\}, \tag{1.2.26a}$$

$$\mathcal{N}(\mathbf{M}^+ - \mathbf{M}^-) = \operatorname{span}\{\mathbf{p}_i : max(\lambda_i, 0) - min(\lambda_i, 0) = 0,\ 1 \le i \le k\} = \mathcal{N}(\mathbf{M}), \tag{1.2.26b}$$

$$\mathcal{N}(\operatorname{sgn}(\mathbf{M})) = \operatorname{span}\{\mathbf{p}_i : \operatorname{sgn}(\lambda_i) = 0,\ 1 \le i \le k\} = \mathcal{N}(\mathbf{M}), \tag{1.2.26c}$$

$$\mathcal{N}(\mathbf{M}^+) = \operatorname{span}\{\mathbf{p}_i : max(\lambda_i, 0) = 0,\ 1 \le i \le k\} \subseteq \mathcal{N}(\mathbf{M}), \tag{1.2.26d}$$

$$\mathcal{N}(\mathbf{M}^-) = \operatorname{span}\{\mathbf{p}_i : min(\lambda_i, 0) = 0,\ 1 \le i \le k\} \subseteq \mathcal{N}(\mathbf{M}), \tag{1.2.26e}$$

which proves (1.2.25b).

Applying Lemma 1.2.5 to $\mathbf{M}$ and $\operatorname{sgn}(\mathbf{M})$ further yields

$$\mathcal{R}(\mathbf{M}) = \operatorname{span}\{\mathbf{p}_i : \lambda_i \neq 0, 1 \le i \le k\} = \mathcal{R}(\mathbf{M}), \tag{1.2.27a}$$

$$\mathcal{R}(\operatorname{sgn}(\mathbf{M})) = \operatorname{span}\{\mathbf{p}_i : \operatorname{sgn}(\lambda_i) \neq 0, 1 \le i \le k\} = \mathcal{R}(\mathbf{M}), \tag{1.2.27b}$$

which proves (1.2.25c).

Further, by the definition of $\mathbf{M}^+$ and $\mathbf{M}^-$ and $\operatorname{sgn}(\mathbf{M})$ in Definition 1.2.9, we have

$$\begin{aligned}\mathbf{M}^+\operatorname{sgn}(\mathbf{M}) &= \mathbf{P}\ \operatorname{diag}(\max(\lambda_1, 0), \ldots, \max(\lambda_k, 0))\ \operatorname{diag}(\operatorname{sgn}(\lambda_1), \ldots, \operatorname{sgn}(\lambda_k))\mathbf{P}^{-1} \\ &= \mathbf{M}^+, \end{aligned} \tag{1.2.28a}$$

$$\begin{aligned}\mathbf{M}^-\operatorname{sgn}(\mathbf{M}) &= \mathbf{P}\ \operatorname{diag}(\min(\lambda_1, 0), \ldots, \max(\lambda_k, 0))\ \operatorname{diag}(\operatorname{sgn}(\lambda_1), \ldots, \operatorname{sgn}(\lambda_k))\mathbf{P}^{-1} \\ &= -\mathbf{M}^-, \end{aligned} \tag{1.2.28b}$$

yielding (1.2.25d).

If $\mathbf{M}$ is symmetric, then $\mathbf{P}^{-1} = \mathbf{P}^t$, which, combined with the definition of $\mathbf{M}^+$ and $\mathbf{M}^-$ and $\operatorname{sgn}(\mathbf{M})$ in Definition 1.2.9, yields that $\mathbf{M}^+$ and $\mathbf{M}^-$ and $\operatorname{sgn}(\mathbf{M})$ are symmetric, which, together with the fact that $max(a,0),\ -min(a,0) \geq 0$ for all $a \in \mathbb{R}$, yields (1.2.25e) and (1.2.25f). □

In Chapter 3 and Chapter 4 we will use the following result from [68], page 563.

Lemma 1.2.11. (Singular value decomposition) *For every real square matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$ there exist orthogonal matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{k\times k}$, and a diagonal matrix $\mathbf{D} \in \mathbb{R}^{k\times k}$ such that*

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^t. \tag{1.2.29}$$

Definition 1.2.12. *For every real square matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$ with singular value decomposition*

$$\mathbf{M} = \mathbf{U}\ diag(\lambda_1, \dots, \lambda_k)\mathbf{V}^t, \tag{1.2.30a}$$

the pseudoinverse is defined by

$$\mathbf{M}^\dagger = \mathbf{U}\ diag(\lambda_1^\dagger, \dots, \lambda_k^\dagger)\mathbf{V}^t, \tag{1.2.30b}$$

where

$$x^\dagger = \begin{cases} x^{-1}, & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases} \quad x \in \mathbb{R}. \tag{1.2.30c}$$

The following result is shown in [39].

Lemma 1.2.13. *For every square matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$, $\mathbf{M}\mathbf{M}^\dagger = (\mathbf{M}\mathbf{M}^\dagger)^t$ is the orthogonal projection onto $\mathcal{R}(\mathbf{M})$ and $\mathbf{M}^\dagger\mathbf{M} = (\mathbf{M}^\dagger\mathbf{M})^t$ is the orthogonal projection onto $\mathcal{N}(\mathbf{M})^\perp$.*

This immediately yields another lemma:

Lemma 1.2.14. *Let $\mathbf{I}$ denote the identity matrix. For every square matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$, $\mathbf{I} - \mathbf{M}^\dagger\mathbf{M}$ is the projection onto $\mathcal{N}(\mathbf{M})$.*

If, additionally, $\mathbf{M}$ is symmetric, then

$$(\mathbf{I} - \mathbf{M}^\dagger\mathbf{M})\mathbf{M}^s = \mathbf{0},\ s = +, -. \tag{1.2.31}$$

Proof. By Lemma 1.2.13,

$$(\mathbf{I} - \mathbf{M}^\dagger\mathbf{M})\mathbf{v} = \mathbf{v},\ \forall\ \mathbf{v} \in \mathcal{N}(\mathbf{M}), \tag{1.2.32a}$$

$$(\mathbf{I} - \mathbf{M}^\dagger\mathbf{M})\mathbf{v} = \mathbf{0},\ \forall\ \mathbf{v} \in \mathcal{N}(\mathbf{M})^\perp, \tag{1.2.32b}$$

which yields that $\mathbf{I} - \mathbf{M}^\dagger\mathbf{M}$ is the projection onto $\mathcal{N}(\mathbf{M})$.

Now let $\mathbf{M}$ be symmetric. Then (1.2.25f), Lemma 1.2.12 and (1.2.25b) yields

$$\mathcal{R}(\mathbf{M}^s) = \mathcal{R}((\mathbf{M}^s)^t) = \mathcal{N}(\mathbf{M}^s)^\perp \subseteq \mathcal{N}(\mathbf{M})^\perp,\ s = +, -, \tag{1.2.33}$$

which, together with (1.2.32b), yields (1.2.31). □

In Chapter 5 we will use the following properties.

Definition 1.2.15. *Two matrices* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{k\times k}$ *are* similar *if and only if there exists an invertible matrix* $\mathbf{P} \in \mathbb{R}^{k\times k}$ *such that* $\mathbf{PAP}^{-1} = \mathbf{B}$.

Lemma 1.2.16. *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{k\times k}$ *be* similar*, such that* $\mathbf{PAP}^{-1} = \mathbf{B}$*. The following holds for every* $\mathbf{v} \in \mathbb{R}^k$*:*

i) $\mathbf{v} \in \mathcal{N}(\mathbf{A})$ *if and only if* $\mathbf{Pv} \in \mathcal{N}(\mathbf{B})$,

ii) $\mathbf{v} \in \mathcal{R}(\mathbf{A})$ *if and only if* $\mathbf{Pv} \in \mathcal{R}(\mathbf{B})$,

iii) $\mathbf{v} \in \mathcal{N}(\mathbf{A}^t)$ *if and only if* $\mathbf{P}^{-t}\mathbf{v} \in \mathcal{N}(\mathbf{B}^t)$*, and*

iv) $\mathbf{v} \in \mathcal{N}(\mathbf{A})^\perp$ *if and only if* $\mathbf{P}^{-t}\mathbf{v} \in \mathcal{R}(\mathbf{B}^t)$.

Proof. Since $\mathbf{B}(\mathbf{Pv}) = (\mathbf{PAP}^{-1})\mathbf{Pv} = \mathbf{PAv}$, we obtain

$$\mathbf{v} \in \mathcal{N}(\mathbf{A}) \;\Leftrightarrow\; \mathbf{Av} = \mathbf{0} \;\Leftrightarrow\; \mathbf{PAv} = \mathbf{0} \;\Leftrightarrow\; \mathbf{B}(\mathbf{Pv}) = \mathbf{0} \;\Leftrightarrow\; \mathbf{Pv} \in \mathcal{N}(\mathbf{B}), \tag{1.2.34}$$

which yields *i*.

Since $\mathbf{PA} = \mathbf{PA}(\mathbf{P}^{-1}\mathbf{P}) = \mathbf{BP}$, we obtain

$$\mathbf{v} \in \mathcal{R}(\mathbf{A}) \;\Leftrightarrow\; \exists \mathbf{w} :\; \mathbf{v} = \mathbf{Aw} \;\Leftrightarrow\; \exists \mathbf{w} :\; \mathbf{Pv} = \mathbf{PAw} \tag{1.2.35a}$$
$$\Leftrightarrow\; \exists \mathbf{w} :\; (\mathbf{Pv}) = \mathbf{B}(\mathbf{Pw}) \;\Leftrightarrow\; \mathbf{Pv} \in \mathcal{R}(\mathbf{B}), \tag{1.2.35b}$$

which yields *ii*.

Since $\mathbf{B}^t(\mathbf{P}^{-t}\mathbf{v}) = (\mathbf{P}^{-1}\mathbf{A}^t\mathbf{P}^t)\mathbf{P}^{-t}\mathbf{v} = \mathbf{P}^{-1}\mathbf{A}^t\mathbf{v}$, we obtain

$$\mathbf{v} \in \mathcal{N}(\mathbf{A}^t) \;\Leftrightarrow\; \mathbf{A}^t\mathbf{v} = \mathbf{0} \;\Leftrightarrow\; \mathbf{P}^{-t}\mathbf{A}^t\mathbf{v} = \mathbf{0} \tag{1.2.36a}$$
$$\Leftrightarrow\; \mathbf{B}^t(\mathbf{P}^{-t}\mathbf{v}) = \mathbf{0} \;\Leftrightarrow\; \mathbf{P}^{-t}\mathbf{v} \in \mathcal{N}(\mathbf{B}^t), \tag{1.2.36b}$$

which yields *iii*.

Since $\mathbf{P}^{-t}\mathbf{A}^t = \mathbf{P}^{-t}\mathbf{A}^t(\mathbf{P}^t\mathbf{P}^{-t}) = \mathbf{B}^t\mathbf{P}^{-t}$, we obtain

$$\mathbf{v} \in \mathcal{N}(\mathbf{A})^\perp = \mathcal{R}(\mathbf{A}^t) \;\Leftrightarrow\; \exists \mathbf{w} :\; \mathbf{v} = \mathbf{A}^t\mathbf{w} \tag{1.2.37a}$$
$$\Leftrightarrow\; \exists \mathbf{w} :\; \mathbf{P}^{-t}\mathbf{v} = \mathbf{P}^{-t}\mathbf{Aw} \tag{1.2.37b}$$
$$\Leftrightarrow\; \exists \mathbf{w} :\; (\mathbf{P}^{-t}\mathbf{v}) = \mathbf{B}^t(\mathbf{P}^{-t}\mathbf{w}) \tag{1.2.37c}$$
$$\Leftrightarrow\; \mathbf{P}^{-t}\mathbf{v} \in \mathcal{R}(\mathbf{B}^t), \tag{1.2.37d}$$

which yields *iv*. □

Lemma 1.2.17. *Let* $\mathbf{M} \in \mathbb{R}^{k\times k}$ *be real diagonalizable and* similar *to* $\mathbf{N}$*, such that* $\mathbf{RMR}^{-1} = \mathbf{N}$*. Then*

$$\mathbf{N}^+ = \mathbf{RM}^+\mathbf{R}^{-1}, \quad \mathbf{N}^- = \mathbf{RM}^-\mathbf{R}^{-1}, \text{ and} \tag{1.2.38a}$$

$$sgn(\mathbf{N}) = \mathbf{R}\ sgn(\mathbf{M})\mathbf{R}^{-1}. \tag{1.2.38b}$$

If additionally, $\mathbf{R}^t\mathbf{RM}$ *is symmetric, then* $\mathbf{N}$ *is symmetric.*

Proof. Since $\mathbf{M}$ is real diagonalizable by $\mathbf{M} = \mathbf{P\Lambda P}^{-1}$, $\mathbf{N}$ is diagonalizable as

$$\mathbf{N} = \mathbf{Q\Lambda Q}^{-1}, \text{ where } \mathbf{Q} = \mathbf{RP}. \tag{1.2.39a}$$

By Definition 1.2.9,

$$\mathbf{N}^+ = \mathbf{Q\Lambda}^+\mathbf{Q}^{-1} = \mathbf{RM}^+\mathbf{R}^{-1}, \tag{1.2.39b}$$

$$\mathbf{N}^- = \mathbf{Q\Lambda}^-\mathbf{Q}^{-1} = \mathbf{RM}^-\mathbf{R}^{-1}, \tag{1.2.39c}$$

and

$$\operatorname{sgn}(\mathbf{N}) = \mathbf{Q}\operatorname{sgn}(\mathbf{\Lambda})\mathbf{Q}^{-1} = \mathbf{R}\operatorname{sgn}(\mathbf{M})\mathbf{R}^{-1}. \tag{1.2.39d}$$

Now let us assume that $\mathbf{R}^t\mathbf{RM}$ is symmetric and show that $\mathbf{N}$ is also symmetric by writing

$$\mathbf{N} = \mathbf{RMR}^{-1} = \mathbf{R}^{-t}(\mathbf{R}^t\mathbf{RM})\mathbf{R}^{-1}, \tag{1.2.40}$$

which, since $\mathbf{R}^t\mathbf{RM}$ is symmetric, implies that $\mathbf{N}^t = \mathbf{N}$. □

The following definition and lemma can be found, *e.g.*, in [41], pg. 331:

Definition 1.2.18. *The Drazin inverse of a matrix* $\mathbf{M} \in \mathbb{R}^{k\times k}$ *is defined as the unique matrix* $\mathbf{M}^\ddagger$ *such that*

$$\mathbf{M}^\ddagger\mathbf{M}\mathbf{M}^\ddagger = \mathbf{M}^\ddagger, \quad \mathbf{M}\mathbf{M}^\ddagger = \mathbf{M}^\ddagger\mathbf{M}, \quad \mathbf{M}^{k+1}\mathbf{M}^\ddagger = \mathbf{M}^k, \tag{1.2.41}$$

where $k = index(\mathbf{M})$*. The index of* $\mathbf{M}$ *is the smallest nonnegative integer* k *such that* $rank(\mathbf{M}^k) = rank(\mathbf{M}^{k+1})$*.*

Lemma 1.2.19. *For every diagonalizable matrix* $\mathbf{M} \in \mathbb{R}^{k\times k}$*, such that*

$$\mathbf{M} = \mathbf{P}\ diag(\lambda_1, \ldots, \lambda_k)\mathbf{P}^{-1}, \tag{1.2.42a}$$

the Drazin inverse is

$$\mathbf{M}^\ddagger = \mathbf{P}\ diag(\lambda_1^\dagger, \ldots, \lambda_k^\dagger)\mathbf{P}^{-1}, \tag{1.2.42b}$$

with $x^\dagger$ *defined in (1.2.30c).*

Lemma 1.2.20. *Let* $\mathbf{M}$ *be diagonalizable and let* $\mathbf{M}^\ddagger$ *satisfy (1.2.42). The following holds for every* $\mathbf{v} \in \mathbb{R}^k$*:*

i) $\mathbf{M}^\ddagger\mathbf{M}\mathbf{v}$ *is the projection of* $\mathbf{v}$ *into* $\mathcal{R}(\mathbf{M})$*, and*

ii) $(\mathbf{I}-\mathbf{M}^\ddagger\mathbf{M})\mathbf{v}$ *is the projection of* $\mathbf{v}$ *into* $\mathcal{N}(\mathbf{M})$*.*

Furthermore,

$$(\mathbf{M}^\ddagger)^t = (\mathbf{M}^t)^\ddagger, \tag{1.2.43a}$$

and

$$\mathcal{R}(\mathbf{M}^\ddagger) = \mathcal{R}(\mathbf{M}), \tag{1.2.43b}$$

We note that this projection is oblique, *i.e.* not orthogonal, in general.

Proof. By Lemma 1.2.19, $\mathbf{M}^\ddagger$ and $\mathbf{M}$ are commuting, since

$$\mathbf{M}^\ddagger\mathbf{M} = \mathbf{P}\mathbf{\Lambda}^\dagger\mathbf{\Lambda}\mathbf{P}^{-1} = \mathbf{M}\mathbf{M}^\ddagger, \tag{1.2.44}$$

thus

$$\mathbf{M}^\ddagger\mathbf{M}\mathbf{v} = \mathbf{M}(\mathbf{M}^\ddagger\mathbf{v}) \in \mathcal{R}(\mathbf{M}). \tag{1.2.45}$$

By Lemma 1.2.5, $\mathcal{R}(\mathbf{M}) = \operatorname{span}\{\mathbf{p}_i : \lambda_i \neq 0, 1 \leq i \leq k\}$, where $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_k)$. Thus, we can write any $\mathbf{v} \in \mathcal{R}(\mathbf{M})$ as

$$\mathbf{v} = \sum_{i:\ \lambda_i \neq 0} \alpha_i \mathbf{p}_i. \tag{1.2.46}$$

Therefore,

$$\mathbf{M}^\ddagger\mathbf{M}\mathbf{v} = \sum_{i:\ \lambda_i \neq 0} \alpha_i \mathbf{M}^\ddagger\mathbf{M}\mathbf{p}_i = \sum_{i:\ \lambda_i \neq 0} \alpha_i \lambda_i^{-1}\lambda_i\mathbf{p}_i = \mathbf{v},\ \forall\ \mathbf{v} \in \mathcal{R}(\mathbf{M}). \tag{1.2.47}$$

which combined with (1.2.45), yields *i.*

By Lemma 1.2.19,

$$\mathbf{M}(\mathbf{I}-\mathbf{M}^\ddagger\mathbf{M}) = \mathbf{P}(\mathbf{\Lambda}-\mathbf{\Lambda}\mathbf{\Lambda}^\ddagger\mathbf{\Lambda})\mathbf{P}^{-1} = \mathbf{0}, \tag{1.2.48}$$

which proves

$$(\mathbf{I}-\mathbf{M}^\ddagger\mathbf{M})\mathbf{v} \in \mathcal{N}(\mathbf{M}). \tag{1.2.49}$$

Further note that

$$(\mathbf{I}-\mathbf{M}^\ddagger\mathbf{M})\mathbf{v} = \mathbf{v} - \mathbf{M}^\ddagger(\mathbf{M}\mathbf{v}) = \mathbf{v},\ \forall\ \mathbf{v} \in \mathcal{N}(\mathbf{M}), \tag{1.2.50}$$

which combined with (1.2.49), yields *ii.*

Equation (1.2.43a) follows directly from Lemma 1.2.19, since $\mathbf{\Lambda}^\ddagger = (\mathbf{\Lambda}^\ddagger)^t$.

Since $\mathbf{M}$ is diagonalizable, we can apply Lemma 1.2.5 to $\mathbf{M}$ and $\mathbf{M}^\ddagger$ to obtain (1.2.43b). □

In Chapter 6, we will define the Lax-Friedrichs numerical flux using the matrices defined below.

Definition 1.2.21. *For every diagonalizable matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$ such that*

$$\mathbf{M} = \mathbf{P}\,diag(\lambda_1, \ldots, \lambda_k)\mathbf{P}^{-1}, \tag{1.2.51a}$$

we define

$$\breve{\mathbf{M}}^{+} = \frac{\mathbf{M} + C\mathbf{I}}{2}, \quad \breve{\mathbf{M}}^{-} = \frac{\mathbf{M} - C\mathbf{I}}{2}, \tag{1.2.51b}$$

where

$$C = \max_{1\le j\le k} |\lambda_j|. \tag{1.2.51c}$$

We further define

$$\mathbf{M}^{(0)} = C\mathbf{M}^{\ddagger}, \quad \mathbf{M}^{(1)} = C^{\dagger}\mathbf{M}, \tag{1.2.51d}$$

where $\mathbf{M}^{\ddagger}$ satisfies (1.2.42) and $C^{\dagger}$ is defined in (1.2.30c).

Lemma 1.2.22. *Let $\mathbf{M} \in \mathbb{R}^{k\times k}$ be a diagonalizable matrix, and let $\breve{\mathbf{M}}^{+}$, $\breve{\mathbf{M}}^{-}$, $\mathbf{M}^{(0)}$ and $\mathbf{M}^{(1)}$ be as in Definition 1.2.21. Then*

$$\breve{\mathbf{M}}^{+} + \breve{\mathbf{M}}^{-} = \mathbf{M}, \tag{1.2.52a}$$

$$\breve{\mathbf{M}}^{+} - \breve{\mathbf{M}}^{-} = C\mathbf{I}, \tag{1.2.52b}$$

$$\mathcal{R}(\mathbf{M}) = \mathcal{R}(\mathbf{M}^{(0)}) = \mathcal{R}(\mathbf{M}^{(1)}), \tag{1.2.52c}$$

$$\breve{\mathbf{M}}^{+} - \breve{\mathbf{M}}^{+}\mathbf{M}^{(1)} + \breve{\mathbf{M}}^{-} + \breve{\mathbf{M}}^{-}\mathbf{M}^{(1)} = \mathbf{0}, \tag{1.2.52d}$$

and, if $\mathbf{M}$ is invertible,

$$\breve{\mathbf{M}}^{+} - \breve{\mathbf{M}}^{+}\mathbf{M}^{(0)} - \breve{\mathbf{M}}^{-} - \breve{\mathbf{M}}^{-}\mathbf{M}^{(0)} = \mathbf{0}. \tag{1.2.52e}$$

If, additionally, $\mathbf{M}$ is symmetric then

$$\mathbf{M}^{(0)}, \mathbf{M}^{(1)} \text{ are symmetric}, \tag{1.2.52f}$$

$$\breve{\mathbf{M}}^{+} \text{ and } -\breve{\mathbf{M}}^{-} \text{ are symmetric positive semi-definite.} \tag{1.2.52g}$$

Proof. Equations (1.2.52a) and (1.2.52b) follow directly from Definition 1.2.21.

In order to prove (1.2.52c), note that $\mathbf{M}$ is diagonalizable, thus we can apply Lemma 1.2.5 to $\mathbf{M}$, $\mathbf{M}^{(0)}$ and $\mathbf{M}^{(1)}$ to obtain

$$\mathcal{R}(\mathbf{M}) = \operatorname{span}\{\mathbf{p}_i : \lambda_i \neq 0, 1 \le i \le k\}, \tag{1.2.53a}$$

$$\mathcal{R}(\mathbf{M}^{(0)}) = \operatorname{span}\{\mathbf{p}_i : C\lambda_1^{\dagger} \neq 0, 1 \le i \le k\} = \mathcal{R}(\mathbf{M}), \tag{1.2.53b}$$

$$\mathcal{R}(\mathbf{M}^{(1)}) = \operatorname{span}\{\mathbf{p}_i : C^{\dagger}\lambda_i \neq 0, 1 \le i \le k\} = \mathcal{R}(\mathbf{M}), \tag{1.2.53c}$$

which proves (1.2.52c).

To prove (1.2.52d), note that

$$\boldsymbol{\Sigma}^{(1)} = 2\mathbf{P}^{-1}\left(\breve{\mathbf{M}}^{+} - \breve{\mathbf{M}}^{+}\mathbf{M}^{(1)} + \breve{\mathbf{M}}^{-} + \breve{\mathbf{M}}^{-}\mathbf{M}^{(1)}\right)\mathbf{P} \tag{1.2.54a}$$

is a diagonal matrix with zero diagonal values, since

$$(\boldsymbol{\Sigma})_{i,i}^{(1)} = (\lambda_i + C) - (\lambda_i + C)C^{\dagger}\lambda_i + (\lambda_i - C) + (\lambda_i - C)C^{\dagger}\lambda_i = 0,\ 1 \le i \le k. \tag{1.2.54b}$$

To prove (1.2.52e), we note that $\mathbf{M}^{(0)} = C\mathbf{M}^{-1}$ for invertible $\mathbf{M}$, thus

$$\boldsymbol{\Sigma}^{(0)} = 2\mathbf{P}^{-1}\left(\breve{\mathbf{M}}^{+} - \breve{\mathbf{M}}^{+}\mathbf{M}^{(0)} - \breve{\mathbf{M}}^{-} - \breve{\mathbf{M}}^{-}\mathbf{M}^{(0)}\right)\mathbf{P} \tag{1.2.55a}$$

is a diagonal matrix with zero diagonal values, since

$$(\boldsymbol{\Sigma})_{i,i}^{(0)} = (\lambda_i + C) - (\lambda_i + C)\frac{C}{\lambda_i} - (\lambda_i - C) - (\lambda_i - C)\frac{C}{\lambda_i} = 0,\ 1 \le i \le k.$$

If $\mathbf{M}$ is symmetric, then $\mathbf{P}^{-1} = \mathbf{P}^t$, which, combined with the definition of $\breve{\mathbf{M}}^{+}$ and $\breve{\mathbf{M}}^{-}$, $\mathbf{M}^{(0)}$ and $\mathbf{M}^{(1)}$ in Definition 1.2.21, yields that $\breve{\mathbf{M}}^{+}$ and $\breve{\mathbf{M}}^{-}$, $\mathbf{M}^{(0)}$ and $\mathbf{M}^{(1)}$ are symmetric, which proves (1.2.52f).

Since $\mathbf{M}^{+}$ and $\mathbf{M}^{-}$ are symmetric and $C \ge \lambda_i$, $1 \le i \le k$, we obtain

$$\breve{\mathbf{M}}^{+} = \frac{1}{2}\mathbf{P}\ \mathrm{diag}(\lambda_1 + C, \ldots, \lambda_k + C)\mathbf{P}^t \ge 0, \tag{1.2.56}$$

$$-\breve{\mathbf{M}}^{-} = \frac{1}{2}\mathbf{P}\ \mathrm{diag}(C - \lambda_1, \ldots, C - \lambda_k)\mathbf{P}^t \ge 0, \tag{1.2.57}$$

which yields (1.2.52g). □

1.2.2 Notations

By [18], the L^q-norm of a function $\mathbf{v} : D \to \mathbb{R}^k$ on a domain D is defined by

$$\|\mathbf{v}\|_{q,D} = \left(\int_D \sum_{i=1}^{k} |\mathbf{v}_i|^q \, d\mathbf{x}\right)^{1/q},\ 1 \le q < \infty, \tag{1.2.58}$$

and, for $q = \infty$,

$$\|\mathbf{v}\|_{\infty,D} = \operatorname*{ess\,sup}_{\mathbf{x}\in D} \|\mathbf{v}(\mathbf{x})\|. \tag{1.2.59}$$

These are the norms of the Banach spaces

$$L^q(D) = \left\{\mathbf{v}(\mathbf{x}) :\ \|\mathbf{v}\|_{q,D} \le \infty\right\},\ 1 \le q \le \infty. \tag{1.2.60}$$

Hölder's inequality yields

$$\int_D \mathbf{v}^t\mathbf{w}\,d\mathbf{x} \leq \|\mathbf{v}\|_{q,D}\|\mathbf{w}\|_{r,D},\ \forall\ \mathbf{v},\mathbf{w}\in L^r(D),\ 1\leq q\leq r\leq\infty,\ \frac{1}{q}+\frac{1}{r}=1. \tag{1.2.61}$$

Let $|D|=\int_D 1\,d\mathbf{x}$ denote the volume of D. Then

$$\|f\|_{q,D}\leq |D|^{1/q-1/r}\|f\|_{r,D},\ \forall\ \mathbf{v}\in L^r(D),\ 1\leq q<r\leq\infty, \tag{1.2.62}$$

and the Cauchy-Schwarz inequality

$$\int_D \mathbf{v}^t\mathbf{w}\,d\mathbf{x} \leq \|\mathbf{v}\|_{2,D}\|\mathbf{w}\|_{2,D},\ \forall\ \mathbf{v},\mathbf{w}\in L^2(D), \tag{1.2.63}$$

which both are a consequence of Hölder's inequality.

We denote the polynomials in $\mathbf{x}\in\mathbb{R}^d$ of total degree at most p by

$$\mathbb{P}^p=\left\{\sum\nolimits_{|\boldsymbol{\alpha}|\leq p} c_{\boldsymbol{\alpha}}\mathbf{x}^{\boldsymbol{\alpha}}:\ c_{\boldsymbol{\alpha}}\in\mathbb{R}\right\},\ p\geq 0. \tag{1.2.64}$$

The inverse inequality in [18] states that, for any real $1\leq q_1\leq q_2\leq\infty$, there exists a positive constant C independent of $|D|$ such that

$$\|f\|_{q_2,D}\leq C|D|^{1/q_2-1/q_1}\|f\|_{q_1,D},\ \forall\ f(\mathbf{x})\in\mathbb{P}_p. \tag{1.2.65}$$

Let S denote any surface in $\mathbb{R}^k$ with smooth parametrization

$$S=\{\mathbf{x}=\mathbf{x}(\mathbf{s}):\ \mathbf{s}=(s_1,\ldots,s_{k-1})\in\hat{S}\},\ \hat{S}\in\mathbb{R}^{k-1}. \tag{1.2.66}$$

Then we define the integral of a function of $\mathbf{x}$ over S as

$$\int_S \mathbf{f}(\mathbf{x})\,d\mathbf{s}=\int_{\hat{S}}\mathbf{f}(\mathbf{x}(\mathbf{s}))\left|\bigwedge\left(\tfrac{\partial\mathbf{x}}{\partial s_1},\ldots,\tfrac{\partial\mathbf{x}}{\partial s_{k-1}}\right)\right| ds_1\ldots ds_{k-1}, \tag{1.2.67}$$

where $|\bigwedge(\mathbf{v}_1,\ldots,\mathbf{v}_n)|$ denotes the hypervolume of the region bounded by its arguments.

We denote a multi-index by $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_k),\ \alpha_i\geq 0$ integers, and define

$$|\boldsymbol{\alpha}|=\sum_{i=1}^k\alpha_i,\quad \boldsymbol{\alpha}!=\prod_{i=1}^k\alpha_i!,\quad D^{\boldsymbol{\alpha}}=\frac{\partial^{|\boldsymbol{\alpha}|}}{\partial\mathbf{x}^{\boldsymbol{\alpha}}}=\prod_{i=1}^k\frac{\partial^{\alpha_i}}{\partial x_i^{\alpha_i}}. \tag{1.2.68}$$

Finally, we abbreviate the partial derivative of a function f with respect to the variable x by

$$f_{,x}=\frac{\partial f}{\partial x}. \tag{1.2.69}$$

1.3 Problem Statement

Let d be the space dimension, $\mathbf{x} = (x_1, \ldots, x_d)^t$ the space variable defined on a domain $\Omega \in \mathbb{R}^d$, and t the time variable defined on $[0, T]$. Let $\mathbf{A}_1, \ldots, \mathbf{A}_d$ be constant matrices in $\mathbb{R}^{m\times m}$, where m is the size of the system. We seek to find $\mathbf{u} : (0, T) \times \Omega \to \mathbb{R}^m$, $\mathbf{u} = (u_1, \ldots, u_m)^t$ that satisfies the linear first-order system

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{i=1}^{d} \mathbf{A}_i \frac{\partial \mathbf{u}}{\partial x_i} = \mathbf{g}(t, \mathbf{x}),\ \mathbf{x} \in \Omega,\ 0 < t < T, \tag{1.3.1a}$$

with *source term* $\mathbf{g} : (0, T) \times \Omega \to \mathbb{R}^m$, subject to the initial and boundary conditions

$$\mathbf{u}(0, \mathbf{x}) = \mathbf{u}_0(\mathbf{x}),\ \mathbf{x} \in \Omega, \tag{1.3.1b}$$

$$\left(\sum_{i=1}^{d} \nu_i \mathbf{A}_i^{\bar{\mu}_i}\right) \mathbf{u}(t, \mathbf{x}) = \left(\sum_{i=1}^{d} \nu_i \mathbf{A}_i^{\bar{\mu}_i}\right) \mathbf{u}_B(t, \mathbf{x}),\ \mathbf{x} \in \partial\Omega,\ 0 < t < T, \tag{1.3.1c}$$

where $\partial\Omega$ denotes the boundary of Ω, $\boldsymbol{\nu}$ denotes the unit outward normal on $\partial\Omega$, and

$$\mu_i = sign(\nu_i), \quad \bar{\mu}_i = sign(-\nu_i), \quad 1 \le i \le d, \tag{1.3.2}$$

where

$$sign(x) = \begin{cases} +, & \text{if } x \ge 0, \\ -, & \text{if } x < 0. \end{cases} \tag{1.3.3}$$

Using the divergence theorem, problem (1.3.1) satisfies

$$\frac{d}{dt} \int_D \mathbf{u}\, d\mathbf{x} = \int_D \mathbf{g}\, d\mathbf{x} + \int_{\partial D} \sum_{i=1}^{d} \nu_i \mathbf{A}_i \mathbf{u}\, d\mathbf{s}, \tag{1.3.4}$$

for every domain D in Ω with ∂D denoting the boundary of D with outward unit normal $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)$, *i.e.*, the rate of change of $\int_D \mathbf{u}\, d\mathbf{x}$ is equal to the quantity created by the source on D and the flux through the boundary ∂D. Thus, if the support of $\mathbf{u}$ is contained in Ω and no source term is present, $\int_\Omega \mathbf{u}\, d\mathbf{x}$ is conserved. For this reason, the system (1.3.1a) is said to be *in conservative form*, if $\mathbf{g} = \mathbf{0}$, and *in balanced form*, if $\mathbf{g} \neq \mathbf{0}$.

Throughout this dissertation, we shall only consider *hyperbolic* systems, which are defined in [16] as follows:

Definition 1.3.1. *The system (1.3.1a) is* hyperbolic *if and only if the following two properties hold.*

i) The matrices $\mathbf{A}(\boldsymbol{\mu}) = \sum_{i=1}^{d} \mu_i \mathbf{A}_i$, $\forall\ \boldsymbol{\mu} \in \mathbb{R}^d$ *are diagonalizable with real eigenvalues,*

$$\mathbf{A}(\boldsymbol{\mu}) = \mathbf{P}(\boldsymbol{\mu})\ diag(\lambda_1(\boldsymbol{\mu}), \ldots, \lambda_m(\boldsymbol{\mu}))\ \mathbf{P}(\boldsymbol{\mu})^{-1},\ \lambda_1(\boldsymbol{\mu}), \ldots, \lambda_m(\boldsymbol{\mu}) \in \mathbb{R},\ and \tag{1.3.5}$$

ii) Their diagonalization is well-conditioned:

$$\sup_{\|\boldsymbol{\mu}\|=1} \|\mathbf{P}(\boldsymbol{\mu})\| \cdot \|\mathbf{P}(\boldsymbol{\mu})^{-1}\| < \infty. \tag{1.3.6}$$

Definition 1.3.2. *The system (1.3.1) is* Friedrichs symmetric*, or simply* symmetric*, if and only if the matrices* $\mathbf{A}_i$*,* $1 \leq i \leq d$*, are symmetric.*

Definition 1.3.3. *The system (1.3.1) is* Friedrichs symmetrizable*, or simply* symmetrizable*, if and only if there is a symmetric positive definite matrix* $\mathbf{S}_0$ *such that the matrices*

$$\mathbf{S}_i = \mathbf{S}_0\mathbf{A}_i, \ 1 \leq i \leq d, \tag{1.3.7}$$

are symmetric.

Definition 1.3.4. *The system (1.3.1) is* commuting and real diagonalizable *if and only if the matrices* $\mathbf{A}_i$*,* $1 \leq i \leq d$*, are commuting and real diagonalizable.*

Lemma 1.3.5. *If the system (1.3.1) is commuting and real diagonalizable, then it is symmetrizable.*

Proof. Let (1.3.1) be commuting and real diagonalizable. By Lemma 1.2.8, there exists a eigenvector matrix $\mathbf{P}$ and diagonal matrices $\boldsymbol{\Lambda}_i$ such that

$$\mathbf{A}_i = \mathbf{P}\boldsymbol{\Lambda}_i\mathbf{P}^{-1}, \ 1 \leq i \leq d. \tag{1.3.8}$$

Now choose $\mathbf{S}_0 = \mathbf{P}^{-t}\mathbf{P}^{-1}$, which is positive definite. Then

$$\mathbf{S}_0\mathbf{A}_i = \mathbf{P}^{-t}\boldsymbol{\Lambda}_i\mathbf{P}^{-1}, \ 1 \leq i \leq d, \tag{1.3.9}$$

are symmetric. □

Lemma 1.3.6. *If the system (1.3.1) is symmetrizable, then it is hyperbolic.*

Proof. The proof of this lemma is taken from [16], page 14:

Let (1.3.1) be symmetrizable. Then $\mathbf{S}_0^{-1}$ is positive definite and admits a (unique) square root $\mathbf{R}$ symmetric positive definite (see [64], page 78). Let us denote $\mathbf{S}(\boldsymbol{\mu}) = \sum_{i=1}^{d} \mu_i\mathbf{S}_i$. Then

$$\mathbf{A}(\boldsymbol{\mu}) = \mathbf{S}_0^{-1}\mathbf{S}(\boldsymbol{\mu}) = \mathbf{R}(\mathbf{R}\mathbf{S}(\boldsymbol{\mu})\mathbf{R})\mathbf{R}^{-1}. \tag{1.3.10}$$

The matrix $\mathbf{R}\mathbf{S}(\boldsymbol{\mu})\mathbf{R}$ is real symmetric and thus may be written as $\mathbf{Q}(\boldsymbol{\mu})^t\mathbf{S}(\boldsymbol{\mu})\mathbf{Q}(\boldsymbol{\mu})$, where $\mathbf{Q}$ is orthogonal. Then $\mathbf{A}(\boldsymbol{\mu})$ is conjugated to $\mathbf{D}(\boldsymbol{\mu})$, $\mathbf{A}(\boldsymbol{\mu}) = \mathbf{P}(\boldsymbol{\mu})^{-1}\mathbf{D}(\boldsymbol{\mu})\mathbf{P}(\boldsymbol{\mu})$, with $\mathbf{P}(\boldsymbol{\mu}) = \mathbf{Q}(\boldsymbol{\mu})\mathbf{R}^{-1}$ and $\mathbf{P}(\boldsymbol{\mu})^{-1} = \mathbf{R}\mathbf{Q}(\boldsymbol{\mu})^t$. Since our matrix norm is invariant under left or right multiplication by unitary matrices, we have

$$\|\mathbf{P}(\boldsymbol{\mu})\|\|\mathbf{P}(\boldsymbol{\mu})^{-1}\| = \|\mathbf{R}\|\|\mathbf{R}^{-1}\|, \tag{1.3.11}$$

a number independent of $\boldsymbol{\mu}$. The diagonalization is thus well conditioned. □

Note. *Note that if the system (1.3.1) is symmetric, then it is symmetrizable with* $\mathbf{S}_0 = \mathbf{I}$*. Then Lemma 1.3.5 and 1.3.6 shows that, if the system 1.3.1 is symmetric or commuting and real diagonalizable, then it is symmetrizable and therefore hyperbolic.*

1.4 Research Goals

In this dissertation we aim to develop efficient and asymptotically exact *a posteriori* estimates of the DG discretization error for first-order linear symmetric or symmetrizable hyperbolic systems. Such estimates are necessary to verify the numerical accuracy of the solution and to guide and stop mesh adaptivity processes.

We develop a new and modified discontinuous Galerkin scheme for the space discretization of linear multi-dimensional hyperbolic systems of conservation laws. We choose an enriched polynomial space $\mathcal{P}_p$, $\mathbb{P}_p \subset \mathcal{P}_p \subset \mathbb{P}_{p+1}$, as a basis for the function space $\mathcal{V}_p^h$ on each element. We perform a local error analysis which shows that the leading term of the discretization error lies in a polynomial subspace spanned by a linear combination of Legendre polynomials of order p and $p+1$. For special hyperbolic systems, where the coefficient matrices are nonsingular, we show that the leading term of the error is spanned by $(p+1)^{th}$-degree Radau polynomials. We also establish new pointwise and averaged $\mathcal{O}(h^{p+2})$ superconvergence results.

We then turn our attention to the construction of a new implicit residual-based *a posteriori* error estimation procedure. By solving a small system of equations based on the local residual of the PDE locally on each element, we compute an cost-efficient estimate of the discretization error . For systems with invertible matrices, the error can be estimated by a static problem. For general systems, however, part of the error has to be computed by solving a transient system of equations. Local error analysis suggests that, for smooth solutions, both error estimates are asymptotically correct, that is they converge to the real error under h- and/or p-refinement. We first show these results for linear symmetric systems that satisfy certain assumptions, then for general linear symmetric systems. We then generalize these results to linear symmetrizable systems by considering an equivalent symmetric formulation, which requires us to make small modifications in the error estimation procedure. Numerical results confirm the results of our analysis, for both the symmetric and the symmetrizable case in one, two and three space dimensions. Examples include the linearized Euler's equations, Maxwell's equations, and the acoustic wave equation, as well as several other systems.

We further investigate the behavior of the discretization error for the Lax-Friedrichs numerical flux. We observe that, while no superconvergence results can be obtained, the error estimation results can be recovered in most cases. We further develop simple h- and p-refinement techniques to show that the error estimates can be successfully used to guide the refinement and coarsening process. Finally, we present numerical results where we apply our DG formulation to some nonlinear problems.

1.5 Outline

The thesis is organized as follows: In Chapter 1, we give an overview of the topics discussed in this dissertation, in particular hyperbolic systems of conservation laws, discontinuous Galerkin (DG) methods, and *a posteriori* error estimation. We introduce some notation and results from linear algebra in §1.2 that will be used within this dissertation. We state the initial-boundary value problem for a system of conservation laws in §1.3.

In Chapter 2, we present the discontinuous Galerkin formulation for linear systems of hyperbolic conservation laws in multiple space dimensions. In § 2.2, we introduce the approximation operators for the initial and boundary conditions; in § 2.3, we introduce a time-stepping method used to integrate in time.

In Chapter 3, we discuss linear symmetrizable hyperbolic systems of conservation laws in two spatial dimensions that satisfy certain assumptions. We present a local error analysis on the element $\omega = (0, h)^2$ and establish an asymptotic expansion of the local discretization error. In §3.2.1 we present new pointwise and average superconvergence results. We construct an *a posteriori* error estimation procedure in §3.2.2, where we split the error into two parts and estimate each part separately. Then we show that the error estimates are asymptotically exact under mesh refinement for smooth solutions.

In Chapter 4, we show that the results of Chapter 3 hold for general linear symmetric hyperbolic systems in multiple space dimensions.

In Chapter 5, we generalize these results further for linear symmetrizable hyperbolic systems by showing that each symmetrizable system can be reduced to a symmetric system. While the superconvergence results extend straightforward, we have to slightly modify the error estimation procedure for symmetrizable systems.

In Chapter 6, we investigate the behavior of the local discretization error for the Lax-Friedrichs numerical flux. While no superconvergence results can be shown, we are able to develop *a posteriori* error estimates for the Lax-Friedrichs flux under certain assumptions.

In Chapter 7, we show how the error estimation results can be applied to guide both h- and p-adaptive processes. We demonstrate the usefulness of our error estimation results by a numerical example. Finally, we apply our DG method to some nonlinear hyperbolic systems in §7.4.

At the end of each chapter we present some numerical results for one-, two- and three-dimensional hyperbolic systems that validate our theory.

We conclude with a few remarks and a discussion of our results in Chapter 8.

Chapter 2

Discontinuous Galerkin Formulation for Hyperbolic Conservation Laws

In this chapter, we present a discontinuous Galerkin formulation for linear systems of hyperbolic conservation laws in multiple space dimensions. In § 2.2, we introduce the approximation operators for initial and boundary conditions; in § 2.3, we introduce an explicit time-stepping method used to integrate in time.

2.1 DG Formulation

In order to discretize (1.3.1), we will use the following polynomial spaces:

Definition 2.1.1. *Let $\mathcal{P}_p$ denote the polynomials in $\mathbf{x}$ with coefficients in $\mathbb{R}^m$ of total degree at most $p+1$ and of degree at most p in each space variable $x_1, \ldots, x_d$, that is*

$$\mathcal{P}_p = \left\{ \sum\nolimits_{\boldsymbol{\alpha} \in \mathcal{A}_p} \mathbf{c}_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} : \ \mathbf{c}_{\boldsymbol{\alpha}} \in \mathbb{R}^m \right\}, \ p \geq 0, \tag{2.1.1a}$$

where

$$\mathcal{A}_p = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^d : \ |\boldsymbol{\alpha}| \leq p+1, \ \max_{1 \leq i \leq d} \boldsymbol{\alpha}_i \leq p \right\}. \tag{2.1.1b}$$

Then $[\mathbb{P}_p]^m \subseteq \mathcal{P}_p \subset [\mathbb{P}_{p+1}]^m$. An illustration of $\mathcal{P}_p$ for $d = 2$ is given in Figure 2.1.1.

Now we are ready to define a partition and a finite element space on Ω.

Definition 2.1.2. *Let the partition $\mathcal{T}_h$ of the domain $\Omega = (0,1)^d$ into a uniform mesh consisting of N^d square elements of size $h = N^{-1}$ be defined as*

$$\mathcal{T}_h = \left\{ \prod\nolimits_{i=1}^{d} (n_i h, n_i h + h) : \ 0 \leq n_i < N, \ 1 \leq i \leq d \right\}. \tag{2.1.2}$$

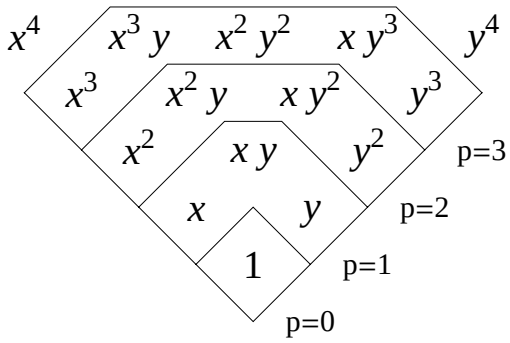

Figure 2.1.1: Polynomial basis of $\mathcal{P}_p$ for $p = 0, 1, 2, 3$ and $d = 2$

Then the finite element space $\mathcal{V}_p^h$ *on* $[0,T] \times \Omega$ *is defined by*

$$\mathcal{V}_p^h = \{\mathbf{v}(t,\mathbf{x}) : \ \mathbf{v}|_{\omega_h} \in \mathcal{P}_p, \ \omega_h \in \mathcal{T}_h, 0 \le t \le T\}. \tag{2.1.3}$$

A weak formulation of (1.3.1) is obtained by multiplying (1.3.1a) by a test function $\mathbf{v}$, integrating over an arbitrary element $\omega_h \in \mathcal{T}_h$ and applying Green's identity to write

$$\int_{\omega_h} \mathbf{v}^t \left(\frac{\partial \mathbf{u}}{\partial t} - \mathbf{g} \right) d\mathbf{x} = \sum_{i=1}^{d} \int_{\omega_h} \frac{\partial \mathbf{v}^t}{\partial x_i} \mathbf{A}_i \mathbf{u} \, d\mathbf{x} - \int_{\partial\omega_h} \mathbf{v}^t \nu_i \mathbf{A}_i \mathbf{u} \, d\mathbf{s}, \ \omega_h \in \mathcal{T}_h, \ 0 \le t \le T, \tag{2.1.4}$$

where $\partial\omega_h$ denotes the boundary of ω_h and $\boldsymbol{\nu}$ its outward unit normal.

Before we discretize (2.1.4) we need to define a *numerical flux* of $\mathbf{u}_h$ on $\partial\omega_h$, since $\mathcal{V}_p^h$ allows discontinuities on the boundary $\partial\omega_h$ for any $\omega_h \in \mathcal{T}_h$. We will use the Steger-Warming flux splitting, as defined in [69].

Definition 2.1.3. *Let the traces of* $\mathbf{u}_h \in \mathcal{V}_p^h$ *on* $\partial\omega_h$ *be defined by*

$$\mathbf{u}_h^+(t,\mathbf{x}) = \lim_{\epsilon \to 0^+} \mathbf{u}_h(t, \mathbf{x} - \epsilon\boldsymbol{\nu}), \quad \mathbf{u}_h^-(t,\mathbf{x}) = \lim_{\epsilon \to 0^+} \mathbf{u}_h(t, \mathbf{x} + \epsilon\boldsymbol{\nu}), \quad \mathbf{x} \in \partial\omega_h, \tag{2.1.5a}$$

where $\boldsymbol{\nu}$ *denotes the unit outward normal on* $\partial\omega_h$*, and define*

$$\mu_i = sign(\nu_i), \quad \bar{\mu}_i = sign(-\nu_i), \quad 1 \le i \le d, \tag{2.1.5b}$$

where

$$sign(x) = \begin{cases} +, & \text{if } x \ge 0, \\ -, & \text{if } x < 0. \end{cases} \tag{2.1.5c}$$

Then the Steger-Warming numerical flux *of* $\mathbf{u}_h \in \mathcal{V}_p^h$ *on the boundary* $\partial\omega_h$ *for* $\omega_h \in \mathcal{T}_h$ *will be defined by replacing the flux* $\sum_{i=1}^d \nu_i \mathbf{A}_i \mathbf{u}_h$ *on* $\partial\omega_h$ *by the numerical flux*

$$\mathbf{h}(\mathbf{u}_h^+, \mathbf{u}_h^-, \boldsymbol{\nu}) = \sum_{i=1}^{d} \nu_i (\mathbf{A}_i^{\mu_i} \mathbf{u}_h^+ + \mathbf{A}_i^{\bar{\mu}_i} \mathbf{u}_h^-), \ \mathbf{x} \in \partial\omega_h, \ \omega_h \in \mathcal{T}_h, \tag{2.1.6}$$

where $\mathbf{A}_i^+$, $\mathbf{A}_i^-$, $1 \le i \le d$, *are defined in Definition 1.2.9.*

Then the discontinuous Galerkin method consists of finding $\mathbf{u}_h \in \mathcal{V}_p^h$ that satisfies

$$\int_{\omega_h} \mathbf{v}^t \left(\frac{\partial \mathbf{u}_h}{\partial t} - \mathbf{g} \right) d\mathbf{x} = \sum_{i=1}^{d} \left(\int_{\omega_h} \frac{\partial \mathbf{v}^t}{\partial x_i} \mathbf{A}_i \mathbf{u}_h \, d\mathbf{x} - \int_{\partial\omega_h} \mathbf{v}^t \nu_i (\mathbf{A}_i^{\mu_i} \mathbf{u}_h^+ + \mathbf{A}_i^{\bar{\mu}_i} \mathbf{u}_h^-) \, d\mathbf{s} \right), \quad (2.1.7)$$
$$\omega_h \in \mathcal{T}_h, \ \mathbf{v} \in \mathcal{P}_p, \ 0 < t < T.$$

We will define approximations of the initial and boundary conditions $\mathbf{u}_0$ and $\mathbf{u}_B$ by functions in $\mathcal{V}_p^h$ in §2.2.

This yields an ODE in time for $\mathbf{u}_h$. In our numerical experiments, we use a temporal error tolerance much smaller than the spatial error by applying a high-order Runge-Kutta method to integrate in time. However, for the purpose of analyzing the behavior of the spatial discretization error, we assume the evolution in time to be exact.

2.2 Approximation of the Initial and Boundary Conditions

In order to approximate the initial conditions $\mathbf{u}_0$ on every element $\omega_h \in \mathcal{T}_h$ and the boundary conditions $\mathbf{u}_B$ on every edge on the boundary of Ω by functions in $\mathcal{P}_p$, we define some special approximation operators such that the resulting approximation error is consistent with the discontinuous Galerkin discretization error. For simplicity, we will only consider the approximation on the element $\omega = (0, h)^d$. We need the following definitions:

Definition 2.2.1. *We denote the* reference element *by $\Delta = (0, 1)^d$, its boundary by Γ, and the unit outward normal on Γ by $\boldsymbol{\nu}$. We split $\Gamma = \bigcup_{i=1}^{d} \Gamma_i$, where*

$$\Gamma_i = \Gamma_i^- \cup \Gamma_i^+, \quad \Gamma_i^- = \{\boldsymbol{\xi} \in \Delta : \ \xi_i = 0\}, \quad \Gamma_i^+ = \{\boldsymbol{\xi} \in \Delta : \ \xi_i = 1\}, \quad 1 \leq i \leq d. \quad (2.2.1)$$

For a function on Γ_i, we denote the surface integral by

$$\int_{\Gamma_i^s} f(\boldsymbol{\xi}) \, d\boldsymbol{\sigma} = \int_{(0,1)^{d-1}} f(\boldsymbol{\xi}) \, d\hat{\boldsymbol{\xi}}, \quad \boldsymbol{\xi} \in \Gamma_i^s, \ 1 \leq i \leq d, \ s = +, -, \quad (2.2.2)$$

where $\hat{\boldsymbol{\xi}} = (\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_d)^t$.

For $\omega = (0, h)^d$ there is an affine transformation $\mathbf{x} : \Delta \to \omega$, $\mathbf{x}(\boldsymbol{\xi}) = h\boldsymbol{\xi}$. We split $\partial\omega = \bigcup_{i=1}^{d} \gamma_i$, where

$$\gamma_i = \gamma_i^- \cup \gamma_i^+, \quad \gamma_i^\pm = \mathbf{x}(\Gamma_i^\pm), \ 1 \leq i \leq d. \quad (2.2.3)$$

An illustration of Δ for $d = 2$ is shown in Figure 2.2.1.

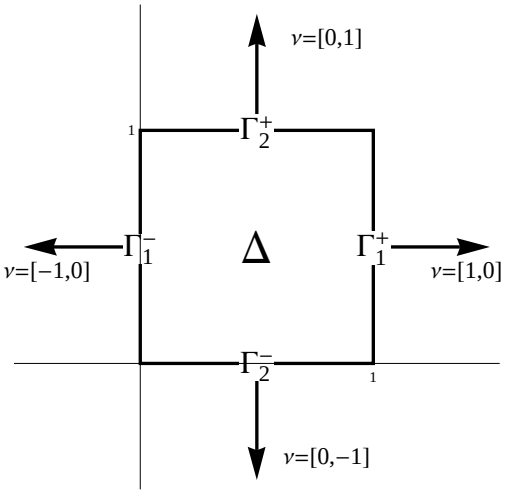

Figure 2.2.1: The reference element $\Delta = (0,1)^d$ for $d = 2$ with boundary Γ and outer normal unit vectors $\boldsymbol{\nu}$.

Definition 2.2.2. *The* Legendre polynomials of degree p *are defined, for instance, by Rodrigues' formula on* $[-1,1]$ *in [1]:*

$$\tilde{L}_p(x) = \frac{1}{2^p p!}\frac{d^p}{dx^p}[(x^2-1)^p],\ p \geq 0. \tag{2.2.4}$$

Shifted Legendre polynomials of degree p *on* $[0,1]$ *are defined by*

$$L_p(\xi) = \tilde{L}_p(2\xi - 1),\ p \geq 0. \tag{2.2.5}$$

$L_p(\xi)$ *is orthogonal to all polynomials of degree not exceeding* $p-1$ *and satisfies*

$$\int_0^1 L_p(\xi)L_q(\xi)d\xi = \frac{\delta_{pq}}{2j+1},\ \int_0^1 L_p(\xi)L'_{p+1}(\xi)d\xi = 2, \tag{2.2.6}$$

where δ_{pq} *is the Kronecker delta, which is equal to 1 if* $p = q$ *and 0 otherwise.*

The shifted left *and* right Radau polynomials of degree p *on* $[0,1]$ *are defined by*

$$R_p^-(\xi) = L_p(\xi) + L_{p-1}(\xi) \text{ and } R_p^+(\xi) = L_p(\xi) - L_{p-1}(\xi),\ p \geq 1. \tag{2.2.7}$$

Now consider the p^{th}-degree Taylor polynomial of a function $\mathbf{v}(\mathbf{x}) \in [C^{p+2}(\bar{\omega})]^m$ about $\mathbf{x} = \mathbf{0}$,

$$T_{p+1}\mathbf{v}(\mathbf{x}) = \sum_{|\alpha|=0}^{p+1} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}}\mathbf{v}(\mathbf{0})\mathbf{x}^{\boldsymbol{\alpha}}, \tag{2.2.8}$$

and define the *dual set of* i by

$$D(i) = \{1, 2, \ldots, i-1, i+1, \ldots, d\},\ 1 \leq i \leq d. \tag{2.2.9}$$

Then we can define the operators Π and π to approximate the initial conditions and π_i, $1 \leq i \leq d$, to approximate the boundary conditions as follows.

Definition 2.2.3. *We define the approximation operators* $\Pi : [C^{p+2}(\bar{\omega})]^m \to \mathcal{P}_p$ *and* $\pi : [C^{p+2}(\bar{\omega})]^m \to \mathcal{P}_p$ *as*

$$\Pi \mathbf{v}(\mathbf{x}) = T_{p+1}\mathbf{v}(\mathbf{x}) - h^{p+1} \sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right) \mathbf{c}_j, \tag{2.2.10a}$$

and

$$\pi \mathbf{v}(\mathbf{x}) = T_{p+1}\mathbf{v}(\mathbf{x}) - h^{p+1} \sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right) \mathbf{c}_j - L_p\left(\frac{x_j}{h}\right) sgn(\mathbf{A}_j)\mathbf{c}_j, \tag{2.2.10b}$$

where

$$\mathbf{c}_j = \frac{1}{a_{p+1}} \frac{1}{(p+1)!} \frac{\partial^{p+1}\mathbf{v}(\mathbf{0})}{\partial x_j^{p+1}},\ 1 \leq j \leq d, \tag{2.2.10c}$$

and a_{p+1} *denotes the coefficient of* ξ^{p+1} *in* $L_{p+1}(\xi)$.

We further define the approximation operator $\pi_i : [C^{p+2}(\bar{\gamma}_i)]^m \to \mathcal{P}_p$ *for* $1 \leq i \leq d$ *on* γ_i *as*

$$\pi_i \mathbf{v}(\mathbf{x})\Big|_{x_i=a} = T_{p+1}\mathbf{v}(\mathbf{x}) - h^{p+1} \sum_{j \in D(i)} L_{p+1}\left(\frac{x_j}{h}\right) \mathbf{c}_j - L_p\left(\frac{x_j}{h}\right) sgn(\mathbf{A}_j)\mathbf{c}_j,\ 1 \leq i \leq d, \tag{2.2.10d}$$

with $a = 0$ *for* γ_i^- *and* $a = h$ *for* γ_i^+, *and where the coefficients* $\mathbf{c}_j$ *in* π_i *satisfy the conditions of (2.2.10c) for* $j \in D(i)$, $1 \leq i \leq d$.

These approximations satisfy the following lemma:

Lemma 2.2.4. *Let* $\omega = (0, h)^d$, $\mathbf{v}(\mathbf{x}) \in [C^{p+2}(\bar{\omega})]^m$ *and* $\boldsymbol{\xi} = h^{-1}\mathbf{x}$. *Let* $\pi\mathbf{v} \in \mathcal{P}_p$ *on* ω *and* $\pi_i\mathbf{v} \in \mathcal{P}_p$ *on the boundary* γ_i, $1 \leq i \leq d$, *defined by (2.2.10b) and (2.2.10d), where* $\mathbf{c}_i$ *satisfy (2.2.10c). Then there exists a positive constant* C *independent of* h *such that*

$$\left\| \mathbf{v}(\mathbf{x}) - \Pi\mathbf{v}(\mathbf{x}) - h^{p+1} \sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right) \mathbf{c}_j \right\|_{\infty,\omega} \leq Ch^{p+2}, \tag{2.2.11a}$$

$$\left\| \mathbf{v}(\mathbf{x}) - \pi\mathbf{v}(\mathbf{x}) - h^{p+1} \sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right) \mathbf{c}_j - L_p\left(\frac{x_j}{h}\right) sgn(\mathbf{A}_j)\mathbf{c}_j \right\|_{\infty,\omega} \leq Ch^{p+2}, \tag{2.2.11b}$$

and

$$\left\| \mathbf{v}(\mathbf{x}) - \pi_i\mathbf{v}(\mathbf{x}) - h^{p+1} \sum_{j \in D(i)} L_{p+1}\left(\frac{x_j}{h}\right) \mathbf{c}_j - L_p\left(\frac{x_j}{h}\right) sgn(\mathbf{A}_i)\mathbf{c}_j \right\|_{\infty,\gamma_i} \leq Ch^{p+2}, \tag{2.2.11c}$$

$$1 \leq i \leq d.$$

Proof. Applying Maclaurin series to $\mathbf{v} \in [C^{p+2}(\bar{\omega})]^m$ yields

$$\mathbf{v}(\mathbf{x}) = T_{p+1}\mathbf{v}(\mathbf{x}) + \sum_{|\alpha|=p+2} \mathbf{R}_\alpha(\mathbf{x})\mathbf{x}^\alpha, \tag{2.2.12}$$

where the remainder can be bounded on ω as

$$\|\mathbf{v}(\mathbf{x}) - T_{p+1}\mathbf{v}(\mathbf{x})\|_{\infty,\omega} = \left\| \sum_{|\boldsymbol{\alpha}|=p+2} \mathbf{R}_{\boldsymbol{\alpha}}(\mathbf{x})\mathbf{x}^\alpha \right\|_{\infty,\omega} \leq Ch^{p+2}. \tag{2.2.13}$$

From the definition of $\pi\mathbf{v}$ in equation (2.2.10b) we observe that

$$\mathbf{v}(\mathbf{x}) - T_{p+1}\mathbf{v}(\mathbf{x}) = \mathbf{v}(\mathbf{x}) - \Pi\mathbf{v}(\mathbf{x}) - h^{p+1}\sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j \tag{2.2.14a}$$

$$= \mathbf{v}(\mathbf{x}) - \pi\mathbf{v}(\mathbf{x}) - h^{p+1}\sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j - L_p\left(\frac{x_j}{h}\right)\operatorname{sgn}(\mathbf{A}_j)\mathbf{c}_j. \tag{2.2.14b}$$

Substituting (2.2.14) into (2.2.13) we establish (2.2.11b) and (2.2.11a).

Following the same line of reasoning we establish (2.2.11c), which concludes the proof. □

We note that the approximation $T_{p+1}\mathbf{v}$ defined in (2.2.10b) is only used for the analysis. In practice we first project $\mathbf{v}$ onto $[\mathbb{P}_{p+1}]^m$ as

$$\mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}|\leq p+1} \frac{\int_\Delta \mathbf{v}(\boldsymbol{\xi})\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\,d\boldsymbol{\xi}}{\int_\Delta \psi_{\boldsymbol{\alpha}}^2(\boldsymbol{\xi})\,d\boldsymbol{\xi}}\psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \text{ where } \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \prod_{i=1}^{d} L_{\alpha_i}(\xi_i). \tag{2.2.15a}$$

Then define the L^2-projection onto $\mathcal{P}_p$ as

$$\Pi\mathbf{v}(\mathbf{x}) = \mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) - \sum_{i=1}^{d} L_{p+1}\left(\frac{x_i}{h}\right)\bar{\mathbf{c}}_i, \tag{2.2.15b}$$

and a *corrected L^2-projection onto $\mathcal{P}_p$* as

$$\pi\mathbf{v}(\mathbf{x}) = \mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) - \sum_{i=1}^{d} \left(L_{p+1}\left(\frac{x_i}{h}\right)\bar{\mathbf{c}}_i - L_p\left(\frac{x_i}{h}\right)\operatorname{sgn}(\mathbf{A}_i)\bar{\mathbf{c}}_i\right). \tag{2.2.15c}$$

where the coefficients $\bar{\mathbf{c}}_i$ are defined as

$$\bar{\mathbf{c}}_i = \frac{\int_\Delta \mathbf{v}(\boldsymbol{\xi})L_{p+1}(\xi_i)\,d\boldsymbol{\xi}}{\int_\Delta L_{p+1}^2(\xi_i)\,d\boldsymbol{\xi}}, \quad 1 \leq i \leq d. \tag{2.2.15d}$$

Similarly, define the L^2-projection on γ_i^s as

$$\mathcal{L}_{p+1}^{i,s}\mathbf{v}(\mathbf{x}) = \sum_{\substack{|\boldsymbol{\alpha}|\leq p+1\\ \alpha_i=0}} \frac{\int_{\gamma_i^s} \mathbf{v}(\mathbf{x})\psi_{\boldsymbol{\alpha}}\left(\frac{x_i}{h}\right)\,d\mathbf{s}}{\int_{\gamma_i^s}\psi_{\boldsymbol{\alpha}}^2\left(\frac{x_i}{h}\right)\,d\mathbf{s}}\psi_{\boldsymbol{\alpha}}\left(\frac{x_i}{h}\right),\ s=+,-,\ 1\leq i\leq d, \tag{2.2.16a}$$

and let $\pi_i^s\mathbf{v}$ be a *corrected L^2-projection* onto $\mathcal{P}_p$, defined by

$$\pi_i^s\mathbf{v}(\mathbf{x}) = \mathcal{L}_{p+1}^{i,s}\mathbf{v}(\mathbf{x}) - \sum_{j\in D(i)}\left(L_{p+1}\left(\frac{x_j}{h}\right)\bar{\mathbf{c}}_{ij}^s - L_p\left(\frac{x_j}{h}\right)\operatorname{sgn}(\mathbf{A}_j)\bar{\mathbf{c}}_{ij}^s\right), \tag{2.2.16b}$$

where the coefficients $\bar{\mathbf{c}}_{ij}^s$ are defined as

$$\bar{\mathbf{c}}_{ij}^s = \frac{\int_{\Gamma_i^s}\mathbf{v}(h\boldsymbol{\xi})L_{p+1}(\xi_j)\,d\boldsymbol{\sigma}}{\int_{\Gamma_i^s}L_{p+1}^2(\xi_j)\,d\boldsymbol{\sigma}},\quad s=+,-,\ j\in D(i),\ 1\leq i\leq d. \tag{2.2.16c}$$

We note that the approximation capability of Π, π and π_i is not affected by choosing either T_{p+1} or $\mathcal{L}_{p+1}$ as approximation operators, as shown in the following lemma.

Lemma 2.2.5. *Let T_{p+1}, $\mathcal{L}_{p+1}$ and $\mathcal{L}_{p+1}^{i,s}$, respectively, be defined in (2.2.8), (2.2.15a) and (2.2.16a). If $\mathbf{c}_i$, $\bar{\mathbf{c}}_i$ and $\bar{\mathbf{c}}_{ij}^s$ are defined in (2.2.10c), (2.2.15d) and (2.2.16c), then for $\mathbf{v}(\mathbf{x})\in[C^{p+2}(\bar{\omega})]^m$,*

$$\|\mathcal{L}_{p+1}\mathbf{v} - T_{p+1}\mathbf{v}\|_{\infty,\omega}\leq Ch^{p+2}, \tag{2.2.17a}$$

$$\|\bar{\mathbf{c}}_j - h^{p+1}\mathbf{c}_j\|\leq Ch^{p+2},\ 1\leq j\leq d, \tag{2.2.17b}$$

and

$$\|\mathcal{L}_{p+1}^{i,s}\mathbf{v} - T_{p+1}\mathbf{v}\|_{\infty,\gamma_i^s}\leq Ch^{p+2}, \tag{2.2.18a}$$

$$\|\bar{\mathbf{c}}_{ij}^s - h^{p+1}\mathbf{c}_j\|\leq Ch^{p+2},\ j\in D(i),\ 1\leq i\leq d,\ s=+,-. \tag{2.2.18b}$$

Proof. To prove (2.2.17a) note that the L^2-projection satisfies

$$\|\mathbf{v}-\mathcal{L}_{p+1}\mathbf{v}\|_{2,\omega} = \min_{\mathbf{p}\in[\mathbb{P}_{p+1}]^m}\|\mathbf{v}-\mathbf{p}\|_{2,\omega}\leq\|\mathbf{v}-T_{p+1}\mathbf{v}\|_{2,\omega}. \tag{2.2.19}$$

By inequality (1.2.62) and (2.2.13),

$$\|\mathbf{v}-T_{p+1}\mathbf{v}\|_{2,\omega}\leq|\omega|^{\frac{1}{2}}\|\mathbf{v}-T_{p+1}\mathbf{v}\|_{\infty,\omega}\leq Ch^{p+2}|\omega|^{\frac{1}{2}}. \tag{2.2.20}$$

Thus, the inverse inequality (1.2.65), combined with (2.2.19) and (2.2.20), yields

$$\|\mathcal{L}_{p+1}\mathbf{v}-T_{p+1}\mathbf{v}\|_{\infty,\omega}\leq C'|\omega|^{-\frac{1}{2}}\|\mathcal{L}_{p+1}\mathbf{v}-T_{p+1}\mathbf{v}\|_{2,\omega}\leq C''h^{p+2}, \tag{2.2.21}$$

which proves (2.2.17a).

By the definitions of $\bar{\mathbf{c}}_j$ and $\mathcal{L}_{p+1}\mathbf{v}$ in (2.2.15d) and (2.2.15a), we obtain

$$\mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) - \sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right)\bar{\mathbf{c}}_j \in \mathcal{P}_p. \tag{2.2.22}$$

By the definitions of $\mathbf{c}_j$ and $T_{p+1}\mathbf{v}$ in (2.2.10c) and (2.2.8), we obtain

$$T_{p+1}\mathbf{v}(\mathbf{x}) - h^{p+1}\sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j \in \mathcal{P}_p. \tag{2.2.23}$$

Subtracting (2.2.23) from (2.2.22), multiplying (2.2.23) by $L_{p+1}\left(\frac{x_i}{h}\right)$ and integrating w.r.t. x on ω, yields

$$\int_\omega \left(\mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) - T_{p+1}\mathbf{v}(\mathbf{x}) - \sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right)(\bar{\mathbf{c}}_j - h^{p+1}\mathbf{c}_j)\right)L_{p+1}\left(\frac{x_i}{h}\right)\,d\mathbf{x} = 0, \tag{2.2.24}$$

where we used the fact that $\int_\omega \mathbf{p}L_{p+1}\left(\frac{x_i}{h}\right)\,d\mathbf{x} = 0$ for all $\mathbf{p} \in \mathcal{P}_p$.

We reorder terms in (2.2.24) to obtain

$$\|\bar{\mathbf{c}}_i - h^{p+1}\mathbf{c}_i\|\left\|L_{p+1}\left(\frac{x_i}{h}\right)\right\|_{2,\omega}^2 = \int_\omega \left(\mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) - T_{p+1}\mathbf{v}(\mathbf{x})\right)L_{p+1}\left(\frac{x_i}{h}\right)\,d\mathbf{x}. \tag{2.2.25}$$

Applying the Cauchy-Schwarz inequality to (2.2.25) yields and dividing by $\|L_{p+1}(\frac{x_i}{h})\|_{2,\omega}^2$ infers that

$$\|\bar{\mathbf{c}}_i - h^{p+1}\mathbf{c}_i\| \leq \frac{\|\mathcal{L}_{p+1}\mathbf{v} - T_{p+1}\mathbf{v}\|_{2,\omega}}{\left\|L_{p+1}\left(\frac{x_i}{h}\right)\right\|_{2,\omega}} = |\omega|^{-\frac{1}{2}}(2p+3)\left\|\mathcal{L}_{p+1}\mathbf{v} - T_{p+1}\mathbf{v}\right\|_{2,\omega}. \tag{2.2.26}$$

Applying inequality (1.2.62) to (2.2.26) and applying (2.2.17a), we obtain

$$\left\|\bar{\mathbf{c}}_i - h^{p+1}\mathbf{c}_i\right\| \leq (2p+3)\left\|\mathcal{L}_{p+1}\mathbf{v} - T_{p+1}\mathbf{v}\right\|_{\infty,\omega} \leq Ch^{p+2},\ 1 \leq i \leq d. \tag{2.2.27}$$

which proves (2.2.17b).

Similarly we can prove (2.2.18). □

On a general element $\omega = \prod_{i=1}^{d}(n_i, n_i + h),\ n_i \in \mathbb{R}$, we define the linear transformation from Δ to ω by $\mathbf{x}(\boldsymbol{\xi}) = (n_1 + h\xi_1, \ldots, n_d + h\xi_d)^t$ and let $\gamma_i^s = \mathbf{x}(\Gamma_i^s)$, $s = +, -$, $1 \leq i \leq d$ denote the faces of ω.

We also define the approximation operators π_p^ω, Π_p^ω, $\mathcal{L}_{p+1}^\omega$ and $\pi_p^{\gamma_i^s}$ such that $\pi = \pi_p^\omega$, $\Pi = \Pi_p^\omega$, $\mathcal{L}_{p+1} = \mathcal{L}_{p+1}^\omega$ and $\pi_p^{\gamma_i^s} = \pi_i^s$ for $\omega = (0, h)^d$ and polynomial order p.

Then the DG formulation for the initial-boundary value problem (1.3.1) consists of finding $\mathbf{u}_h \in \mathcal{V}_p^h$ that satisfies

$$\int_\omega \mathbf{v}^t \left(\frac{\partial \mathbf{u}_h}{\partial t} - \mathbf{g} \right) d\mathbf{x} = \sum_{i=1}^{d} \left(\int_\omega \frac{\partial \mathbf{v}}{\partial x_i} \mathbf{A}_i \mathbf{u}_h \, d\mathbf{x} - \int_{\partial\omega} \nu_i \mathbf{v}^t (\mathbf{A}_i^{\mu_i} \mathbf{u}_h^+ + \mathbf{A}_i^{\bar{\mu}_i} \mathbf{u}_h^-) \, d\mathbf{s} \right), \quad \forall\ \mathbf{v} \in \mathcal{V}_p^h,\ \omega \in \mathcal{T}_h,\ 0 < t < T, \tag{2.2.28a}$$

subject to the initial and boundary conditions

$$\mathbf{u}_h(0, \mathbf{x}) = \pi_p^\omega \mathbf{u}_0(\mathbf{x}) \text{ or } \mathbf{u}_h(0, \mathbf{x}) = \Pi_p^\omega \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \omega,\ \omega \in \mathcal{T}_h \tag{2.2.28b}$$

$$(\nu_1 \mathbf{A}_1^{\bar{\mu}_1} + \nu_2 \mathbf{A}_2^{\bar{\mu}_2}) \mathbf{u}_h^-(t, \mathbf{x}) = (\nu_1 \mathbf{A}_1^{\bar{\mu}_1} + \nu_2 \mathbf{A}_2^{\bar{\mu}_2}) \pi_p^{\gamma_i^s} \mathbf{u}_B(t, \mathbf{x}), \quad \mathbf{x} \in \gamma_i^s \cap \partial\Omega,\ 1 \le i \le d,\ s = +, -,\ \omega \in \mathcal{T}_h,\ 0 < t < T. \tag{2.2.28c}$$

2.3 Time Integration

Embedded Runge-Kutta methods that solve ordinary differential equations of the form

$$\frac{dy}{dt} = f(t, y), \tag{2.3.1a}$$

are defined by the general formula

$$y_{n+1} = y_n + \Delta t \sum_{i=1}^{s} b_i k_i, \quad y_{n+1}^* = y_n + \Delta t \sum_{i=1}^{s} b_i^* k_i, \tag{2.3.1b}$$

$$k_i = f\left(t_n + c_i \Delta t, y_n + \Delta t \sum_{j=1}^{s} a_{ij} k_j \right), \tag{2.3.1c}$$

where y_{n+1}^* is of higher order of convergence than y_{n+1}, and Δt is chosen such that the temporal discretization error estimate $\|e_{n+1}\|_{2,\Omega} = \|y_{n+1} - y_{n+1}^*\|_{2,\Omega} \le tol$, where the tolerance tol is prescribed by the user.

We use a time tolerance of $tol \le \frac{1}{3} \|\mathbf{u}(T, \mathbf{x}) - \mathcal{L}_{p+1} \mathbf{u}(T, \mathbf{x})\|_{2,\Omega}$ to insure that the temporal error at time T will be far smaller than the spatial discretization error.

To specify a particular method, one needs to provide the number of stages s and the coefficients $a_{ij}, b_j, b_j^*, c_i,\ 1 \le i, j \le s$. These data are usually arranged in a extended Butcher tableau,

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \\ & b_1^* & b_2^* & \dots & b_s^* \end{array}. \tag{2.3.1d}$$

Here we use the explicit Dormand-Prince method, as described in [37], whose extended Butcher tableau is given by

0							
1/5	1/5						
3/10	3/40	9/40					
4/5	44/45	−56/15	32/9				
8/9	19372/6561	−25360/2187	64448/6561	−212/729			
1	9017/3168	−355/33	46732/5247	49/176	−5103/18656		
1	35/384	0	500/1113	125/192	−2187/6784	11/84	
	5179/57600	0	7571/16695	393/640	−92097/339200	187/2100	1/40
	35/384	0	500/1113	125/192	−2187/6784	11/84	0

(2.3.1e)

The first row of b-coefficients gives the fourth-order accurate solution, and the second row b^* has order five.

One could have used a TVD Runge-Kutta scheme. However, since we are only interested in the spatial discretization error, this will not affect our results.

Chapter 3

Error Analysis for Linear Symmetric Hyperbolic Systems

In this chapter we will investigate the asymptotic error behavior for the DG Method on one element $\omega = (0,h)^2$ for $d = 2$ applied to linear symmetric hyperbolic systems under the assumptions of Lemma 3.1.1 below.

Thus, let $\mathbf{A}_1$, $\mathbf{A}_2$ be symmetric matrices satisfying the conditions of Lemma 3.1.1, and let $\mathbf{u} \in [C^2([0,T], C^{p+2}(\bar{\Omega}))]^m$ be the solution of

$$\mathbf{u}_{,t} + \mathbf{A}_1\mathbf{u}_{,x_1} + \mathbf{A}_2\mathbf{u}_{,x_2} = \mathbf{g}(t,\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad 0 < t < T, \tag{3.0.1a}$$

subject to the initial and boundary conditions

$$\mathbf{u}(0,\mathbf{x}) = \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{3.0.1b}$$

$$(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})\mathbf{u}(t,\mathbf{x}) = (\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})\mathbf{u}_B(t,\mathbf{x}), \quad \mathbf{x} \in \partial\Omega,\ 0 < t < T. \tag{3.0.1c}$$

Then the DG formulation on the element ω consists of finding $\mathbf{u}_h \in \mathcal{P}_p$ that satisfies

$$\begin{aligned}&\int_\omega \mathbf{v}^t \left(\mathbf{u}_{h,t} + \mathbf{A}_1\mathbf{u}_{h,x_1} + \mathbf{A}_2\mathbf{u}_{h,x_2} - \mathbf{g}\right) d\mathbf{x} \\ &\quad + \int_{\partial\omega} \mathbf{v}^t(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})(\mathbf{u}_h^- - \mathbf{u}_h^+)\, d\mathbf{s} = 0,\ \forall\ \mathbf{v} \in \mathcal{P}_p,\ 0 < t < T,\end{aligned} \tag{3.0.2a}$$

subject to the initial and boundary conditions

$$\mathbf{u}_h(0,\mathbf{x}) = \pi\mathbf{u}_0(\mathbf{x}) \text{ or } \mathbf{u}_h(0,\mathbf{x}) = \Pi\mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \omega, \tag{3.0.2b}$$

$$\begin{aligned}(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})\mathbf{u}_h^-(t,\mathbf{x}) = (\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})\pi_i\mathbf{u}(t,\mathbf{x}),& \\ \mathbf{x} \in \gamma_i, 1 \le i \le d,\ 0 < t < T,&\end{aligned} \tag{3.0.2c}$$

where $\mathbf{u} = \mathbf{u}_B$ on the boundary of Ω.

We will perform a local error analysis by writing the local error as a series and show that its leading term can be expressed as a linear combination of Legendre polynomials of degree p and $p+1$. We apply these asymptotic results to observe that projections of the error are pointwise superconvergent in some cases and establish superconvergence results for some integrals of the error. We further apply these asymptotic results and solve relatively small local problems to compute efficient and asymptotically exact estimates of the finite element error. Finally, we present some computational results for one- and two-dimensional systems.

3.1 Local Error Analysis

We will denote the dual of j as $j' = 3 - j$ for $j = 1, 2$.

Then we are ready to state and prove several preliminary lemmas.

Lemma 3.1.1. *Let $\mathbf{A}_1$ and $\mathbf{A}_2$ be symmetric matrices such that the $m \times (m-r)$ matrix $\mathbf{P}_{j,2}$, $j = 1, 2$, denotes the matrix of all $(m-r)$ orthogonal eigenvectors associated with the zero eigenvalue of $\mathbf{A}_j$. If the matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ satisfy either of the following assumptions*

i) $\mathbf{A}_j$ is invertible for $j = 1$ or $j = 2$,

ii) $\mathcal{N}(\mathbf{P}_{j,2}^t \mathbf{A}_{j'} \mathbf{P}_{j,2}) = \{\mathbf{0}\}$, for $j = 1$ or $j = 2$,

then $\mathcal{N}(\mathbf{A}_1) \bigcap \mathcal{N}(\mathbf{A}_2) = \{\mathbf{0}\}$.

Proof. If either $\mathbf{A}_1$ or $\mathbf{A}_2$ is invertible, the proof is straightforward. Now let us prove the lemma if one of the remaining conditions are satisfied. Without loss of generality, we assume $j = 1$ and let $\mathbf{q}$ be an arbitrary vector in $\mathcal{N}(\mathbf{A}_1) \bigcap \mathcal{N}(\mathbf{A}_2)$. Thus, $\mathbf{A}_1\mathbf{q} = \mathbf{0}$ and $\mathbf{A}_2\mathbf{q} = \mathbf{0}$ which are equivalent to

$$\mathbf{P}_1^t \mathbf{A}_1 \mathbf{P}_1 \mathbf{w} = \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{w} = \mathbf{0}, \text{ and } \mathbf{P}_1^t \mathbf{A}_2 \mathbf{P}_1 \mathbf{w} = \mathbf{0}, \tag{3.1.1}$$

where $\mathbf{w} = \mathbf{P}_1^t \mathbf{q}$ with $\mathbf{P}_1 = [\mathbf{P}_{1,1}, \mathbf{P}_{1,2}]$ containing the eigenvectors of $\mathbf{A}_1$.

Thus, if we split $\mathbf{w}$ as $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2)^t$, where $\mathbf{w}_1 \in \mathbb{R}^r$, then $\mathbf{\Lambda}\mathbf{w}_1 = \mathbf{0}$ yields $\mathbf{w}_1 = \mathbf{0}$.

On the other-hand $\mathbf{P}_1^t \mathbf{A}_2 \mathbf{P}_1 \mathbf{w} = \mathbf{0}$ yields $\mathbf{P}_{1,2}^t \mathbf{A}_2 \mathbf{P}_{1,2} \mathbf{w}_2 = \mathbf{0}$ which, in turns, leads to $\mathbf{w}_2 = \mathbf{0}$. Therefore, $\mathbf{w} = \mathbf{q} = \mathbf{0}$. □

Here we note that the converse of the previous lemma is not true.

Lemma 3.1.2. *Let* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *be symmetric matrices satisfying one of the assumptions of Lemma 3.1.1. If* $\mathbf{q} \in \mathcal{P}_p$ *satisfies*

$$\int_\Delta \mathbf{v}^t(\mathbf{A}_1\mathbf{q}_{,\xi_1} + \mathbf{A}_2\mathbf{q}_{,\xi_2})\,d\boldsymbol{\xi} - \int_\Gamma \mathbf{v}^t(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})\mathbf{q}\,d\boldsymbol{\sigma} = 0,\ \mathbf{v} \in \mathcal{P}_p, \tag{3.1.2}$$

then $\mathbf{q} = \mathbf{0}$ *on* Δ.

Proof. The proof follows the same line of reasoning for all possible cases. Here, we present the proof for assumption 2 of Lemma 3.1.1 with $j = 1$ only.

First we integrate (3.1.2) by parts to obtain

$$-\int_\Delta (\mathbf{v}_{,\xi_1}^t\mathbf{A}_1 + \mathbf{v}_{,\xi_2}^t\mathbf{A}_2)\mathbf{q}\,d\boldsymbol{\xi} + \int_\Gamma \mathbf{v}^t(\nu_1\mathbf{A}_1^{\mu_1} + \nu_2\mathbf{A}_2^{\mu_2})\mathbf{q}\,d\boldsymbol{\sigma} = 0. \tag{3.1.3}$$

Adding (3.1.2) to (3.1.3) and testing against $\mathbf{v} = \mathbf{q}$ we note that by the symmetry of $\mathbf{A}_1$ and $\mathbf{A}_2$, the double integrals on Δ cancel out. Thus, $\mathbf{q}$ satisfies

$$\begin{aligned}&\int_{\Gamma_1} \mathbf{q}^t\nu_1(\mathbf{A}_1^{\mu_1} - \mathbf{A}_1^{\bar{\mu}_1})\mathbf{q}\,d\boldsymbol{\sigma} + \int_{\Gamma_2} \mathbf{q}^t\nu_2(\mathbf{A}_2^{\mu_2} - \mathbf{A}_2^{\bar{\mu}_2})\mathbf{q}\,d\boldsymbol{\sigma}\\ &= \int_{\Gamma_1} \mathbf{q}^t(\mathbf{A}_1^{+} - \mathbf{A}_1^{-})\mathbf{q}\,d\boldsymbol{\sigma} + \int_{\Gamma_2} \mathbf{q}^t(\mathbf{A}_2^{+} - \mathbf{A}_2^{-})\mathbf{q}\,d\boldsymbol{\sigma} = 0.\end{aligned} \tag{3.1.4}$$

Since $(\mathbf{A}_i^+ - \mathbf{A}_i^-)$ is symmetric positive semi-definite it admits a Cholesky factorization $(\mathbf{A}_i^+ - \mathbf{A}_i^-) = \mathbf{L}_i^t\mathbf{L}_i$. Thus, (3.1.4) can be written as

$$\int_{\Gamma_1} \|\mathbf{L}_1\mathbf{q}\|^2\,d\boldsymbol{\sigma} + \int_{\Gamma_2} \|\mathbf{L}_2\mathbf{q}\|^2\,d\boldsymbol{\sigma} = 0. \tag{3.1.5}$$

Thus, $\mathbf{L}_i\mathbf{q} = \mathbf{0}$ on Γ_i which yields

$$\mathbf{L}_i^t(\mathbf{L}_i\mathbf{q}) = (\mathbf{A}_i^+ - \mathbf{A}_i^-)\mathbf{q} = \mathbf{0} \text{ on } \Gamma_i,\ 1 \le i \le d. \tag{3.1.6}$$

which combined with flux property (1.2.25b) leads to

$$\mathbf{A}_i^{\pm}\mathbf{q}|_{\Gamma_i} = \mathbf{0},\ i = 1, 2. \tag{3.1.7}$$

Testing against $\mathbf{v} = \mathbf{A}_1\mathbf{q}_{,\xi_1} + \mathbf{A}_2\mathbf{q}_{,\xi_2}$ in (3.1.2) and combining the resulting equation with (3.1.7) lead to

$$\int_\Delta \|\mathbf{A}_1\mathbf{q}_{\xi_1} + \mathbf{A}_2\mathbf{q}_{\xi_2}\|^2\,d\boldsymbol{\xi} = \mathbf{0}, \tag{3.1.8}$$

which in turn yields

$$\mathbf{A}_1\mathbf{q}_{\xi_1} + \mathbf{A}_2\mathbf{q}_{\xi_2} = \mathbf{0},\ \boldsymbol{\xi} \in \Delta. \tag{3.1.9}$$

Next, we let

$$\mathbf{B}_1 = \mathbf{P}_1^t \mathbf{A}_1 \mathbf{P}_1 = \begin{bmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \text{ and } \mathbf{B}_2 = \mathbf{P}_1^t \mathbf{A}_2 \mathbf{P}_1 = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}, \tag{3.1.10}$$

where $\mathbf{P}_1 = [\mathbf{P}_{1,1}, \mathbf{P}_{1,2}]$ and $\mathbf{B}_{ij} = \mathbf{P}_{1,i}^t \mathbf{A}_2 \mathbf{P}_{1,j},\ i,j = 1,2.$

Applying $\mathbf{A}_1\mathbf{q}|_{\Gamma_1} = \mathbf{0}$ we write

$$\mathbf{B}_1\mathbf{w}|_{\Gamma_1} = \mathbf{\Lambda}\mathbf{w}_1|_{\Gamma_1} = \mathbf{0}, \tag{3.1.11}$$

where $\mathbf{w} = \mathbf{P}^t\mathbf{q} = (\mathbf{w}_1, \mathbf{w}_2)^t$. This establishes $\mathbf{w}_1|_{\Gamma_1} = \mathbf{0}$.

On the other-hand $\mathbf{A}_2\mathbf{q}|_{\Gamma_2} = \mathbf{0}$ leads to

$$\mathbf{B}_2\mathbf{w}|_{\Gamma_2} = \mathbf{0}, \tag{3.1.12}$$

which can be split as

$$(\mathbf{B}_{11}\mathbf{w}_1 + \mathbf{B}_{12}\mathbf{w}_2)|_{\Gamma_2} = \mathbf{0}, \tag{3.1.13a}$$
$$(\mathbf{B}_{21}\mathbf{w}_1 + \mathbf{B}_{22}\mathbf{w}_2)|_{\Gamma_2} = \mathbf{0}. \tag{3.1.13b}$$

Pre-multiplying (3.1.9) by $\mathbf{P}^t$ yields

$$\mathbf{\Lambda}\mathbf{w}_{1,\xi_1} + \quad \mathbf{B}_{11}\mathbf{w}_{1,\xi_2} + \mathbf{B}_{12}\mathbf{w}_{2,\xi_2} \quad = \mathbf{0}, \tag{3.1.14a}$$
$$\mathbf{B}_{21}\mathbf{w}_{1,\xi_2} + \mathbf{B}_{22}\mathbf{w}_{2,\xi_2} \quad = \mathbf{0}, \quad \boldsymbol{\xi} \in \Delta. \tag{3.1.14b}$$

Since $\mathbf{B}_{22}$ is invertible, we can express $\mathbf{w}_{2,\xi_2}$ as

$$\mathbf{w}_{2,\xi_2} = -\mathbf{B}_{22}^{-1}\mathbf{B}_{21}\mathbf{w}_{1,\xi_2}, \tag{3.1.15a}$$

which leads to

$$\mathbf{\Lambda}\mathbf{w}_{1,\xi_1} + (\mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21})\mathbf{w}_{1,\xi_2} = \mathbf{0}, \quad \boldsymbol{\xi} \in \Delta. \tag{3.1.15b}$$

Combining the fact that $\mathbf{w}_1|_{\Gamma_1} = \mathbf{0}$ and (3.1.15b) we establish that

$$\frac{\partial^k \mathbf{w}_1}{\partial \xi_1^l \partial \xi_2^{k-l}} = \mathbf{0}, \quad \boldsymbol{\xi} \in \Delta,\ k \geq 0,\ l = 0, \cdots, k. \tag{3.1.16}$$

Since $\mathbf{w}_1$ is a polynomial, $\mathbf{w}_1 = \mathbf{0}$.

Combining (3.1.13b) and (3.1.15a) we establish that $\mathbf{w}_2|_\Delta = \mathbf{0}$. Thus, $\mathbf{q} = \mathbf{0},\ \boldsymbol{\xi} \in \Delta.$ which completes the proof of the lemma. □

Now we are ready to state the main theorem for the local spatial discretization error.

Theorem 3.1.3. *Let $\mathbf{A}_1$ and $\mathbf{A}_2$ be symmetric matrices satisfying the conditions of Lemma 3.1.1 and let $\mathbf{u} \in [C^2([0,T], C^{p+2}(\bar{\omega}))]^m$ and $\mathbf{u}_h \in \mathcal{P}_p$ be the solutions of (3.0.1) and (3.0.2), respectively, on $\omega = (0,h)^2$. Here $\mathbf{u}_h$ is computed by approximating the boundary conditions as*

$$\mathbf{u}_h^-(t,\mathbf{x}) = \pi_i \mathbf{u}(t,\mathbf{x}), \quad 0 < t < T,\ \mathbf{x} \in \gamma_i,\ 1 \le i \le d, \tag{3.1.17}$$

and using as initial conditions either $\pi\mathbf{u}_0$ or $\Pi\mathbf{u}_0$.
Then the local finite element error on ω, at $t = \mathcal{O}(1)$ and $p \ge 1$, can be written as

$$\mathbf{e}(t, h\boldsymbol{\xi}) = \mathbf{u}(t, h\boldsymbol{\xi}) - \mathbf{u}_h(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{j=1}^{2} (L_{p+1}(\xi_j)\mathbf{c}_j(t) - L_p(\xi_j)\mathbf{d}_j(t)) + \mathcal{O}(h^{p+2}), \tag{3.1.18a}$$

where

$$\mathbf{A}_i^+ \mathbf{c}_i(t) - \mathbf{A}_i^+ \mathbf{d}_i(t) = \mathbf{0}, \quad \mathbf{A}_i^- \mathbf{c}_i(t) + \mathbf{A}_i^- \mathbf{d}_i(t) = \mathbf{0},\ 1 \le i \le d. \tag{3.1.18b}$$

Proof. Subtracting the weak DG formulation (3.0.2) from (3.0.1), the local discretization error $\mathbf{e} = \mathbf{u} - \mathbf{u}_h$ satisfies the DG orthogonality condition

$$\int_\omega \mathbf{v}^t (\mathbf{e}_{,t} + \mathbf{A}_1 \mathbf{e}_{,x_1} + \mathbf{A}_2 \mathbf{e}_{,x_2})\, d\mathbf{x} + \int_{\partial\omega} \mathbf{v}^t (\nu_1 \mathbf{A}_1^{\bar{\mu}_1} + \nu_2 \mathbf{A}_2^{\bar{\mu}_2})(\mathbf{e}^- - \mathbf{e}^+)\, d\mathbf{s} = 0,\ \mathbf{v} \in \mathcal{P}_p. \tag{3.1.19}$$

Apply the scalings $\tau = \frac{t}{T}$ and $\boldsymbol{\xi} = \frac{\mathbf{x}}{h}$ to write $\hat{\mathbf{e}}(\tau, \boldsymbol{\xi}) = \mathbf{e}(T\tau, h\boldsymbol{\xi})$ to obtain

$$\int_\Delta \mathbf{v}^t (\frac{h}{T}\hat{\mathbf{e}}_{,\tau} + \mathbf{A}_1 \hat{\mathbf{e}}_{,\xi_1} + \mathbf{A}_2 \hat{\mathbf{e}}_{,\xi_2})\, d\boldsymbol{\xi} + \int_\Gamma \mathbf{v}^t (\nu_1 \mathbf{A}_1^{\bar{\mu}_1} + \nu_2 \mathbf{A}_2^{\bar{\mu}_2})(\hat{\mathbf{e}}^- - \hat{\mathbf{e}}^+)\, d\boldsymbol{\sigma} = 0,\ \mathbf{v} \in \mathcal{P}_p. \tag{3.1.20}$$

In order to investigate the asymptotic behavior of the local error, we start by writing the Maclaurin series of $\hat{\mathbf{e}}$ with respect to the mesh parameter h as

$$\hat{\mathbf{e}}(\tau, \boldsymbol{\xi}) = \sum_{k=0}^{p+1} h^k \mathbf{q}_k(\tau, \boldsymbol{\xi}) + \mathcal{O}(h^{p+2}), \tag{3.1.21}$$

where

$$\mathbf{q}_k(\tau, \boldsymbol{\xi}) = \frac{1}{k!} \frac{d^k (\mathbf{u} - \mathbf{u}_h)(\tau T, \boldsymbol{\xi} h, h)}{dh^k}\bigg|_{h=0}. \tag{3.1.22}$$

We used the fact that $\mathbf{u}_h(t, \mathbf{x}, h)$ is a function of t, $\mathbf{x}$ and h.

Next, from the definition of $\pi_i \mathbf{u}$ in Lemma 2.2.4, $\hat{\mathbf{e}}^-$ satisfies

$$\hat{\mathbf{e}}^-(\tau, \boldsymbol{\xi}) = (\mathbf{u} - \pi_1 \mathbf{u})(\tau, \xi) = \mathbf{r}_2(\tau, \xi_2) h^{p+1} + \mathcal{O}(h^{p+2}) \text{ for } \boldsymbol{\xi} \in \Gamma_1, \tag{3.1.23a}$$

$$\hat{\mathbf{e}}^-(\tau, \boldsymbol{\xi}) = (\mathbf{u} - \pi_2 \mathbf{u})(\tau, \xi) = \mathbf{r}_1(\tau, \xi_1) h^{p+1} + \mathcal{O}(h^{p+2}) \text{ for } \boldsymbol{\xi} \in \Gamma_2, \tag{3.1.23b}$$

where

$$\mathbf{r}_i(\tau, \xi_i) = L_{p+1}(\xi_i)\hat{\mathbf{c}}_i(\tau) - L_p(\xi_i)\hat{\mathbf{d}}_i(\tau), \tag{3.1.23c}$$

with, using property (1.2.25d),

$$\hat{\mathbf{c}}_i(\tau) = \mathbf{c}_i(t)\Big|_{t=\tau T} = \frac{1}{a_{p+1}}\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial x_i^{p+1}}\mathbf{u}(t,0)\Big|_{t=\tau T}, \tag{3.1.23d}$$

$$\mathbf{A}_i^+\hat{\mathbf{c}}_i(\tau) - \mathbf{A}_i^+\hat{\mathbf{d}}_i(\tau) = \mathbf{0}, \text{ and } \mathbf{A}_i^-\hat{\mathbf{c}}_i(\tau) + \mathbf{A}_i^-\hat{\mathbf{d}}_i(\tau) = \mathbf{0}. \tag{3.1.23e}$$

Substituting (3.1.21), (3.1.23a) and (3.1.23b) in (3.1.20) yields

$$\begin{aligned}\sum_{k=0}^{p+1} h^k \int_\Delta \mathbf{v}^t(\frac{h}{T}\mathbf{q}_{k,\tau} + \mathbf{A}_1\mathbf{q}_{k,\xi_1} + \mathbf{A}_2\mathbf{q}_{k,\xi_2})\,d\boldsymbol{\xi} - \int_\Gamma \mathbf{v}^t(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})\mathbf{q}_k\,d\boldsymbol{\sigma}\\ = \quad -h^{p+1}\left(\int_{\Gamma_1}\mathbf{v}^t\nu_1\mathbf{A}_1^{\bar{\mu}_1}\mathbf{r}_2\,d\boldsymbol{\sigma} + \int_{\Gamma_2}\mathbf{v}^t\nu_2\mathbf{A}_2^{\bar{\mu}_2}\mathbf{r}_1\,d\boldsymbol{\sigma}\right) + \mathcal{O}(h^{p+2}).\end{aligned} \tag{3.1.24}$$

Now we will use induction to prove that $\mathbf{q}_k = \mathbf{0},\ k = 0, 1, \cdots, p$ by first assuming that $T = \mathcal{O}(1)$ and setting to zero all terms having the same power of h. Thus, the $\mathcal{O}(1)$ term $\mathbf{q}_0$ satisfies the orthogonality condition

$$\int_\Delta \mathbf{v}^t(\mathbf{A}_1\mathbf{q}_{0,\xi_1} + \mathbf{A}_2\mathbf{q}_{0,\xi_2})\,d\boldsymbol{\xi} - \int_\Gamma \mathbf{v}^t(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})\mathbf{q}_0\,d\boldsymbol{\sigma} = 0,\ \mathbf{v} \in \mathcal{P}_p. \tag{3.1.25}$$

By Lemma 3.1.2, $\mathbf{q}_0 = \mathbf{0}$.

Now assume that $\mathbf{q}_j = \mathbf{0},\ j = 0, 1, \cdots, k-1 < p$. Thus, the $\mathcal{O}(h^k)$ term is written as

$$\int_\Delta \mathbf{v}^t(\mathbf{A}_1\mathbf{q}_{k,\xi_1} + \mathbf{A}_2\mathbf{q}_{k,\xi_2})\,d\boldsymbol{\xi} - \int_\Gamma \mathbf{v}^t(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})\mathbf{q}_k\,d\boldsymbol{\sigma} = 0,\ \mathbf{v} \in \mathcal{P}_p. \tag{3.1.26}$$

By Lemma 3.1.2, $\mathbf{q}_k = \mathbf{0}$ on Δ. Thus, by induction, $\mathbf{q}_k = \mathbf{0},\ 0 \le k \le p$.

The $\mathcal{O}(h^{p+1})$ term satisfies the orthogonality condition

$$\begin{aligned}&\int_\Delta \mathbf{v}^t(\mathbf{A}_1\mathbf{q}_{p+1,\xi_1} + \mathbf{A}_2\mathbf{q}_{p+1,\xi_2})\,d\boldsymbol{\xi}\\ &\quad = \int_{\Gamma_1}\mathbf{v}^t\nu_1\mathbf{A}_1^{\bar{\mu}_1}(\mathbf{q}_{p+1} - \mathbf{r}_2)\,d\boldsymbol{\sigma} + \int_{\Gamma_2}\mathbf{v}^t\nu_2\mathbf{A}_2^{\bar{\mu}_2}(\mathbf{q}_{p+1} - \mathbf{r}_1)\,d\boldsymbol{\sigma},\end{aligned} \tag{3.1.27}$$

By equation (3.1.21),

$$\begin{aligned}\mathbf{q}_{p+1}(\tau,\boldsymbol{\xi}) &= \frac{1}{(p+1)!}\frac{d^{p+1}(\mathbf{u}-\mathbf{u}_h)(\tau T, \boldsymbol{\xi}h)}{dh^{p+1}}\Big|_{h=0}\\ &= \sum_{|\alpha|=p+1}\frac{1}{\alpha!}D^\alpha(\mathbf{u}-\mathbf{u}_h)(\tau T,0)\boldsymbol{\xi}^\alpha\\ &= \sum_{i=1}^{2}\frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial x_i^{p+1}}(\mathbf{u}-\mathbf{u}_h)(\tau T,0)\xi_i^{p+1} + \mathbf{p}_1(\tau,\boldsymbol{\xi}),\end{aligned} \tag{3.1.28}$$

where, for a fixed τ, $\mathbf{p}_1(\tau,\boldsymbol{\xi}) \in \mathcal{P}_p$. By equation (3.1.23c)

$$\begin{aligned}\sum_{i=1}^{2} \mathbf{r}_i(\tau,\xi_i) &= \sum_{i=1}^{2}(L_{p+1}(\xi_i)\hat{\mathbf{c}}_i - L_p(\xi_i)\hat{\mathbf{d}}_i) \\ &= \sum_{i=1}^{2} \frac{1}{(p+1)!}\frac{\partial^{p+1}}{\partial x_i^{p+1}}\mathbf{u}(\tau T,0)\xi_i^{p+1} + \mathbf{p}_2(\tau,\boldsymbol{\xi}), \end{aligned} \tag{3.1.29}$$

where $\mathbf{p}_2 \in \mathcal{P}_p$. We further note that since $\mathbf{u}_h \in \mathcal{P}_p$, $\frac{\partial^{p+1}\mathbf{u}_h}{\partial x_i^{p+1}} = \mathbf{0}$, which in turn leads to

$$\mathbf{q}_{p+1}(\tau,\boldsymbol{\xi}) - \mathbf{r}_1(\tau,\xi_1) - \mathbf{r}_2(\tau,\xi_2) = \mathbf{p}_1(\tau,\boldsymbol{\xi}) - \mathbf{p}_2(\tau,\boldsymbol{\xi}) = \mathbf{p}(\tau,\boldsymbol{\xi}) \ \in \mathcal{P}_p. \tag{3.1.30}$$

Noting that $\mathbf{r}_2$ is independent of ξ_1, $\mathbf{r}_1$ is independent of ξ_2, solving (3.1.30) for $\mathbf{q}_{p+1}$ and substituting into (3.1.27) yields

$$\begin{aligned}&\int_\Delta \mathbf{v}^t(\mathbf{A}_1(\mathbf{p}+\mathbf{r}_1)_{,\xi_1} + \mathbf{A}_2(\mathbf{p}+\mathbf{r}_2)_{,\xi_2})\,d\boldsymbol{\xi} \\ &= \int_{\Gamma_1} \mathbf{v}^t\nu_1\mathbf{A}_1^{\bar{\mu}_1}(\mathbf{p}+\mathbf{r}_1)\,d\boldsymbol{\sigma} + \int_{\Gamma_2} \mathbf{v}^t\nu_2\mathbf{A}_2^{\bar{\mu}_2}(\mathbf{p}+\mathbf{r}_2)\,d\boldsymbol{\sigma},\ \mathbf{v}\in\mathcal{P}_p. \end{aligned} \tag{3.1.31}$$

Integrating $\int_\Delta \mathbf{v}^t(\mathbf{A}_1\mathbf{r}_{1,\xi_1} + \mathbf{A}_2\mathbf{r}_{2,\xi_2})\,d\boldsymbol{\xi}$ by parts, (3.1.31) becomes

$$\begin{aligned}&\int_\Delta \mathbf{v}^t(\mathbf{A}_1\mathbf{p}_{,\xi_1} + \mathbf{A}_2\mathbf{p}_{,\xi_2})\,d\boldsymbol{\xi} - \int_\Delta \mathbf{v}^t_{,\xi_1}\mathbf{A}_1\mathbf{r}_1 + \mathbf{v}^t_{,\xi_2}\mathbf{A}_2\mathbf{r}_2\,d\boldsymbol{\xi} \\ &= \int_{\Gamma_1} \mathbf{v}^t\nu_1(\mathbf{A}_1^{\bar{\mu}_1}\mathbf{p} - \mathbf{A}_1^{\mu_1}\mathbf{r}_1)\,d\boldsymbol{\sigma} + \int_{\Gamma_2} \mathbf{v}^t\nu_2(\mathbf{A}_2^{\bar{\mu}_2}\mathbf{p} - \mathbf{A}_2^{\mu_2}\mathbf{r}_2)\,d\boldsymbol{\sigma},\ \mathbf{v}\in\mathcal{P}_p. \end{aligned} \tag{3.1.32}$$

Since $\mathbf{r}_i(\tau,\xi_i) = L_{p+1}(\xi_i)\hat{\mathbf{c}}_i(\tau) - L_p(\xi_i)\hat{\mathbf{d}}_i(\tau)$, by the orthogonality of Legendre polynomials we have

$$\int_\Delta \mathbf{v}^t_{,\xi_1}\mathbf{A}_1\mathbf{r}_1 + \mathbf{v}^t_{,\xi_2}\mathbf{A}_2\mathbf{r}_2\,d\boldsymbol{\xi} = 0,\ \mathbf{v}\in\mathcal{P}_p. \tag{3.1.33}$$

Using (3.1.23c), (3.1.23d), $L_p(0) = (-1)^p$ and $L_p(1) = 1$, we further show that

$$\mathbf{A}_i^+\mathbf{r}_i(\tau,1) = \mathbf{A}_i^+(L_{p+1}(1)\hat{\mathbf{c}}_i(\tau) - L_p(1)\hat{\mathbf{d}}_i(\tau)) = \mathbf{A}_i^+\hat{\mathbf{c}}_i(\tau) - \mathbf{A}_i^+\hat{\mathbf{d}}_i(\tau) = \mathbf{0}, \tag{3.1.34a}$$

$$\mathbf{A}_i^-\mathbf{r}_i(\tau,0) = \mathbf{A}_i^-(L_{p+1}(0)\hat{\mathbf{c}}_i(\tau) - L_p(0)\hat{\mathbf{d}}_i(\tau)) = (-1)^p(\mathbf{A}_i^-\hat{\mathbf{c}}_i(\tau) + \mathbf{A}_i^-\hat{\mathbf{d}}_i(\tau)) = \mathbf{0}. \tag{3.1.34b}$$

Thus, we have established that

$$\mathbf{A}_i^{\mu_i}\mathbf{r}_i\Big|_{\Gamma_i} = \mathbf{0},\ 1\le i\le d. \tag{3.1.35}$$

Combining (3.1.33) and (3.1.35) with the orthogonality condition (3.1.32) leads to

$$\int_\Delta \mathbf{v}^t(\mathbf{A}_1\mathbf{p}_{,\xi_1} + \mathbf{A}_2\mathbf{p}_{,\xi_2})\,d\boldsymbol{\xi} = \int_\Gamma \mathbf{v}^t(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})\mathbf{p}\,d\boldsymbol{\sigma},\ \mathbf{v}\in\mathcal{P}_p. \tag{3.1.36}$$

By Lemma 3.1.2, $\mathbf{p} = \mathbf{0}$ on Δ, which, when combined with (3.1.30), yields

$$\mathbf{q}_{p+1}(\tau, \xi) = \mathbf{r}_1(\tau, \xi_1) + \mathbf{r}_2(\tau, \xi_2). \tag{3.1.37}$$

This completes the proof. □

3.2 Superconvergence and *A Posteriori* Error Analysis

In this section we investigate pointwise superconvergence for DG solutions and describe procedures to compute asymptotically correct *a posteriori* DG error estimates under mesh refinement.

3.2.1 Superconvergence

In order for the DG solution $\mathbf{u}_h$ to be $\mathcal{O}(h^{p+2})$-superconvergent at few points in element ω, the leading error term shown in Theorem 3.1.3 has to be zero at these points. This pointwise superconvergence happens only for special hyperbolic problems as shown in the following theorem.

Theorem 3.2.1. *Under the conditions of Theorem 3.1.3 with $p \geq 1$ we let $\bar{\xi}^s_j,\ j = 1, \cdots, p+1$, denote the roots of $R^s_{p+1}(\xi),\ s = +, -$ shifted to $[0, 1]$. Thus,*

i) If $\mathbf{z}$ is a unit vector in the union of the spaces $\mathcal{R}(\mathbf{A}^s_1) \bigcap \mathcal{R}(\mathbf{A}^\sigma_2),\ s, \sigma = +, -$, then the projection $\mathbf{z}^t\mathbf{e}(t, \mathbf{x})$ of the local error onto span$\{\mathbf{z}\}$ is $\mathcal{O}(h^{p+2})$ superconvergent at the points $(t, \underline{x}) = (t, h\bar{\xi}^s_k, h\bar{\xi}^\sigma_l),\ 1 \leq k, l \leq p+1,\ t = \mathcal{O}(1)$, i.e.,

$$\mathbf{z}^t\mathbf{e}(t, h\bar{\xi}^s_k, h\bar{\xi}^\sigma_l) = \mathcal{O}(h^{p+2}). \tag{3.2.1}$$

ii) Moreover, if $\gamma_i(a) = \{\mathbf{x} \in (0, h)^2 : x_i = a\},\ 0 \leq a \leq h$ and if $\mathbf{v} \in \mathcal{P}_{p-1}$ is a unit vector with respect to the C^∞ norm, then, at $a = h\bar{\xi}^s_k,\ 1 \leq k \leq p+1$, we have the superconvergence of the following error averages

$$\frac{1}{h} \int_{\gamma_i(h\bar{\xi}^s_l)} \mathbf{v}^t \mathbf{A}^s_i \mathbf{e}\, d\mathbf{s} = \mathcal{O}(h^{p+2}),\ i = 1, 2,\ s = +, -, \tag{3.2.2}$$

and

$$\frac{1}{h} \int_{\gamma^s_i} \mathbf{v}^t (\mathbf{A}^{\mu_i}_i \mathbf{e} + \mathbf{A}^{\bar{\mu}_i}_i \mathbf{e}^-)\, d\mathbf{s} = \mathcal{O}(h^{p+2}),\ i = 1, 2,\ s = +, -. \tag{3.2.3}$$

Proof. We prove (3.2.1) by assuming that there exists a unit vector $\mathbf{z} \in \mathcal{R}(\mathbf{A}_i^+)$ for $i = 1$ and $i = 2$, *i.e.*, there exists $\mathbf{v}_i$ such that $\mathbf{A}_i^+ \mathbf{v}_i = \mathbf{z},\ i = 1, 2$.

Left pre-multiplying (3.1.18a) by $\mathbf{z}^t$ and evaluating the resulting function at the points $(t, h\bar{\xi}_k^+, h\bar{\xi}_l^+)$, $1 \le k, l \le p+1$, we obtain

$$\begin{aligned} \mathbf{z}^t \mathbf{e}(t, h\bar{\xi}_k^+, h\bar{\xi}_l^+) &= h^{p+1}(L_{p+1}(\bar{\xi}_k^+)\mathbf{z}^t\mathbf{c}_1 - L_p(\bar{\xi}_k^+)\mathbf{z}^t\mathbf{d}_1) \\ &+ h^{p+1}(L_{p+1}(\bar{\xi}_l^+)\mathbf{z}^t\mathbf{c}_2 - L_p(\bar{\xi}_l^+)\mathbf{z}^t\mathbf{d}_2) + \mathcal{O}(h^{p+2}). \end{aligned} \tag{3.2.4}$$

Applying (3.1.18b) yields

$$\mathbf{z}^t\mathbf{d}_i = \mathbf{v}_i^t\mathbf{A}_i^+\mathbf{d}_i = \mathbf{v}_i^t\mathbf{A}_i^+\mathbf{c}_i = \mathbf{z}^t\mathbf{c}_i. \tag{3.2.5}$$

Combining (3.2.4) and (3.2.5) we prove that

$$\mathbf{z}^t \mathbf{e}(t, h\bar{\xi}_k^+, h\bar{\xi}_l^+) = \mathcal{O}(h^{p+2}). \tag{3.2.6}$$

Following the same line of reasoning we establish (3.2.1) for all other cases.

Let $\mathbf{v} \in \mathcal{P}_{p-1}$ be a unit vector in $[C^\infty]^m$ norm and apply the orthogonality of Legendre polynomials and the relations (3.1.18a) to obtain

$$\begin{aligned} \frac{1}{h}\int_{\gamma_1(h\bar{\xi}_k^+)} \mathbf{v}^t\mathbf{A}_1^+\mathbf{e}\, d\mathbf{s} &= \frac{1}{h}\int_0^h \mathbf{v}^t(h\bar{\xi}_k^+, x_2)\mathbf{A}_1^+\mathbf{e}(t, h\bar{\xi}_k^+, x_2)\, dx_2 \\ &= \int_0^1 \mathbf{v}^t(h\bar{\xi}_k^+, h\xi_2)\mathbf{A}_1^+\mathbf{e}(t, h\bar{\xi}_k^+, h\xi_2)\, d\xi_2 \\ &= h^{p+1}\int_0^1 \mathbf{v}^t\mathbf{A}_1^+\left(L_{p+1}(\bar{\xi}_k^+)\mathbf{c}_1 - L_p(\bar{\xi}_k^+)\mathbf{d}_1 + L_{p+1}(\xi_2)\mathbf{c}_2 - L_p(\xi_2)\mathbf{d}_2\right)\, d\xi_2 + \mathcal{O}(h^{p+2}) \\ &= h^{p+1}\int_0^1 \mathbf{v}^t\left(L_{p+1}(\bar{\xi}_k^+) - L_p(\bar{\xi}_k^+)\right)\mathbf{A}_1^+\mathbf{d}_1\, d\xi_2 + \mathcal{O}(h^{p+2}) \\ &= \mathcal{O}(h^{p+2}). \end{aligned} \tag{3.2.7}$$

The estimate (3.2.2) holds on $\gamma_i(h\bar{\xi}_k^\pm),\ i = 1, 2$ for $\mathbf{A}_i^\pm$.

By virtue of (3.1.17) we note that on γ_1^+

$$\mathbf{e}^-(t, \mathbf{x}) = \mathbf{u}(t, \mathbf{x}) - \pi_1\mathbf{u}(t, \mathbf{x}) = h^{p+1}[L_{p+1}(\frac{x_2}{h})\mathbf{c}_2 + L_p(\frac{x_2}{h})\mathbf{d}_2] + \mathcal{O}(h^{p+2}). \tag{3.2.8}$$

Since 1 and 0, respectively, are shifted roots of R_{p+1}^+ and R_{p+1}^-, applying (3.2.2) we will prove $\mathcal{O}(h^{p+2})$ superconvergence of the flux for $i = 1$.

$$\begin{aligned} &\frac{1}{h}\int_{\gamma_1^+} \mathbf{v}^t(\mathbf{A}_1^+\mathbf{e} + \mathbf{A}_1^-\mathbf{e}^-)\, d\mathbf{s} \\ =&\frac{1}{h}\left[\int_{\gamma_1^+} \mathbf{v}^t\mathbf{A}_1^+\mathbf{e}\, d\mathbf{s} + h^{p+1}\int_{\gamma_1^+} \mathbf{v}^t[L_{p+1}(\frac{x_2}{h})\mathbf{A}_1^-\mathbf{c}_2 + L_p(\frac{x_2}{h})\mathbf{A}_1^-\mathbf{d}_2]ds\right] + \mathcal{O}(h^{p+2}). \end{aligned} \tag{3.2.9}$$

Now, combining (3.2.9) and (3.2.2) yields (3.2.3). We presented the proof for $i = 1$ with $a = h$. Other cases can be treated using the same line of reasoning and are omitted. □

3.2.2 *A Posteriori* Error Estimation

In this section we present an *a posteriori* error estimation procedure which consists of computing asymptotically exact local and global error estimates of the DG error. In Theorem 3.1.3 we showed that the local discretization error for the DG method on a physical element $\omega = (0, h)^2$ can be written as

$$\mathbf{e}(t, h\boldsymbol{\xi}) \;=\; \mathbf{u}(t, h\boldsymbol{\xi}) - \mathbf{u}_h(t, h\boldsymbol{\xi}) \;=\; h^{p+1} \sum_{j=1}^{2} (L_{p+1}(\xi_j)\mathbf{c}_j(t) - L_p(\xi_j)\mathbf{d}_j(t)) + \mathcal{O}(h^{p+2}), \quad (3.2.10a)$$

where $\mathbf{c}_i(t)$, $\mathbf{d}_i(t)$, $i = 1, 2$ satisfy

$$\mathbf{A}_i^+ \mathbf{c}_i(t) - \mathbf{A}_i^+ \mathbf{d}_i(t) = \mathbf{0}, \quad \mathbf{A}_i^- \mathbf{c}_i(t) + \mathbf{A}_i^- \mathbf{d}_i(t) = \mathbf{0}, \; i = 1, 2. \quad (3.2.10b)$$

In this following we consider problems where the matrices $\mathbf{A}_i$, $i = 1, 2$ may be singular and satisfy assumptions of Lemma 3.1.1.

Let us subtract equations (3.2.10b) and solve for $\mathbf{d}_i$ in terms of $\mathbf{c}_i$, using Lemma 1.2.13 and property (1.2.25d), to write

$$\mathbf{d}_i = \mathbf{A}_i^\dagger (\mathbf{A}_i^+ - \mathbf{A}_i^-)\mathbf{c}_i + \mathbf{d}_i^{\boldsymbol{\Phi}} = (\mathbf{A}_i^\dagger \mathbf{A}_i)\mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i^\perp + \mathbf{d}_i^{\boldsymbol{\Phi}} = \mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i^\perp + \mathbf{d}_i^{\boldsymbol{\Phi}}. \quad (3.2.11a)$$

Moreover, from the direct sum $\mathbb{R}^m = \mathcal{N}(\mathbf{A}_i) \bigoplus \mathcal{N}(\mathbf{A}_i)^\perp$ we split $\mathbf{c}_i$ as

$$\mathbf{c}_i = \mathbf{c}_i^\perp + \mathbf{c}_i^{\boldsymbol{\Phi}}, \quad (3.2.11b)$$

where $\mathbf{c}_i^{\boldsymbol{\Phi}}, \; \mathbf{d}_i^{\boldsymbol{\Phi}} \in \mathcal{N}(\mathbf{A}_i)$ and $\mathbf{c}_i^\perp, \; \mathbf{d}_i^\perp \in \mathcal{N}(\mathbf{A}_i)^\perp$.

Hence, the leading term of the spatial discretization error (3.2.10a) can be split into two parts as

$$\mathbf{e} = \mathbf{e}^\perp + \mathbf{e}^{\boldsymbol{\Phi}} + \mathcal{O}(h^{p+2}), \quad (3.2.12a)$$

where

$$\mathbf{e}^\perp(t, h\boldsymbol{\xi}) \;=\; h^{p+1} \sum_{j=1}^{2} \left[L_{p+1}(\xi_j) - L_p(\xi_j)\mathrm{sgn}(\mathbf{A}_j) \right] \mathbf{c}_j^\perp(t), \quad (3.2.12b)$$

$$\mathbf{e}^{\boldsymbol{\Phi}}(t, h\boldsymbol{\xi}) \;=\; h^{p+1} \sum_{j=1}^{2} (L_{p+1}(\xi_j)\mathbf{c}_j^{\boldsymbol{\Phi}}(t) - L_p(\xi_j)\mathbf{d}_j^{\boldsymbol{\Phi}}(t)). \quad (3.2.12c)$$

We note that for invertible matrices $\mathbf{A}_i, \; i = 1, 2$, the error component $\mathbf{e}^{\boldsymbol{\Phi}}(t, \mathbf{x}) = \mathbf{0}$.

Next, we develop an *a posteriori* error estimation procedure for estimating both $\mathbf{e}^{\perp}$ and $\mathbf{e}^{\Psi}$ (if needed). We end the section by proving that, for smooth solutions, our local error estimates converge to the true error under mesh refinement. Up to this point we are not able to prove the asymptotic exactness of our global *a posteriori* error estimates. However, computational results for several hyperbolic systems shown in § 3.3 suggest that our global *a posteriori* error estimates are asymptotically exact under mesh refinement for smooth solutions.

The *a posteriori* error estimation procedure to compute estimates for $\mathbf{e}^{\perp}$ consists of determining

$$\mathbf{E}^{\perp}(t, h\boldsymbol{\xi}) = \sum_{j=1}^{2} \left[L_{p+1}(\xi_j) - L_p(\xi_j)\mathrm{sgn}(\mathbf{A}_j)\right] \boldsymbol{\gamma}_j^{\perp}(t)),\ \boldsymbol{\gamma}_j^{\perp} \in \mathcal{N}(\mathbf{A}_j)^{\perp}, \tag{3.2.13a}$$

such that

$$\int_{\omega} L_p\left(\frac{x_i}{h}\right) \mathbf{w}^t \left[\mathbf{u}_{h,t} + \mathbf{A}_1(\mathbf{u}_h + \mathbf{E}^{\perp})_{,x_1} + \mathbf{A}_2(\mathbf{u}_h + \mathbf{E}^{\perp})_{,x_2} - \mathbf{g}\right] d\mathbf{x} = 0, \tag{3.2.13b}$$
$$\mathbf{w} \in \mathcal{N}(\mathbf{A}_i)^{\perp},\ i = 1,\ 2.$$

By Lemma 1.2.13, $\mathsf{P}_i = \mathbf{A}_i^{\dagger}\mathbf{A}_i$ is symmetric and projects any vector in $\mathbb{R}^m$ into $\mathcal{N}(\mathbf{A}_i)^{\perp}$. Therefore, the columns of P_i span $\mathcal{N}(\mathbf{A}_i)^{\perp}$. Testing against all columns of P_i, we can replace (3.2.13b) by the system

$$\int_{\omega} L_p\left(\frac{x_i}{h}\right) \mathsf{P}_i \left[\mathbf{u}_{h,t} + \mathbf{A}_1(\mathbf{u}_h + \mathbf{E}^{\perp})_{,x_1} + \mathbf{A}_2(\mathbf{u}_h + \mathbf{E}^{\perp})_{,x_2} - \mathbf{g}\right] d\mathbf{x} = \mathbf{0},\ i = 1,\ 2, \tag{3.2.14}$$

which can be reduced to

$$\int_{\omega} L_p\left(\frac{x_i}{h}\right) L'_{p+1}\left(\frac{x_i}{h}\right) d\mathbf{x}\ \mathbf{A}_i\boldsymbol{\gamma}_i^{\perp} = \mathfrak{r}_{p,i}^{\perp}, \tag{3.2.15a}$$

where $\mathfrak{r}_{p,i}^{\perp}$ is the projection of the residual defined as

$$\mathfrak{r}_{p,i}^{\perp} = \int_{\omega} L_p\left(\frac{x_i}{h}\right) \mathsf{P}_i(\mathbf{g} - \mathbf{u}_{h,t} - \mathbf{A}_1\mathbf{u}_{h,x_1} - \mathbf{A}_2\mathbf{u}_{h,x_2})\, d\mathbf{x}. \tag{3.2.15b}$$

This can be reduced further to obtain

$$2h\ \mathbf{A}_i\boldsymbol{\gamma}_i^{\perp} = \mathfrak{r}_{p,i}^{\perp},\ i = 1, 2. \tag{3.2.16}$$

Since $\mathfrak{r}_{p,i}^{\perp} \in \mathcal{N}(\mathbf{A}_i)^{\perp}$, we can solve (3.2.16) to find the unique solution $\boldsymbol{\gamma}_i^{\perp} \in \mathcal{N}(\mathbf{A}_i)^{\perp}$,

$$\boldsymbol{\gamma}_i^{\perp} = \frac{1}{2h}\mathbf{A}_i^{\dagger}\mathfrak{r}_{p,i}^{\perp},\ i = 1, 2. \tag{3.2.17}$$

Now we turn to estimating the error component $\mathbf{e}^{\mathbf{\Phi}}$ lying in the polynomial space

$$\mathcal{E}_p = \left\{ \mathbf{v}(\mathbf{x}) = \sum_{i=1}^{2} \left(L_{p+1}\left(\frac{x_i}{h}\right) \mathbf{a}_i - L_p\left(\frac{x_i}{h}\right) \mathbf{b}_i \right) : \ \mathbf{a}_i, \mathbf{b}_i \in \mathcal{N}(\mathbf{A}_i) \right\}, \tag{3.2.18}$$

which contains nonzero elements only if at least $\mathbf{A}_1$ or $\mathbf{A}_2$ is singular.

By Lemma 2.2.4, $\mathbf{u} - \mathbf{u}_h^-$ on $\partial\omega$ and $(\mathbf{u}_0 - \mathbf{u}_h)(0, \mathbf{x})$ on ω satisfy

$$(\mathbf{u} - \mathbf{u}_h^-)(t, h\boldsymbol{\xi}) = h^{p+1}(L_{p+1}(\xi_i)\mathbf{c}_i(t) - L_p(\xi_i)\mathbf{d}_i(t)) + \mathcal{O}(h^{p+2}), \ \mathbf{x} \in \gamma_{i'}, \ i = 1, 2, \tag{3.2.19}$$

$$(\mathbf{u}_0 - \mathbf{u}_h)(0, h\boldsymbol{\xi}) = \sum_{j=1}^{2} h^{p+1}(L_{p+1}(\xi_j)\mathbf{c}_j(0) - L_p(\xi_j)\mathbf{d}_j(0)) + \mathcal{O}(h^{p+2}), \ \mathbf{x} \in \omega, \tag{3.2.20}$$

where $\mathbf{c}_i(t)$, $\mathbf{d}_i(t)$, $i = 1, 2$, satisfy

$$\mathbf{A}_i^+ \mathbf{c}_i(t) - \mathbf{A}_i^+ \mathbf{d}_i(t) = \mathbf{0}, \quad \mathbf{A}_i^- \mathbf{c}_i(t) + \mathbf{A}_i^- \mathbf{d}_i(t) = \mathbf{0}, \ i = 1, 2. \tag{3.2.21}$$

Therefore, we can define $\mathbf{E}^{\mathbf{\Phi}}(0, \mathbf{x})$ on ω and $\mathbf{E}^-(t, \mathbf{x})$ on $\partial\omega$ by

$$\mathbf{E}^{\mathbf{\Phi}}(0, \mathbf{x}) = \mathbf{e}^{\mathbf{\Phi}}(0, \mathbf{x}), \quad \mathbf{x} \in \omega, \tag{3.2.22a}$$

and

$$\mathbf{E}^-(t, \mathbf{x}) = h^{p+1}(L_{p+1}(\xi_i)\mathbf{c}_i(t) - L_p(\xi_i)\mathbf{d}_i(t)), \quad \mathbf{x} \in \gamma_{i'}, \ i = 1, 2, \tag{3.2.22b}$$

where i' denotes the dual of i.

Now let us approximate $\mathbf{e}^{\mathbf{\Phi}}$ by determining

$$\mathbf{E}^{\mathbf{\Phi}}(t, h\boldsymbol{\xi}) = \sum_{j=1}^{2} L_{p+1}(\xi_j)\boldsymbol{\gamma}_j^{\mathbf{\Phi}} - L_p(\xi_j)\boldsymbol{\delta}_j^{\mathbf{\Phi}}, \tag{3.2.23a}$$

such that

$$\begin{aligned} &\int_\omega \mathbf{v}^t \left[(\mathbf{u}_h + \mathbf{E}^{\mathbf{\Phi}})_{,t} + \mathbf{A}_1 \mathbf{u}_{h,x_1} + \mathbf{A}_2 \mathbf{u}_{h,x_2} - \mathbf{g} \right] d\mathbf{x} \\ &+ \int_{\partial\omega} \mathbf{v}^t (\nu_1 \mathbf{A}_1^{\bar{\mu}_1} + \nu_2 \mathbf{A}_2^{\bar{\mu}_2})(\mathbf{u}_h^- + \mathbf{E}^- - \mathbf{u}_h - \mathbf{E}^\perp - \mathbf{E}^{\mathbf{\Phi}}) \, d\mathbf{s} = 0, \ \forall \ \mathbf{v} \in \mathcal{E}_p. \end{aligned} \tag{3.2.23b}$$

By Lemma 1.2.14, $(\mathbf{I} - \mathsf{P}_i)$ is the projection into $\mathcal{N}(\mathbf{A}_i)$. Hence, the columns of $L_{p+1}(\xi_i)(\mathbf{I} - \mathsf{P}_i)$ and $L_p(\xi_i)(\mathbf{I} - \mathsf{P}_i)$ span $\mathcal{E}_p$.

Applying the projection $(\mathbf{I} - \mathsf{P}_i)$ to (3.2.23b) yields the following system

$$\begin{aligned} &\int_\omega L_m\left(\frac{x_i}{h}\right) (\mathbf{I} - \mathsf{P}_i) \left[(\mathbf{u}_h + \mathbf{E}^{\mathbf{\Phi}})_{,t} + \mathbf{A}_1 \mathbf{u}_{h,x_1} + \mathbf{A}_2 \mathbf{u}_{h,x_2} - \mathbf{g} \right] d\mathbf{x} \\ &+ \int_{\partial\omega} L_m\left(\frac{x_i}{h}\right) (\mathbf{I} - \mathsf{P}_i)(\nu_1 \mathbf{A}_1^{\bar{\mu}_1} + \nu_2 \mathbf{A}_2^{\bar{\mu}_2})(\mathbf{u}_h^- + \mathbf{E}^- - \mathbf{u}_h - \mathbf{E}^\perp - \mathbf{E}^{\mathbf{\Phi}}) \, d\mathbf{s} \\ &= 0, \quad m = p, p+1, \ i = 1, 2. \end{aligned} \tag{3.2.24}$$

Equation (1.2.31) yields $(\mathbf{I}-\mathsf{P}_i)\mathbf{A}_i^{\bar{\mu}_i}=\mathbf{0}$, which can be used to write (3.2.24) as

$$\int_\omega L_m\left(\frac{x_i}{h}\right)(\mathbf{I}-\mathsf{P}_i)\mathbf{E}_{,t}^{\mathbf{\Phi}}\,d\mathbf{x}-\int_{\gamma_{i'}} L_m\left(\frac{x_i}{h}\right)(\mathbf{I}-\mathsf{P}_i)(\nu_{i'}\mathbf{A}_{i'}^{\bar{\mu}_{i'}})\mathbf{E}^{\mathbf{\Phi}}\,d\mathbf{s}=\mathfrak{r}_{m,i}^{\mathbf{\Phi}}, \tag{3.2.25a}$$

where $\mathfrak{r}_{m,i}^{\mathbf{\Phi}}$ is the projection of the residual given by

$$\begin{aligned}\mathfrak{r}_{m,i}^{\mathbf{\Phi}} &= -\int_\omega L_m\left(\frac{x_i}{h}\right)(\mathbf{I}-\mathsf{P}_i)\left(\mathbf{u}_{h,t}+\mathbf{A}_1\mathbf{u}_{h,x_1}+\mathbf{A}_2\mathbf{u}_{h,x_2}-\mathbf{g}\right)d\mathbf{x}\\ &\quad-\int_{\gamma_{i'}} L_m\left(\frac{x_i}{h}\right)(\mathbf{I}-\mathsf{P}_i)(\nu_{i'}\mathbf{A}_{i'}^{\bar{\mu}_{i'}})(\mathbf{u}_h^-+\mathbf{E}^--\mathbf{u}_h-\mathbf{E}^\perp)\,d\mathbf{s},\end{aligned} \tag{3.2.25b}$$

with i' denoting the dual of i defined in (2.2.9).

For $m=p+1$, (3.2.25a) can be reduced to

$$\int_\omega L_{p+1}^2\left(\frac{x_i}{h}\right)d\mathbf{x}\;\frac{d}{dt}\boldsymbol{\gamma}_i^{\mathbf{\Phi}}-\int_{\gamma_{i'}} L_{p+1}^2\left(\frac{x_i}{h}\right)(\mathbf{I}-\mathsf{P}_i)(\nu_{i'}\mathbf{A}_{i'}^{\bar{\mu}_{i'}})\boldsymbol{\gamma}_i^{\mathbf{\Phi}}\,d\mathbf{s}=\mathfrak{r}_{p+1,i}^{\mathbf{\Phi}}, \tag{3.2.26}$$

which is equal to

$$h^2\frac{d}{dt}\boldsymbol{\gamma}_i^{\mathbf{\Phi}}+h(\mathbf{I}-\mathsf{P}_i)(\mathbf{A}_{i'}^+-\mathbf{A}_{i'}^-)\boldsymbol{\gamma}_i^{\mathbf{\Phi}}=(2p+3)\mathfrak{r}_{p+1,i}^{\mathbf{\Phi}}. \tag{3.2.27a}$$

For $m=p$, we get similarly

$$h^2\frac{d}{dt}\boldsymbol{\delta}_i^{\mathbf{\Phi}}+h(\mathbf{I}-\mathsf{P}_i)(\mathbf{A}_{i'}^+-\mathbf{A}_{i'}^-)\boldsymbol{\delta}_i^{\mathbf{\Phi}}=(2p+1)\mathfrak{r}_{p,i}^{\mathbf{\Phi}}, \tag{3.2.27b}$$

subject to the initial conditions

$$\boldsymbol{\gamma}_i^{\mathbf{\Phi}}(0)=h^{p+1}\mathbf{c}_i(0),\quad \boldsymbol{\delta}_i^{\mathbf{\Phi}}(0)=h^{p+1}\mathbf{d}_i(0). \tag{3.2.27c}$$

We are ready to state and establish the convergence of the local error estimate for $\mathbf{e}^\perp$ and $\mathbf{e}^{\mathbf{\Phi}}$.

Theorem 3.2.2. *Under the assumptions of Theorem 3.1.3 with $p\geq 1$, let us consider the error estimate*

$$\mathbf{E}^\perp(t,h\boldsymbol{\xi})=\sum_{j=1}^{2}\left[L_{p+1}(\xi_j)-L_p(\xi_j)sgn(\mathbf{A}_j)\right]\frac{1}{2h}\mathbf{A}_j^\dagger\mathfrak{r}_{p,j}^\perp, \tag{3.2.28}$$

where $\mathfrak{r}_{p,j},\ j=1,2$ are defined in (3.2.15b). Then, at $t=\mathcal{O}(1)$,

$$\mathbf{e}^\perp(t,\mathbf{x})=\mathbf{E}^\perp(t,\mathbf{x})+\mathcal{O}(h^{p+2}),\ \mathbf{x}\in\omega. \tag{3.2.29}$$

Proof. Since the true solution $\mathbf{u}$ satisfies equation (3.0.1), we have

$$\int_\omega L_p\left(\frac{x_i}{h}\right)\mathsf{P}_i\left[\mathbf{u}_{,t}+\mathbf{A}_1\mathbf{u}_{,x_1}+\mathbf{A}_2\mathbf{u}_{,x_2}-\mathbf{g}\right]\,d\mathbf{x}=0,\ i=1,2. \tag{3.2.30}$$

Subtracting (3.2.14) from (3.2.30) yields

$$\int_\omega L_p\left(\frac{x_i}{h}\right)\mathsf{P}_i\left[\mathbf{e}_{,t}+\mathbf{A}_1(\mathbf{e}-\mathbf{E}^\perp)_{,x_1}+\mathbf{A}_2(\mathbf{e}-\mathbf{E}^\perp)_{,x_2}\right]\,d\mathbf{x}=0. \tag{3.2.31}$$

Applying (3.2.12a) and $\mathbf{A}_i\mathbf{e}^{\boldsymbol{\Xi}}_{,x_i}=\mathbf{0},\ i=1,2$, (3.2.31) becomes

$$\int_\omega L_p\left(\frac{x_i}{h}\right)\mathsf{P}_i\left[\mathbf{e}_{,t}+\mathbf{A}_1(\mathbf{e}^\perp-\mathbf{E}^\perp)_{,x_1}+\mathbf{A}_2(\mathbf{e}^\perp-\mathbf{E}^\perp)_{,x_2}\right]\,d\mathbf{x}=0. \tag{3.2.32}$$

Applying the linear transformations $t=T\tau,\ T>0$, and $\mathbf{x}=h\boldsymbol{\xi}$, (3.2.32) becomes

$$\int_\Delta L_p(\xi_i)\mathsf{P}_i\left[\frac{h}{T}\hat{\mathbf{e}}_{,\tau}+\mathbf{A}_1(\hat{\mathbf{e}}^\perp-\hat{\mathbf{E}}^\perp)_{,\xi_1}+\mathbf{A}_2(\hat{\mathbf{e}}^\perp-\hat{\mathbf{E}}^\perp)_{,\xi_2}\right]\,d\boldsymbol{\xi}=0. \tag{3.2.33}$$

where $\hat{\mathbf{e}}(\tau,\boldsymbol{\xi})=\mathbf{e}(\tau T,h\boldsymbol{\xi})$. By substituting the definitions of $\mathbf{e}^\perp$ (3.2.12b) and $\mathbf{E}^\perp$ (3.2.13a) into (3.2.33) we get

$$\int_\Delta \mathsf{P}_i\left[\frac{h^{p+2}}{T}L_p^2(\xi_i)\mathrm{sgn}(\mathbf{A}_i)\hat{\mathbf{c}}_{,\tau}+L_p(\xi_i)L'_{p+1}(\xi_i)\mathbf{A}_i(h^{p+1}\hat{\mathbf{c}}_i^\perp-\hat{\boldsymbol{\gamma}}_i^\perp)\right]\,d\boldsymbol{\xi}=\mathcal{O}(h^{p+2}), \tag{3.2.34}$$

where we used the orthogonality properties of Legendre polynomials. This can be further simplified to

$$\frac{h^{p+2}}{T(2p+1)}\mathrm{sgn}(\mathbf{A}_i)\hat{\mathbf{c}}_{,\tau}+2\mathbf{A}_i(h^{p+1}\hat{\mathbf{c}}_i^\perp-\hat{\boldsymbol{\gamma}}_i^\perp)=\mathcal{O}(h^{p+2}), \tag{3.2.35}$$

Thus, at $T=\mathcal{O}(1)$, we have

$$2\mathbf{A}_i(h^{p+1}\mathbf{c}_i^\perp-\boldsymbol{\gamma}_i^\perp)=\mathcal{O}(h^{p+2}). \tag{3.2.36}$$

Since $\mathbf{c}_i^\perp,\ \boldsymbol{\gamma}_i^\perp\in\mathcal{N}(\mathbf{A}_i)^\perp$, it follows that

$$\boldsymbol{\gamma}_i^\perp(t)=h^{p+1}\mathbf{c}_i^\perp(t)+\mathcal{O}(h^{p+2}),\quad i=1,2, \tag{3.2.37}$$

which establishes (3.2.29). □

Next, we will state and prove a technical lemma before stating and proving our last theorem.

Lemma 3.2.3. *Let $\mathbf{A}_1$ and $\mathbf{A}_2$ be symmetric matrices that satisfy the assumptions of Lemma 3.1.1 and $\mathbf{q}\in\mathcal{E}_p$. If $\mathbf{q}$ satisfies the orthogonality condition on the reference element*

$$\int_\Delta \mathbf{v}^t(\mathbf{A}_1\mathbf{q}_{,\xi_1}+\mathbf{A}_2\mathbf{q}_{,\xi_2})\,d\boldsymbol{\xi}-\int_\Gamma \mathbf{v}^t(\nu_1\mathbf{A}_1^{\bar{\mu}_1}+\nu_2\mathbf{A}_2^{\bar{\mu}_2})\mathbf{q}\,d\boldsymbol{\sigma}=0,\ \forall\quad \mathbf{v}\in\mathcal{E}_p, \tag{3.2.38}$$

then $\mathbf{q}=\mathbf{0}$.

Proof. First we integrate equation (3.2.38) by parts to write

$$-\int_\Delta \left(\mathbf{v}^t_{,\xi_1}\mathbf{A}_1 + \mathbf{v}^t_{,\xi_2}\mathbf{A}_2\right)\mathbf{q}\,d\boldsymbol{\xi} + \int_\Gamma \mathbf{v}^t(\nu_1\mathbf{A}_1^{\mu_1} + \nu_2\mathbf{A}_2^{\mu_2})\mathbf{q}\,d\boldsymbol{\sigma} = 0. \tag{3.2.39}$$

Adding (3.2.38) and (3.2.39) and setting $\mathbf{v} = \mathbf{q}$, the integral on Δ vanishes because of the symmetry of $\mathbf{A}_1$ and $\mathbf{A}_2$, and we get

$$\int_{\Gamma_1} \mathbf{q}^t(\mathbf{A}_1^+ - \mathbf{A}_1^-)\mathbf{q}\,d\boldsymbol{\sigma} + \int_{\Gamma_2} \mathbf{q}^t(\mathbf{A}_2^+ - \mathbf{A}_2^-)\mathbf{q}\,d\boldsymbol{\sigma} = 0. \tag{3.2.40}$$

Since $\mathbf{q} \in \mathcal{E}_p$, there are $\mathbf{a}_i, \mathbf{b}_i \in \mathcal{N}(\mathbf{A}_i)$ for $i = 1, 2$, such that

$$\mathbf{q}(\xi) = \sum_{i=1}^{2} L_{p+1}(\xi_i)\mathbf{a}_i - L_p(\xi_i)\mathbf{b}_i. \tag{3.2.41}$$

Combining property (1.2.25b), $(\mathbf{A}_i^+ - \mathbf{A}_i^-)\mathbf{a}_i = (\mathbf{A}_i^+ - \mathbf{A}_i^-)\mathbf{b}_i = \mathbf{0}$ for $i = 1, 2$, and (3.2.40) we obtain

$$\begin{aligned}\frac{1}{2p+3}\mathbf{a}_2^t(\mathbf{A}_1^+ - \mathbf{A}_1^-)\mathbf{a}_2 + \frac{1}{2p+1}\mathbf{b}_2^t(\mathbf{A}_1^+ - \mathbf{A}_1^-)\mathbf{b}_2 \\ +\frac{1}{2p+3}\mathbf{a}_1^t(\mathbf{A}_2^+ - \mathbf{A}_2^-)\mathbf{a}_1 + \frac{1}{2p+1}\mathbf{b}_1^t(\mathbf{A}_2^+ - \mathbf{A}_2^-)\mathbf{b}_1 = 0.\end{aligned} \tag{3.2.42}$$

Since $(\mathbf{A}_i^+ - \mathbf{A}_i^-)$ is positive semi-definite, there exists a matrix $\mathbf{L}_i$ such that $\mathbf{L}_i^t\mathbf{L}_i = (\mathbf{A}_i^+ - \mathbf{A}_i^-)$ for $i = 1, 2$ and (3.2.42) can be written as

$$\frac{1}{2p+3}\|\mathbf{L}_1\mathbf{a}_2\|^2 + \frac{1}{2p+1}\|\mathbf{L}_1\mathbf{b}_2\|^2 + \frac{1}{2p+3}\|\mathbf{L}_2\mathbf{a}_1\|^2 + \frac{1}{2p+1}\|\mathbf{L}_2\mathbf{b}_1\|^2 = 0. \tag{3.2.43}$$

This leads to

$$\mathbf{L}_1\mathbf{a}_2 = \mathbf{L}_1\mathbf{b}_2 = \mathbf{L}_2\mathbf{a}_1 = \mathbf{L}_2\mathbf{b}_1 = \mathbf{0}. \tag{3.2.44}$$

We pre-multiply $\mathbf{L}_1\mathbf{a}_2$ and $\mathbf{L}_1\mathbf{b}_2$ by $\mathbf{L}_1^t$ and $\mathbf{L}_2\mathbf{a}_1$ and $\mathbf{L}_2\mathbf{b}_1$ by $\mathbf{L}_2^t$ to show that

$$(\mathbf{A}_1^+ - \mathbf{A}_1^-)\mathbf{a}_2 = (\mathbf{A}_1^+ - \mathbf{A}_1^-)\mathbf{b}_2 = (\mathbf{A}_2^+ - \mathbf{A}_2^-)\mathbf{a}_1 = (\mathbf{A}_2^+ - \mathbf{A}_2^-)\mathbf{b}_1 = \mathbf{0}. \tag{3.2.45}$$

We combine Lemmas 3.1.1 and 3.1.2, which yield $\mathcal{N}(\mathbf{A}_1) \cap \mathcal{N}(\mathbf{A}_2) = \{\mathbf{0}\}$, the definition of $\mathcal{E}_p$ and property (1.2.25b) to infer that $\mathbf{a}_2 = \mathbf{b}_2 = \mathbf{0}$ and $\mathbf{a}_1 = \mathbf{b}_1 = \mathbf{0}$. Thus, we establish Lemma 3.2.3. □

Now we state and prove the convergence of the error estimate $\mathbf{E}^\Pi$ to $\mathbf{e}^\Pi$ under mesh refinement.

Theorem 3.2.4. *Under the assumptions of Theorem 3.1.3 with $p \geq 1$ we let*

$$\mathbf{E}^{\Phi}(t, h\boldsymbol{\xi}) = \sum_{j=1}^{2} (L_{p+1}(\xi_j)\boldsymbol{\gamma}_j^{\Phi}(t) - L_p(\xi_j)\boldsymbol{\delta}_j^{\Phi}(t)), \tag{3.2.46}$$

where $\boldsymbol{\gamma}_i^{\Phi}, \boldsymbol{\delta}_i^{\Phi},\ i = 1,\ 2$ are solutions of (3.2.27) and (3.2.25b).

Then, at $t = \mathcal{O}(1)$,

$$\mathbf{e}^{\Phi}(t, \mathbf{x}) = \mathbf{E}^{\Phi}(t, \mathbf{x}) + \mathcal{O}(h^{p+2}),\ \forall\ \mathbf{x} \in \omega. \tag{3.2.47}$$

Proof. Since the true solution $\mathbf{u}$ is continuous and $\mathbf{u} = \mathbf{u}^-$ on $\partial\omega$, $\mathbf{u}$ satisfies

$$\begin{aligned} &\int_{\omega} \mathbf{v}^t \left[\mathbf{u}_{,t} + \mathbf{A}_1\mathbf{u}_{,x_1} + \mathbf{A}_2\mathbf{u}_{,x_2} - \mathbf{g}\right] d\mathbf{x} \\ &\quad + \int_{\partial\omega} \mathbf{v}^t(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})(\mathbf{u}^- - \mathbf{u})\, d\mathbf{s} = 0, \quad \forall\ \mathbf{v} \in \mathcal{E}_p. \end{aligned} \tag{3.2.48}$$

Subtracting (3.2.23b) from (3.2.48) gives

$$\begin{aligned} &\int_{\omega} \mathbf{v}^t \left[(\mathbf{e} - \mathbf{E}^{\Phi})_{,t} + \mathbf{A}_1\mathbf{e}_{,x_1} + \mathbf{A}_2\mathbf{e}_{,x_2}\right] d\mathbf{x} \\ &\quad + \int_{\partial\omega} \mathbf{v}^t(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})(\mathbf{e}^- - \mathbf{E}^- - \mathbf{e} + \mathbf{E}^{\perp} + \mathbf{E}^{\Phi})\, d\mathbf{s} = 0. \end{aligned} \tag{3.2.49}$$

Using the definition of $\mathbf{E}^{\perp}$ and $\mathbf{E}^{\Phi}$ and the orthogonality properties of Legendre polynomials one can easily check that the following holds

$$\int_{\omega} \mathbf{v}^t \left[\mathbf{E}_{,t}^{\perp} + \mathbf{A}_1(\mathbf{E}^{\perp} + \mathbf{E}^{\Phi})_{,x_1} + \mathbf{A}_2(\mathbf{E}^{\perp} + \mathbf{E}^{\Phi})_{,x_2}\right] d\mathbf{x} = 0, \quad \forall\ \mathbf{v} \in \mathcal{E}_p. \tag{3.2.50}$$

Subtracting (3.2.50) from (3.2.49) gives for $\boldsymbol{\epsilon} = \mathbf{e} - \mathbf{E}^{\perp} - \mathbf{E}^{\Phi}$ and $\boldsymbol{\epsilon}^- = \mathbf{e}^- - \mathbf{E}^-$

$$\int_{\omega} \mathbf{v}^t \left[\boldsymbol{\epsilon}_{,t} + \mathbf{A}_1\boldsymbol{\epsilon}_{,x_1} + \mathbf{A}_2\boldsymbol{\epsilon}_{,x_2}\right] d\mathbf{x} + \int_{\partial\omega} \mathbf{v}^t(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})(\boldsymbol{\epsilon}^- - \boldsymbol{\epsilon})\, d\mathbf{s} = 0. \tag{3.2.51}$$

Applying the linear transformations $t = T\tau,\ T > 0$, and $\mathbf{x} = h\boldsymbol{\xi}$, (3.2.51) becomes

$$\int_{\Delta} \mathbf{v}^t \left[\frac{h}{T}\boldsymbol{\epsilon}_{,\tau} + \mathbf{A}_1\boldsymbol{\epsilon}_{,\xi_1} + \mathbf{A}_2\boldsymbol{\epsilon}_{,\xi_2}\right] d\mathbf{x} + \int_{\Gamma} \mathbf{v}^t(\nu_1\mathbf{A}_1^{\bar{\mu}_1} + \nu_2\mathbf{A}_2^{\bar{\mu}_2})(\boldsymbol{\epsilon}^- - \boldsymbol{\epsilon})\, d\boldsymbol{\sigma} = 0. \tag{3.2.52}$$

The Maclaurin series of $\boldsymbol{\epsilon}$ with respect to h is

$$\boldsymbol{\epsilon}(t, \boldsymbol{\xi}) = \sum_{k=0}^{p+1} h^k \mathbf{q}_k(t, \boldsymbol{\xi}) + \mathcal{O}(h^{p+2}). \tag{3.2.53}$$

Since $\mathbf{e}^{\perp} - \mathbf{E}^{\perp} = \mathcal{O}(h^{p+2})$, $\mathbf{q}_k \in \mathcal{E}_p$, for $k < p+2$.

Substituting $\boldsymbol{\epsilon}^{-}(t, \boldsymbol{\xi}) = \mathcal{O}(h^{p+2})$ (from the definition of $\mathbf{E}^{-}$ (3.2.22b)) and (3.2.53) into (3.2.51) yields

$$\sum_{k=0}^{p+1} h^k \Bigg(\int_{\Delta} \mathbf{v}^t \left[\frac{h}{T} \mathbf{q}_{k,t} + \mathbf{A}_1 \mathbf{q}_{k,x_1} + \mathbf{A}_2 \mathbf{q}_{k,x_2} \right] d\mathbf{x} - \int_{\Gamma} \mathbf{v}^t (\nu_1 \mathbf{A}_1^{\bar{\mu}_1} + \nu_2 \mathbf{A}_2^{\bar{\mu}_2}) \mathbf{q}_k \, d\boldsymbol{\sigma} \Bigg) = \mathcal{O}(h^{p+2}), \tag{3.2.54}$$

which infers that all terms of the same power in h are zero.

For instance, the $\mathcal{O}(1)$ term leads to the orthogonality condition for $\mathbf{q}_0$

$$\int_{\Delta} \mathbf{v}^t \left[\mathbf{A}_1 \mathbf{q}_{0,x_1} + \mathbf{A}_2 \mathbf{q}_{0,x_2} \right] d\mathbf{x} - \int_{\Gamma} \mathbf{v}^t (\nu_1 \mathbf{A}_1^{\bar{\mu}_1} + \nu_2 \mathbf{A}_2^{\bar{\mu}_2}) \mathbf{q}_0 \, d\boldsymbol{\sigma} = 0, \ \forall \ \mathbf{v} \in \mathcal{E}_p. \tag{3.2.55}$$

By Lemma 3.2.3, $\mathbf{q}_0 = \mathbf{0}$.

Using induction, we assume that $\mathbf{q}_l = \mathbf{0}, \ 0 \leq l \leq k-1$, and apply the $\mathcal{O}(h^k)$ term to obtain the orthogonality condition

$$\int_{\Delta} \mathbf{v}^t \left[\mathbf{A}_1 \mathbf{q}_{k,x_1} + \mathbf{A}_2 \mathbf{q}_{k,x_2} \right] d\mathbf{x} - \int_{\Gamma} \mathbf{v}^t (\nu_1 \mathbf{A}_1^{\bar{\mu}_1} + \nu_2 \mathbf{A}_2^{\bar{\mu}_2}) \mathbf{q}_k \, d\boldsymbol{\sigma} = 0, \ \forall \ \mathbf{v} \in \mathcal{E}_p. \tag{3.2.56}$$

Again, by Lemma 3.2.3, $\mathbf{q}_k = \mathbf{0}$.

Hence, $\mathbf{q}_k = \mathbf{0}, \ k = 0, \cdots, p+1$, *i.e.*,

$$\boldsymbol{\epsilon}(t, \boldsymbol{\xi}) = \mathcal{O}(h^{p+2}), \tag{3.2.57}$$

which completes the proof. □

We conclude this section by noting that the transient estimate $\mathbf{E}^{\maltese}(t, \mathbf{x}) \in \mathcal{N}(\mathbf{A}_1) \bigoplus \mathcal{N}(\mathbf{A}_2)$. Hence, the stationary error estimate $\mathbf{E}^{\perp}(t, \mathbf{x})$ is accurate only for the error component lying in $(\mathcal{N}(\mathbf{A}_1) \bigoplus \mathcal{N}(\mathbf{A}_2))^{\perp}$.

3.3 Computational Examples

We validate our theory on one- and two-dimensional linear symmetric hyperbolic systems. The accuracy of *a posteriori* error estimates is measured by the local effectivity indices

$$\theta_e = \frac{\|\mathbf{E}\|_{L^2(\omega_e)}}{\|\mathbf{e}\|_{L^2(\omega_e)}}, \tag{3.3.1}$$

and the global effectivity index with respect to the L^2 norm

$$\theta = \frac{\|\mathbf{E}\|_{L^2(\Omega)}}{\|\mathbf{e}\|_{L^2(\Omega)}}. \tag{3.3.2}$$

The componentwise effectivity indices are defined as

$$\theta^* = \left(\frac{\|E_1\|_{2,\Omega}}{\|e_1\|_{2,\Omega}}, \cdots, \frac{\|E_m\|_{2,\Omega}}{\|e_m\|_{2,\Omega}} \right)^t. \tag{3.3.3}$$

We also need the componentwise L^2-error

$$\|\mathbf{e}\|^* = \left(\|e_1\|_{2,\Omega}, \cdots, \|e_m\|_{2,\Omega} \right)^t, \tag{3.3.4}$$

where $\mathbf{E} = (E_1, \cdots, E_m)^t$ and $\mathbf{e} = (e_1, \cdots, e_m)^t$.

Ideally, the effectivity indices should approach unity under mesh refinement.

3.3.1 Examples for Superconvergence

We start with an example to validate our superconvergence results of Theorem 3.1.3. Then, we apply our error estimation procedure to several problems to show that computations and theory are in full agreement. We note that, we study only the spatial component of the DG discretization error only and in all numerical computations we integrated in time using the Dormand-Prince method with a very small time-step, as described in §2.3. Therefore we assume the temporal component of the error to be negligible. We use the L^2-projection $\Pi\mathbf{u}_0$ to approximate the initial conditions and π_1, π_2 to approximate the boundary conditions.

Example 3.3.1. *We consider the linearized one-dimensional Euler equations with proper time and space scalings*

$$\begin{pmatrix} \mathsf{p} \\ u \end{pmatrix}_{,t} + \mathbf{A} \begin{pmatrix} \mathsf{p} \\ u \end{pmatrix}_{,x} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ x \in (0,1), \ 0 < t < 1, \ \mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{3.3.5a}$$

and select the initial and boundary conditions such that the exact solution is

$$\begin{pmatrix} \mathsf{p} \\ u \end{pmatrix} = \begin{pmatrix} \sin(t)\cos(x-1) \\ -\cos(t)\sin(x-1) \end{pmatrix}, \tag{3.3.5b}$$

where u and p denote the velocity and pressure, respectively.

Basic linear algebra yields that $\mathbf{z}_1 = (1,1)^t \in \mathcal{R}(\mathbf{A}^+)$, $\mathbf{z}_2 = (1,-1)^t \in \mathcal{R}(\mathbf{A}^-)$. Thus, applying Theorem 3.2.1 leads to

$$\mathbf{z}_1^t \mathbf{e}(t, \xi(x)) = h^{p+1}(L_{p+1}(\xi) - L_p(\xi))c(t) + \mathcal{O}(h^{p+2}), \tag{3.3.6}$$

$$\mathbf{z}_2^t \mathbf{e}(t, \xi(x)) = h^{p+1}(L_{p+1}(\xi) + L_p(\xi))d(t) + \mathcal{O}(h^{p+2}). \tag{3.3.7}$$

Maximum error $E = |(1,1)\mathbf{e}|$ at right Radau points

	$p=0$		$p=1$		$p=2$		$p=3$	
N	E	*order*	E	*order*	E	*order*	E	*order*
10	3.560*e*-2	–	2.149*e*-5	–	1.124*e*-7	–	2.546*e*-10	–
20	1.882*e*-2	0.9192	2.854*e*-6	2.9128	7.043*e*-9	3.9961	8.175*e*-12	4.9610
30	1.286*e*-2	0.9404	8.621*e*-7	2.9524	1.392*e*-9	3.9979	1.084*e*-12	4.9841
40	9.773*e*-3	0.9529	3.694*e*-7	2.9455	4.407*e*-10	3.9987	2.565*e*-13	5.0092

Maximum error $E = |(1,\text{-}1)\mathbf{e}|$ at left Radau points

	$p=0$		$p=1$		$p=2$		$p=3$	
N	E	*order*	E	*order*	E	*order*	E	*order*
10	3.560*e*-2	–	2.149*e*-5	–	1.124*e*-7	–	2.546*e*-10	–
20	1.882*e*-2	0.9192	2.854*e*-6	2.9128	7.043*e*-9	3.9961	8.175*e*-12	4.9610
30	1.286*e*-2	0.9404	8.621*e*-7	2.9524	1.392*e*-9	3.9979	1.084*e*-12	4.9841
40	9.773*e*-3	0.9529	3.694*e*-7	2.9455	4.407*e*-10	3.9987	2.565*e*-13	5.0092

Table 3.3.1: Maximum projected errors $|(1,1)\mathbf{e}|$ at left Radau points and $|(1,\text{-}1)\mathbf{e}|$ at right Radau points and their order of convergence at $t = 1$ for Example 3.3.1.

Therefore, on each element in the mesh $\mathbf{z}_1^t\mathbf{e}$ and $\mathbf{z}_2^t\mathbf{e}$ is $\mathcal{O}(h^{p+2})$ superconvergent at the shifted roots of right Radau polynomial $R_{p+1}^+(x)$ and the left Radau polynomial $R_{p+1}^-(x)$, respectively.

We first solve (3.3.5) on a uniform mesh having $N = 6$ elements for $p = 0, 1, 2, 3$ and plot the projected true errors $\mathbf{z}_i^t\mathbf{e}$, $i = 1, 2$, at $t = 1$ versus x in Figure 3.3.1. We observe that the error plots for $p \geq 1$ intersect the x-axis very close to Radau points marked by $\times$.

In order to show the $\mathcal{O}(h^{p+2})$-superconvergence rates under mesh refinement, we solve (3.3.5) on uniform meshes having $N = 10, 20, 30, 40$ elements for $p = 0, 1, 2, 3$ and show maximum errors $|\mathbf{z}_1^t\mathbf{e}|$ and $|\mathbf{z}_2^t\mathbf{e}|$ at Radau points and their rates of convergence in Table 3.3.1. We observe that the maximum projected errors at Radau points are $\mathcal{O}(h^{p+2})$-superconvergent while the L^2-error is only $\mathcal{O}(h^{p+1})$.

Example 3.3.2. *Let us consider the two-dimensional hyperbolic system*

$$\mathbf{u}_{,t} + \mathbf{A}_1\mathbf{u}_{,x} + \mathbf{A}_2\mathbf{u}_{,y} = \mathbf{g}(t,x,y), \quad (x,y) \in (0,1)^2,\ 0 < t < 1, \tag{3.3.8a}$$

where

$$\mathbf{A}_1 = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix},\ \mathbf{A}_2 = \begin{pmatrix} -5 & 1 & 0 \\ 1 & -5 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \tag{3.3.8b}$$

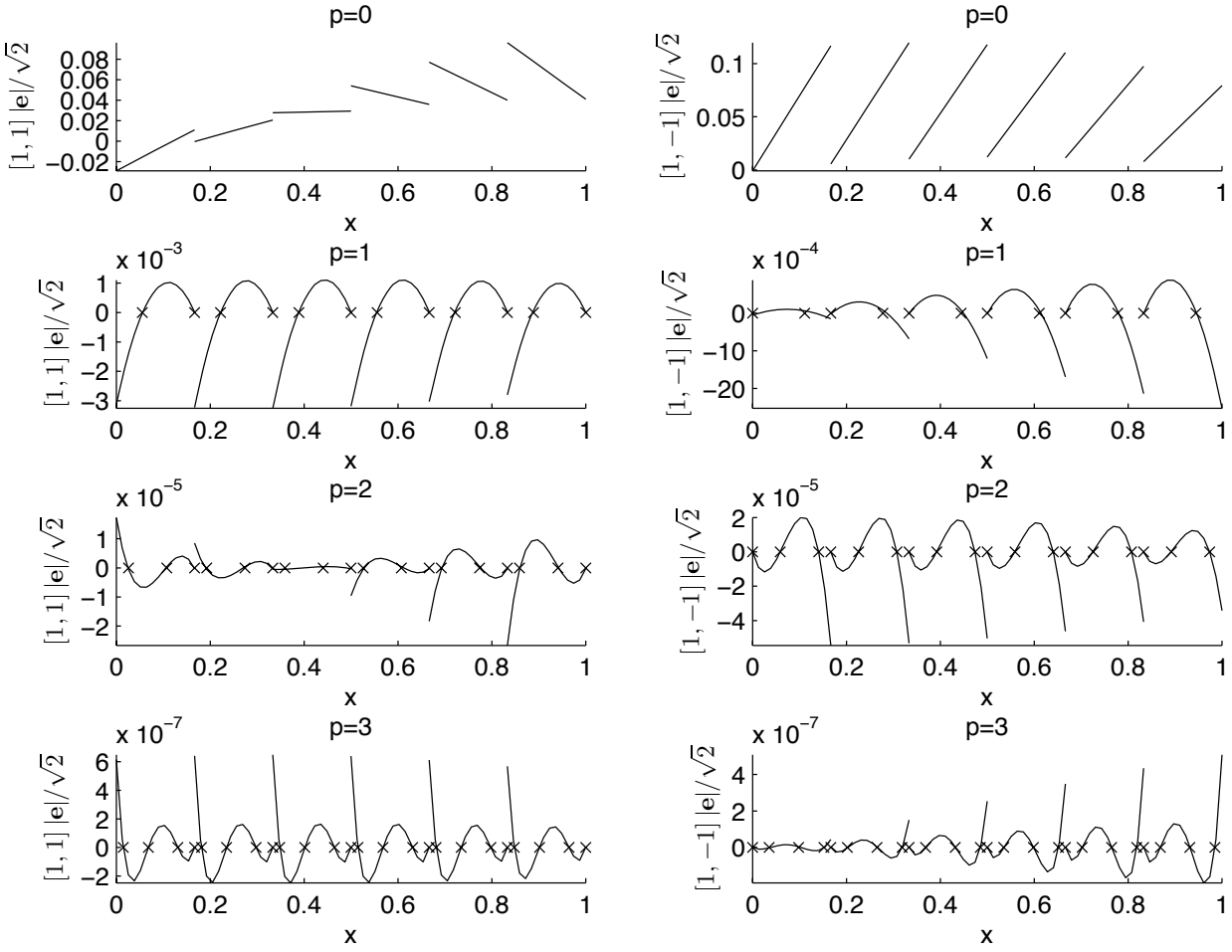

Figure 3.3.1: Projected errors $\frac{1}{\sqrt{2}}(1,1)\mathbf{e}$, $\frac{1}{\sqrt{2}}(1,-1)\mathbf{e}$ versus x at $t = 1$ for Example 3.3.1. Shifted right Radau (left) and left Radau (right) points are marked by $\times$.

and select $\mathbf{g}(t,x,y)$*, initial and boundary conditions such that the true solution is*

$$\mathbf{u} = \begin{pmatrix} \exp(t-x-y) \\ \exp(-t+x-y) \\ \exp(-t-x+y) \end{pmatrix}. \tag{3.3.8c}$$

Basic linear algebra yields that $(0,1,0)^t$, $(1,0,-1)^t$, $(1,0,1)^t$ are eigenvectors of $\mathbf{A}_1$ corresponding to eigenvalues 0, 1, and 3, respectively. On the other-hand, the vectors $(1,1,0)^t$, $(1,-1,0)^t$, and $(0,0,1)^t$ are eigenvectors of $\mathbf{A}_2$ corresponding to eigenvalues -6, -4, and 2, respectively.

We further note that $\mathcal{R}(\mathbf{A}_1^+) = \operatorname{span}\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_3\}$, $\mathcal{R}(\mathbf{A}_1^-) = \{\mathbf{0}\}$, $\mathcal{R}(\mathbf{A}_2^+) = \operatorname{span}\{\boldsymbol{\varepsilon}_3\}$ and $\mathcal{R}(\mathbf{A}_2^-) = \operatorname{span}\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\}$, where $\boldsymbol{\varepsilon}_i$, $i = 1,2,3$, is the canonical basis of $\mathbb{R}^3$ which leads to

$$\mathcal{R}(\mathbf{A}_1^+) \cap \mathcal{R}(\mathbf{A}_2^+) = \operatorname{span}\{\boldsymbol{\varepsilon}_3\}, \tag{3.3.9a}$$
$$\mathcal{R}(\mathbf{A}_1^+) \cap \mathcal{R}(\mathbf{A}_2^-) = \operatorname{span}\{\boldsymbol{\varepsilon}_1\}. \tag{3.3.9b}$$

Applying Theorem 3.2.1, we expect pointwise $\mathcal{O}(h^{p+2})$-superconvergence of the error projections $e_1 = \boldsymbol{\varepsilon}_1^t \mathbf{e}$ and $e_3 = \boldsymbol{\varepsilon}_3^t \mathbf{e}$, *i.e.*, the first and third components of $\mathbf{e}$ are $\mathcal{O}(h^{p+2})$-superconvergent at Radau points.

We solve (3.3.8) on a 4×4 uniform mesh for $p = 1,2,3$ and plot the zero-level curves of the first and third components of the error in Figure 3.3.1. We observe that the zero-level curves pass near shifted Radau points marked by $\times$.

In order to show superconvergence rates we solve (3.3.8) on uniform square meshes having $N = 5^2, 10^2, 15^2, 20^2, 25^2$ elements with $p = 1,2,3$ and present the maximum errors of $|e_1|$ and $|e_3|$ at shifted Radau points over all elements and their order of convergence under mesh refinement. We observe $\mathcal{O}(h^{p+2})$ superconvergence rates which is in full agreement with Theorem 3.2.1.

3.3.2 Examples for *A Posteriori* Error Estimation

Here, we solve several examples to validate our *a posteriori* error estimation procedure. We use π_i, $i = 1,2$, to approximate the boundary conditions and both L^2-projection $\Pi\mathbf{u}_0$ and $\pi\mathbf{u}_0$ to approximate the initial conditions.

Example 3.3.3. We solve the one-dimensional Euler's system (3.3.5) on uniform meshes for $0 < t \leq 1$ with $p = 0,1,2,3$ and $N = 50, 75, 100$, and compute *a posteriori* error estimates $\mathbf{E}^\perp$ by solving the stationary problem (3.2.13). We present the $L^2(\Omega)$ norm of the error and effectivity indices at $t = 1$ in Table 3.3.3 using Πu_0. These results indicate that the

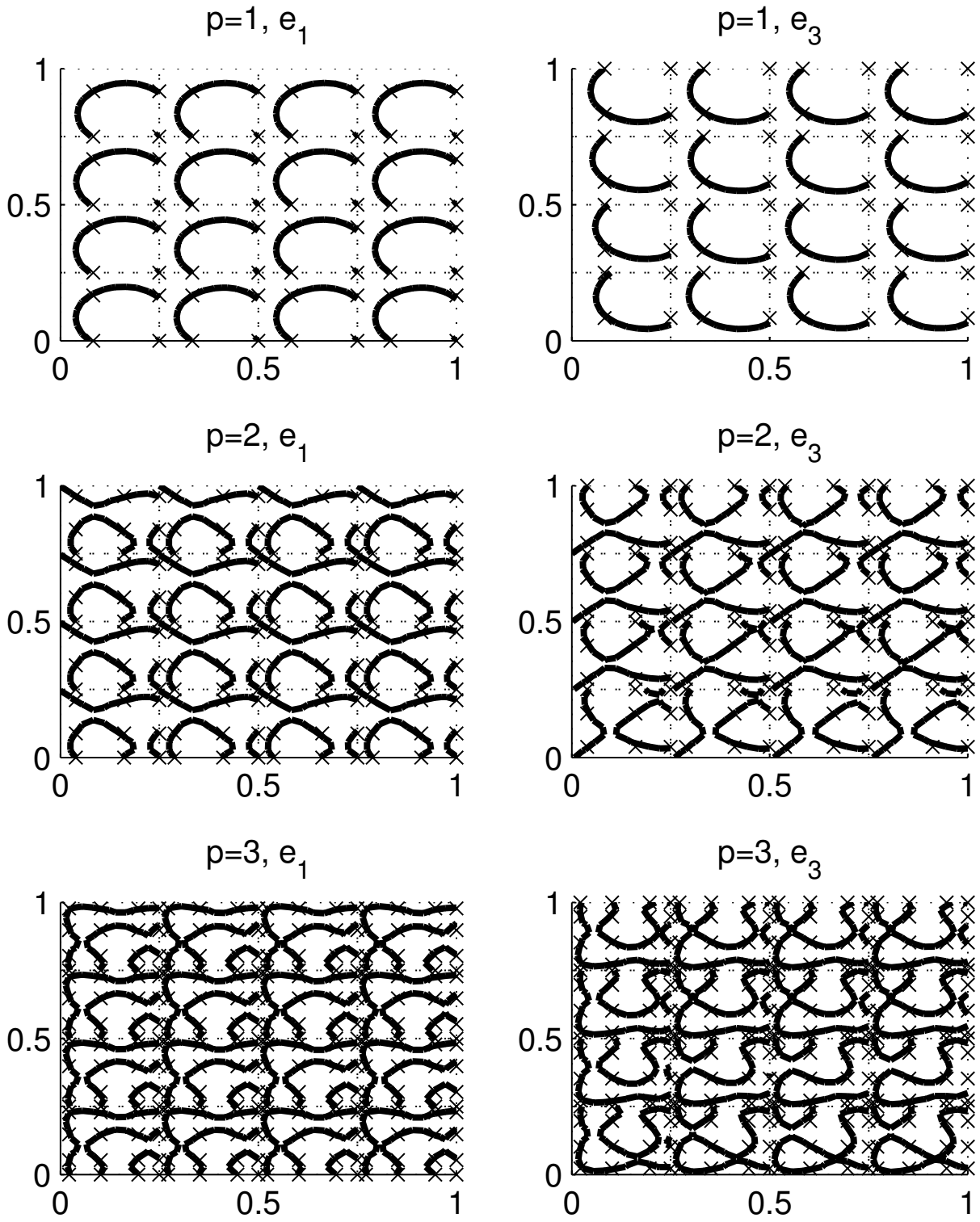

Figure 3.3.2: Zero-level curves of e_1, e_3 at $t = 1$ for Example 3.3.2. Shifted Radau points are marked by $\times$.

Maximum error $E = |e_1|$ at Radau points

	$p=0$		$p=1$		$p=2$		$p=3$	
N	E	*order*	E	*order*	E	*order*	E	*order*
5^2	0.1260	–	1.054*e*-3	–	1.704*e*-5	–	5.682*e*-7	–
10^2	6.899*e*-2	0.8688	1.622*e*-4	2.6997	1.149*e*-6	3.8901	2.156*e*-8	4.7198
15^2	4.790*e*-2	0.8999	5.128*e*-5	2.8407	2.321*e*-7	3.9456	3.036*e*-9	4.8355
20^2	3.678*e*-2	0.9186	2.272*e*-5	2.8303	7.415*e*-8	3.9661	7.450*e*-10	4.8833
25^2	2.973*e*-2	0.9525	1.199*e*-5	2.8638	3.055*e*-8	3.9731	2.491*e*-10	4.9084

Maximum error $E = |e_3|$ at Radau points

	$p=0$		$p=1$		$p=2$		$p=3$	
N	E	*order*	E	*order*	E	*order*	E	*order*
5^2	0.1260	–	1.054*e*-3	–	1.704*e*-5	–	5.682*e*-7	–
10^2	6.899*e*-2	0.8688	1.622*e*-4	2.6997	1.149*e*-6	3.8901	2.156*e*-8	4.7198
15^2	4.790*e*-2	0.8999	5.128*e*-5	2.8407	2.321*e*-7	3.9456	3.036*e*-9	4.8355
20^2	3.678*e*-2	0.9186	2.272*e*-5	2.8303	7.415*e*-8	3.9661	7.450*e*-10	4.8833
25^2	2.973*e*-2	0.9525	1.199*e*-5	2.8638	3.055*e*-8	3.9731	2.491*e*-10	4.9084

Table 3.3.2: Maximum errors for $|e_1|$ at shifted Radau points (ξ_i^+, ξ_j^+) and $|e_3|$ at shifted Radau points (ξ_i^+, ξ_j^-) and $t = 1$ for Example 3.3.2.

stationary error estimates $\mathbf{E}^{\perp}$ are accurate estimates of $\mathbf{e}$ for $p > 1$, which is in full agreement with the theory.

In Figure 3.3.3, we plot the effectivity indices versus time for $N = 100, 200, 300$ and $p = 1, 2, 3$ using $\Pi\mathbf{u}_0$ (solid) and $\pi\mathbf{u}_0$ (dotted). We observe that the effectivity indices for $\Pi\mathbf{u}_0$ slightly oscillate about unity and then get closer to unity with increasing time $t > \mathcal{O}(h)$. On the other hand, the global effectivity indices for $\pi\mathbf{u}_0$ stay close to unity at all times. Thus, we recommend the use of $\pi\mathbf{u}_0$.

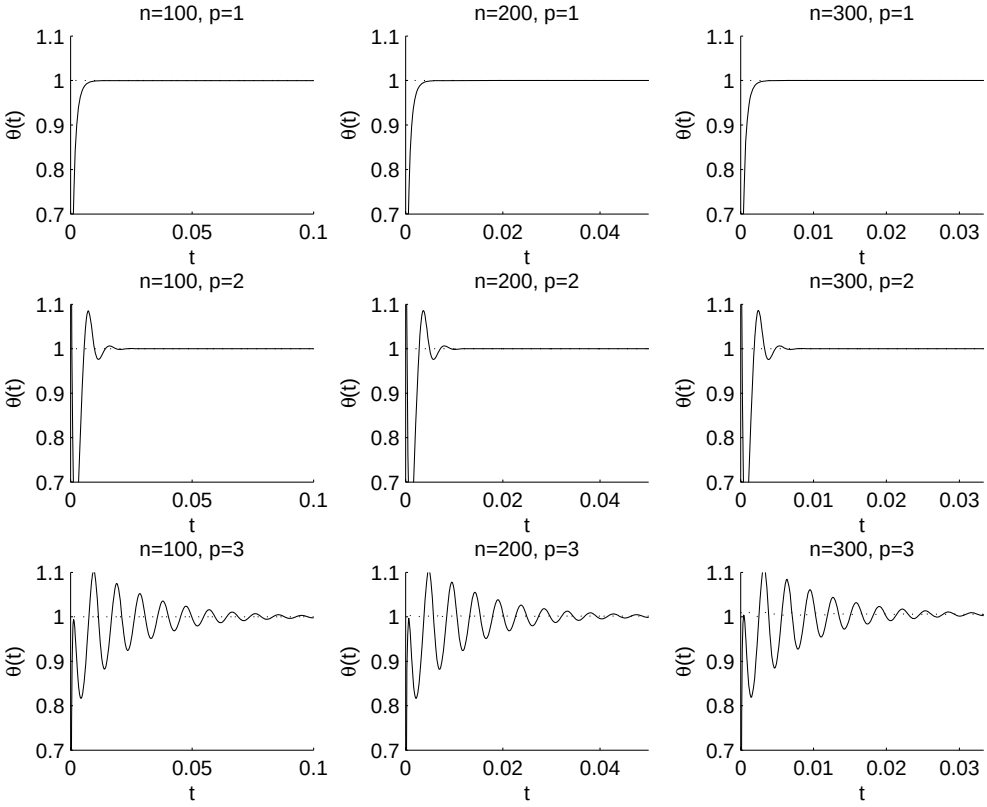

Figure 3.3.3: Global effectivity indices versus t over the interval $[0, \frac{10}{h}]$ for $N = 100, 200, 300$, $p = 1, 2, 3$ using $\pi\mathbf{u}_0$ (dotted) and $\Pi\mathbf{u}_0$ (solid) for Example 3.3.3.

Example 3.3.4. *Let us consider the two-dimensional wave equation*

$$\frac{\partial^2 v}{\partial t^2} = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}, \ (x, y) \in (0, 1)^2, \ 0 < t \leq 1, \tag{3.3.10}$$

which can be written as the first-order linear hyperbolic system

$$\mathbf{u}_{,t} + \mathbf{A}_1 \mathbf{u}_{,x} + \mathbf{A}_2 \mathbf{u}_{,y} = 0, \ (x, y) \in (0, 1)^2, \ 0 < t \leq 1, \tag{3.3.11a}$$

p	N	$\|\mathbf{e}\|_{2,\Omega}$	order	$\|\mathbf{e}-\mathbf{E}^{\perp}\|_{2,\Omega}$	order	$[\min_e \theta_e, \max_e \theta_e]$	θ
0	50	$9.175e-03$	–	$3.713e-03$	–	$[0.343, 1.878]$	0.7945
	75	$6.187e-03$	0.972	$2.518e-03$	0.958	$[0.335, 1.900]$	0.7877
	100	$4.671e-03$	0.977	$1.907e-03$	0.967	$[0.331, 1.914]$	0.7836
1	50	$1.875e-05$	–	$2.161e-07$	–	$[0.992, 1.001]$	0.9997
	75	$8.338e-06$	1.999	$7.084e-08$	2.750	$[0.993, 1.001]$	0.9998
	100	$4.691e-06$	1.999	$3.229e-08$	2.731	$[0.994, 1.001]$	0.9999
2	50	$2.488e-08$	–	$1.160e-10$	–	$[0.997, 1.001]$	0.9999
	75	$7.369e-09$	3.001	$2.401e-11$	3.885	$[0.998, 1.001]$	1.0000
	100	$3.108e-09$	3.001	$7.882e-12$	3.872	$[0.998, 1.001]$	1.0000
3	50	$3.699e-11$	–	$7.066e-14$	–	$[0.999, 1.001]$	1.0000
	75	$7.309e-12$	3.999	$9.712e-15$	4.894	$[0.999, 1.000]$	0.9999
	100	$2.313e-12$	3.999	$3.925e-15$	3.149	$[0.999, 1.000]$	0.9998

Table 3.3.3: L^2 errors $\|\mathbf{e}\|_{2,\Omega}$, $\|\mathbf{e}-\mathbf{E}^{\perp}\|_{2,\Omega}$, their rates of convergence with maximum and minimum local effectivity indices and global effectivity indices for $\mathbf{E}^{\perp}$ at $t=1$ for Example 3.3.3.

where

$$\mathbf{u} = \begin{pmatrix} v_{,t} + v_{,x} \\ v_{,y} \end{pmatrix}, \ \mathbf{A}_1 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \ \mathbf{A}_2 = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \tag{3.3.11b}$$

and select initial and boundary conditions such that the true solution is

$$\mathbf{u} = \begin{pmatrix} \sin(\sqrt{2}t + x + y) - \cos(-\sqrt{2}t + x + y) \\ (\sqrt{2}-1)\sin(\sqrt{2}t + x + y) + (1+\sqrt{2})\cos(-\sqrt{2}t + x + y)) \end{pmatrix}. \tag{3.3.11c}$$

We solve (3.3.11) on uniform meshes having $N = 10^2, 20^2, 30^2$ elements for $p = 0, 1, 2, 3$ using $\Pi\mathbf{u}_0$ and present the L^2 errors and effectivity indices corresponding to the stationary error estimates $\mathbf{E}^{\perp}$ at $t = 1$ in Table 3.3.4. We observe that the effectivity indices converge to unity under mesh refinement. Furthermore, we plot the effectivity indices versus time in Figure 3.3.4 to note that the stationary error estimate $\mathbf{E}^{\perp}$ is asymptotically accurate which is in full agreement with Theorem 3.2.2. We further note that the effectivity indices stay close to unity at all times when using $\pi\mathbf{u}_0$. For $\Pi\mathbf{u}_0$ the effectivity indices oscillates about unity near $t = 0$ before approaching unity.

p	N	$\|\mathbf{e}\|_{2,\Omega}$	order	$\|\mathbf{e}-\mathbf{E}^{\perp}\|_{2,\Omega}$	order	$[\min_e \theta_e, \max_e \theta_e]$	θ
0	10^2	$1.380e-01$	–	$1.073e-01$	–	$[0.180, 2.872]$	0.780
	20^2	$7.322e-02$	0.914	$5.760e-02$	0.898	$[0.165, 3.375]$	0.771
	30^2	$4.997e-02$	0.942	$3.957e-02$	0.925	$[0.157, 3.490]$	0.766
1	10^2	$2.062e-03$	–	$1.039e-04$	–	$[0.966, 1.007]$	0.994
	20^2	$5.119e-04$	2.010	$1.478e-05$	2.813	$[0.980, 1.005]$	0.998
	30^2	$2.270e-04$	2.006	$4.772e-06$	2.789	$[0.986, 1.004]$	0.999
2	10^2	$1.059e-05$	–	$9.583e-07$	–	$[0.966, 1.015]$	1.002
	20^2	$1.331e-06$	2.992	$6.093e-08$	3.975	$[0.985, 1.009]$	1.002
	30^2	$3.953e-07$	2.995	$1.211e-08$	3.984	$[0.991, 1.006]$	1.001
3	10^2	$1.008e-07$	–	$4.781e-09$	–	$[0.978, 1.006]$	0.999
	20^2	$6.282e-09$	4.004	$1.512e-10$	4.982	$[0.992, 1.003]$	1.000
	30^2	$1.240e-09$	4.002	$2.000e-11$	4.989	$[0.996, 1.002]$	1.000

Table 3.3.4: L^2 errors $\|\mathbf{e}\|_{2,\Omega}$, $\|\mathbf{e}-\mathbf{E}^{\perp}\|_{2,\Omega}$ and their order of convergence. Maximum, minimum local and global effectivity indices for Example 3.3.4 at $t=1$ using $\Pi\mathbf{u}_0$.

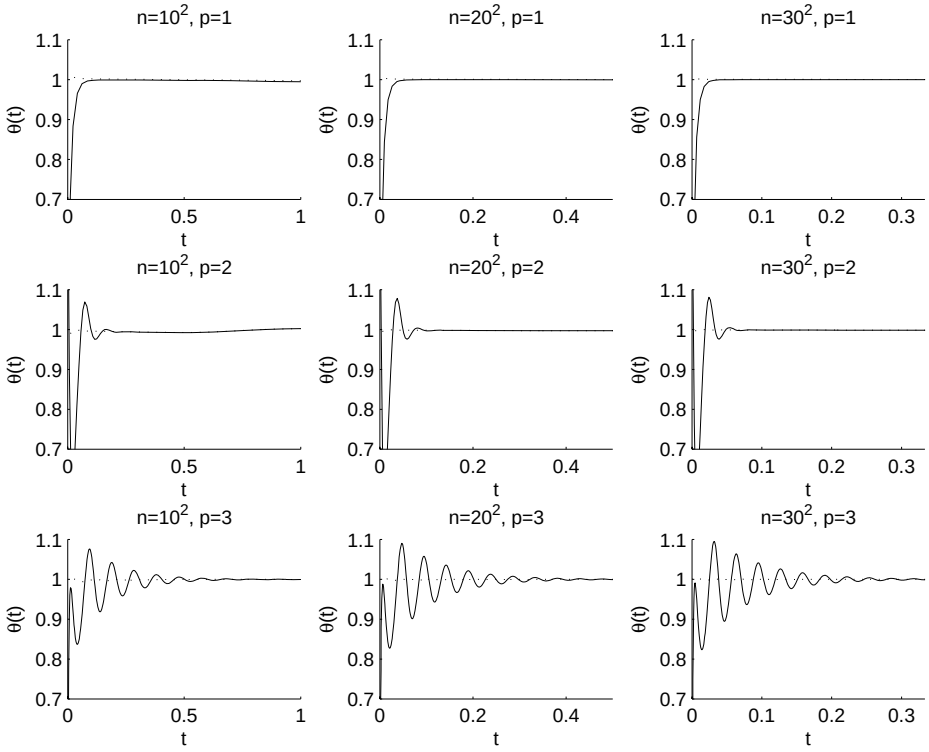

Figure 3.3.4: Global effectivity indices versus $0 \le t \le \frac{10}{h}$ using $\pi\mathbf{u}_0$ (dotted) and $\Pi\mathbf{u}_0$ (solid) for Example 3.3.4.

Chapter 4

Error Analysis for Linear Symmetric Hyperbolic Systems, Revisited

In this chapter we discuss general multi-dimensional symmetric hyperbolic systems. We will investigate the asymptotic behavior of the DG method on one element $\omega = (0, h)^d$.

Thus, let $\mathbf{u} \in [C^2([0,T], C^{p+2}(\bar{\Omega}))]^m$ be the true solution of the linear symmetric hyperbolic system

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{i=1}^{d} \mathbf{A}_i \frac{\partial \mathbf{u}}{\partial x_i} = \mathbf{g}(t, \mathbf{x}), \quad \mathbf{x} \in \Omega,\ 0 < t < T, \tag{4.0.1a}$$

subject to the initial and boundary conditions

$$\mathbf{u}(0, \mathbf{x}) = \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{4.0.1b}$$

$$\left(\sum_{i=1}^{d} \nu_i \mathbf{A}_i^{\bar{\mu}_i} \right) \mathbf{u}(t, \mathbf{x}) = \left(\sum_{i=1}^{d} \nu_i \mathbf{A}_i^{\bar{\mu}_i} \right) \mathbf{u}_B(t, \mathbf{x}), \quad \mathbf{x} \in \partial\Omega,\ 0 < t < T. \tag{4.0.1c}$$

The DG method on ω consists of finding $\mathbf{u}_h \in \mathcal{P}_p$ that satisfies

$$\int_\omega \mathbf{v}^t \left(\frac{\partial \mathbf{u}_h}{\partial t} - \mathbf{g} \right) d\mathbf{x} = \sum_{j=1}^{d} \left(\int_\omega \frac{\partial \mathbf{v}^t}{\partial x_j} \mathbf{A}_j \mathbf{u}_h \, d\mathbf{x} - \int_{\partial\omega} \mathbf{v}^t \nu_j (\mathbf{A}_j^{\mu_j} \mathbf{u}_h^+ + \mathbf{A}_j^{\bar{\mu}_j} \mathbf{u}_h^-) \, d\mathbf{s} \right),$$
$$\forall\ \mathbf{v} \in \mathcal{P}_p,\ 0 < t < T,$$

subject to the initial and boundary conditions

$$\mathbf{u}_h(0, \mathbf{x}) = \pi \mathbf{u}_0(\mathbf{x}) \text{ or } \mathbf{u}_h(0, \mathbf{x}) = \Pi \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \omega,$$
$$(\nu_i \mathbf{A}_i^{\bar{\mu}_i}) \mathbf{u}_h^-(t, \mathbf{x}) = (\nu_i \mathbf{A}_i^{\bar{\mu}_i}) \pi_i \mathbf{u}(t, \mathbf{x}), \quad \mathbf{x} \in \gamma_i,\ 0 < t < T,\ 1 \le i \le d.$$

where $\mathbf{u} = \mathbf{u}_B$ on the boundary of Ω.

We will perform a local error analysis by writing the local error as a series and show that its leading term can be expressed as a linear combination of Legendre polynomials of degree p and $p+1$, even if the conditions of §3 are not satisfied. For special hyperbolic systems where the coefficient matrices are nonsingular we show that the leading term of the error is spanned by $(p+1)^{th}$-degree Radau polynomials. We again apply these asymptotic results to observe that projections of the error are pointwise superconvergent in some cases and establish superconvergence results for some integrals of the error. We solve relatively small local problems to compute efficient and asymptotically exact estimates of the finite element error. Finally, we present computational results in two- and three space dimensions for the wave equation and Maxwell's equations.

4.1 Preliminary Results

To show the asymptotic behavior of the error, we define

$$\bar{\mathcal{P}_p} = \left\{ \mathbf{v}(\boldsymbol{\xi}) \in \mathcal{P}_p : \sum_{i=1}^{d} \mathbf{A}_i \frac{\partial \mathbf{v}}{\partial \xi_i} = \mathbf{0} \text{ on } \Delta,\ \mathbf{A}_i \mathbf{v} = \mathbf{0} \text{ on } \Gamma_i,\ 1 \leq i \leq d \right\}, \tag{4.1.1a}$$

and the orthogonal complement of $\bar{\mathcal{P}_p}$ in $[L^2(\Delta)]^m$, defined by

$$\bar{\mathcal{P}_p}^{\perp} = \left\{ \mathbf{w}(\boldsymbol{\xi}) \in [L^2(\Delta)]^m : \int_{\Delta} \mathbf{v}^t \mathbf{w}\, d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \bar{\mathcal{P}_p} \right\}. \tag{4.1.1b}$$

Let $\Gamma_i(a)$ be a hyperplane through $\bar{\Delta}$, defined by

$$\Gamma_i(a) = \{ \boldsymbol{\xi} \in \bar{\Delta} : \xi_i = a \},\ a \in [0,1]. \tag{4.1.2}$$

Note that $\Gamma_i(0) = \Gamma_i^-$ and $\Gamma_i(1) = \Gamma_i^+$, as defined in 2.2.1.

We can show the following properties of $\bar{\mathcal{P}_p}$:

Lemma 4.1.1. *For all integrable functions* $\mathbf{f} : (0,1) \to \mathbb{R}^m$ *we have*

$$\int_{\Delta} \mathbf{v}^t \mathbf{A}_i \mathbf{f}(\xi_i)\, d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \bar{\mathcal{P}_p},\ 1 \leq i \leq d. \tag{4.1.3}$$

Proof. Let $1 \leq i \leq d$ and $\mathbf{v} \in \bar{\mathcal{P}_p}$ and define the function

$$\mathbf{h}(\xi_i) = \int_{\Gamma_i(\xi_i)} \mathbf{A}_i \mathbf{v}\, d\boldsymbol{\sigma},\ \xi_i \in [0,1]. \tag{4.1.4}$$

By the definition of $\bar{\mathcal{P}_p}$, $\mathbf{A}_j \mathbf{v} = \mathbf{0}$ on Γ_j for all $1 \leq j \leq d$. The divergence theorem yields

$$\int_{\Gamma_i(\xi_i)} \mathbf{A}_j \frac{\partial \mathbf{v}}{\partial \xi_j}\, d\boldsymbol{\sigma} = \int_{\Gamma_j^+ \cap \Gamma_i(\xi_i)} \mathbf{A}_j \mathbf{v}\, d\boldsymbol{\sigma} - \int_{\Gamma_j^- \cap \Gamma_i(\xi_i)} \mathbf{A}_j \mathbf{v}\, d\boldsymbol{\sigma} = \mathbf{0},\ j \in D(i),\ \xi_i \in [0,1]. \tag{4.1.5}$$

We note that by the definition of $\bar{\mathcal{P}}_p$, $\sum_{j=1}^d \mathbf{A}_j \frac{\partial \mathbf{v}}{\partial \xi_j} = \mathbf{0}$ on Ω, which combined with (4.1.5) yields

$$\frac{d}{d\xi_i}\mathbf{h}(\xi_i) = \int_{\Gamma_i(\xi_i)} \mathbf{A}_i \frac{\partial \mathbf{v}}{\partial \xi_i}\, d\boldsymbol{\sigma} = -\sum_{j\in D(i)} \int_{\Gamma_i(\xi_i)} \mathbf{A}_j \frac{\partial \mathbf{v}}{\partial \xi_j}\, d\boldsymbol{\sigma} = \mathbf{0},\ \xi_i \in (0,1). \tag{4.1.6}$$

Since $\mathbf{A}_i\mathbf{v} = \mathbf{0}$ on Γ_i, we obtain

$$\mathbf{h}(0) = \int_{\Gamma_i^-} \mathbf{A}_i\mathbf{v}\, d\boldsymbol{\sigma} = \mathbf{0}, \tag{4.1.7}$$

which together with the Fundamental Theorem of Calculus and (4.1.6) yields

$$\mathbf{h}(\xi_i) = \mathbf{h}(0) + \int_0^{\xi_i} \frac{d}{d\hat{\xi}_i}\mathbf{h}(\hat{\xi}_i)\, d\hat{\xi}_i = \mathbf{0},\ \xi_i \in [0,1]. \tag{4.1.8}$$

Since $\mathbf{f}(\xi_i)$ is constant on $\Gamma_i(\xi_i)$, and $\mathbf{A}_i$ is symmetric, we obtain

$$\int_{\Gamma_i(\xi_i)} \mathbf{v}^t\mathbf{A}_i\mathbf{f}(\xi_i)\, d\boldsymbol{\sigma} = \int_{\Gamma_i(\xi_i)} \mathbf{f}^t(\xi_i)\mathbf{A}_i\mathbf{v}\, d\boldsymbol{\sigma} = \mathbf{f}^t(\xi_i)\mathbf{h}(\xi_i) = 0,\ \xi_i \in [0,1]. \tag{4.1.9}$$

Integrating (4.1.9) over ξ_i from 0 to 1 yields (4.1.3). □

Lemma 4.1.2. *Let* $\mathbf{d} \in \bigoplus_{k=1}^d \mathcal{R}(\mathbf{A}_k)$. *Then*

$$\int_\Delta \mathbf{v}^t\mathbf{d}L_p(\xi_i)\, d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \bar{\mathcal{P}}_p,\ 1 \le i \le d. \tag{4.1.10}$$

Proof. Since $\mathbf{d} \in \bigoplus_{k=1}^d \mathcal{R}(\mathbf{A}_k)$, it can be written as

$$\mathbf{d} = \sum_{k=1}^d \mathbf{A}_k\mathbf{d}_k. \tag{4.1.11}$$

By Lemma 4.1.1 we know that

$$\int_\Delta \mathbf{v}^t\mathbf{A}_i\mathbf{d}_i L_p(\xi_i)\, d\boldsymbol{\xi} = 0,\ 1 \le i \le d. \tag{4.1.12}$$

Now we will show that

$$\int_\Delta \mathbf{v}^t\mathbf{A}_k\mathbf{d}_k L_p(\xi_i)\, d\boldsymbol{\xi} = 0,\ k \in D(i),\ 1 \le i \le d. \tag{4.1.13}$$

First we define the auxiliary function

$$\tilde{\mathbf{h}}_k(\xi_k) = \int_{\Gamma_k(\xi_k)} \mathbf{A}_k\mathbf{v}L_p(\xi_i)\, d\boldsymbol{\sigma},\ \xi_k \in [0,1]. \tag{4.1.14}$$

By the definition of $\bar{\mathcal{P}}_p$, $\mathbf{A}_k \frac{\partial \mathbf{v}}{\partial \xi_k} = -\sum_{j \in D(k)} \mathbf{A}_j \frac{\partial \mathbf{v}}{\partial \xi_j}$, which yields

$$\frac{d}{d\xi_k}\tilde{\mathbf{h}}_k(\xi_k) = \int_{\Gamma_k(\xi_k)} \mathbf{A}_k \frac{\partial \mathbf{v}}{\partial \xi_k} L_p(\xi_i)\, d\boldsymbol{\sigma} = -\sum_{j \in D(k)} \int_{\Gamma_k(\xi_k)} \mathbf{A}_j \frac{\partial \mathbf{v}}{\partial \xi_j} L_p(\xi_i)\, d\boldsymbol{\sigma}, \quad \xi_k \in (0,1). \tag{4.1.15}$$

We will show that $\frac{d}{d\xi_k}\tilde{\mathbf{h}}_k(\xi_k) = 0,\ \xi_k \in [0,1]$.

Since $\mathbf{v} \in \bar{\mathcal{P}}_p \subset \mathcal{P}_p$, we obtain by the orthogonality properties of Legendre polynomials that

$$\int_{\Gamma_k(\xi_k)} \mathbf{A}_i \frac{\partial \mathbf{v}}{\partial \xi_i} L_p(\xi_i)\, d\boldsymbol{\sigma} = \mathbf{0}, \quad \xi_k \in [0,1]. \tag{4.1.16}$$

By the definition of $\bar{\mathcal{P}}_p$, $\mathbf{A}_j \mathbf{v} = \mathbf{0}$ on Γ_j for all $1 \le j \le d$, which together with the divergence theorem yields

$$\int_{\Gamma_k(\xi_k)} \mathbf{A}_j \frac{\partial \mathbf{v}}{\partial \xi_j} L_p(\xi_i)\, d\boldsymbol{\sigma} = \int_{\Gamma_k(\xi_k)\cap\Gamma_j} \nu_j \mathbf{A}_j \mathbf{v} L_p(\xi_i)\, d\boldsymbol{\sigma} = \mathbf{0}, \quad j \in D(k) \cap D(i),\ \xi_k \in [0,1], \tag{4.1.17}$$

where Γ_j is defined in (2.2.1).

Substituting (4.1.16) and (4.1.17) into (4.1.15) yields

$$\frac{d}{d\xi_k}\tilde{\mathbf{h}}_k(\xi_k) = \mathbf{0}, \quad \xi_k \in (0,1). \tag{4.1.18}$$

Since $\mathbf{A}_k \mathbf{v} = \mathbf{0}$ on Γ_k and $\Gamma_k^- = \Gamma_k(0)$, we obtain

$$\tilde{\mathbf{h}}_k(0) = \int_{\Gamma_k^-} \mathbf{A}_k \mathbf{v} L_p(\xi_i)\, d\boldsymbol{\sigma} = 0, \tag{4.1.19}$$

which together with the Fundamental Theorem of Calculus and (4.1.18) yields

$$\tilde{\mathbf{h}}_k(\xi_k) = \mathbf{0}, \quad \xi_k \in [0,1]. \tag{4.1.20}$$

Since $\mathbf{d}_k$ is constant and $\mathbf{A}_k$ is symmetric, we obtain

$$\int_{\Gamma_k(\xi_k)} \mathbf{v}^t \mathbf{A}_k \mathbf{d}_k L_p(\xi_i)\, d\boldsymbol{\sigma} = \int_{\Gamma_k(\xi_k)} \mathbf{d}_k^t \mathbf{A}_k \mathbf{v} L_p(\xi_i)\, d\boldsymbol{\sigma} = \mathbf{d}_k^t \tilde{\mathbf{h}}_k(\xi_k) = 0,\ \xi_k \in [0,1]. \tag{4.1.21}$$

Integrating (4.1.21) from $\xi_k = 0$ to 1 yields (4.1.13).

Combining (4.1.12) and (4.1.13) yields

$$\int_\Delta \mathbf{v}^t \mathbf{d} L_p(\xi_i)\, d\boldsymbol{\xi} = \sum_{k=1}^{d} \int_\Delta \mathbf{v}^t \mathbf{A}_k \mathbf{d}_k L_p(\xi_i)\, d\boldsymbol{\xi} = 0. \tag{4.1.22}$$

This concludes the proof. □

Lemmas 4.1.1 and 4.1.2 can now be used to show the next lemma.

Lemma 4.1.3. *Let* $\mathbf{c} \in \mathbb{R}^m,\ \mathbf{d} \in \bigoplus_{k=1}^d \mathcal{R}(\mathbf{A}_k)$. *Then*

$$\int_\Delta \mathbf{v}^t \left(L_{p+1}(\xi_i)\mathbf{c} - L_p(\xi_i)(sgn(\mathbf{A}_i)\mathbf{c} + \mathbf{d})\right)\, d\boldsymbol{\xi} = 0, \quad \forall\ \mathbf{v} \in \bar{\mathcal{P}}_p,\ 1 \le i \le d. \tag{4.1.23}$$

Proof. Since $\mathbf{A}_i\mathbf{A}_i^\dagger$ is the identity on $\mathcal{R}(\mathbf{A}_i)$ by Lemma 1.2.13 and $\mathcal{R}(\mathbf{A}_i) = \mathcal{R}(\mathrm{sgn}(\mathbf{A}_i))$ by property (1.2.25c), we obtain

$$\mathrm{sgn}(\mathbf{A}_i) = \mathbf{A}_i\mathbf{A}_i^\dagger \mathrm{sgn}(\mathbf{A}_i), \quad 1 \le i \le d, \tag{4.1.24}$$

which, when combined with Lemma 4.1.1, yields

$$\int_\Delta \mathbf{v}^t \mathrm{sgn}(\mathbf{A}_i)\mathbf{c} L_p(\xi_i)\, d\boldsymbol{\xi} = \int_\Delta \mathbf{v}^t \mathbf{A}_i \left(\mathbf{A}_i^\dagger \mathrm{sgn}(\mathbf{A}_i)\mathbf{c} L_p(\xi_i)\right)\, d\boldsymbol{\xi} = 0, \quad \forall\ \mathbf{v} \in \bar{\mathcal{P}}_p,\ 1 \le i \le d. \tag{4.1.25}$$

Lemma 4.1.2 and the orthogonality properties (2.2.6) combined yield

$$\int_\Delta \mathbf{v}^t \left(L_{p+1}(\xi_i)\mathbf{c} - L_p(\xi_i)\mathbf{d}\right)\, d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \bar{\mathcal{P}}_p, \quad 1 \le i \le d. \tag{4.1.26}$$

Adding (4.1.25) and (4.1.26) yields (4.1.23). □

We also need the following lemma:

Lemma 4.1.4. *If* $\mathbf{q} \in \mathcal{P}_p$ *satisfies*

$$\sum_{i=1}^d \int_\Delta \left(\frac{\partial \mathbf{v}^t}{\partial \xi_i}\mathbf{A}_i\mathbf{q}\, d\boldsymbol{\xi} - \int_{\Gamma_i} \mathbf{v}^t \nu_i \mathbf{A}_i^{\mu_i}\mathbf{q}\, d\boldsymbol{\sigma}\right) = 0,\ \mathbf{v} \in \mathcal{P}_p, \tag{4.1.27}$$

then $\mathbf{q} \in \bar{\mathcal{P}}_p$.

Proof. First we integrate (4.1.27) by parts to obtain

$$\sum_{i=1}^d \left(-\int_\Delta \mathbf{v}^t \mathbf{A}_i \frac{\partial \mathbf{q}}{\partial \xi_i}\, d\boldsymbol{\xi} + \int_{\Gamma_i} \mathbf{v}^t \nu_i \mathbf{A}_i^{\bar{\mu}_i}\mathbf{q}\, d\boldsymbol{\sigma}\right) = 0,\ \mathbf{v} \in \mathcal{P}_p. \tag{4.1.28}$$

Adding (4.1.27) to (4.1.28), testing against $\mathbf{v} = -\mathbf{q}$, and using the symmetry of $\mathbf{A}_i$, $1 \le i \le d$, we obtain

$$\sum_{i=1}^d \int_{\Gamma_i} \mathbf{q}^t \nu_i (\mathbf{A}_i^{\mu_i} - \mathbf{A}_i^{\bar{\mu}_i})\mathbf{q}\, d\boldsymbol{\sigma} = \sum_{i=1}^d \int_{\Gamma_i} \mathbf{q}^t (\mathbf{A}_i^+ - \mathbf{A}_i^-)\mathbf{q}\, d\boldsymbol{\sigma} = 0. \tag{4.1.29}$$

$(\mathbf{A}_i^+ - \mathbf{A}_i^-)$ is symmetric positive semi-definite by (1.2.25f), and therefore admits a Cholesky factorization $(\mathbf{A}_i^+ - \mathbf{A}_i^-) = \mathbf{L}_i^t \mathbf{L}_i$. Hence (4.1.29) can be written as

$$\sum_{i=1}^{d} \int_{\Gamma_i} \|\mathbf{L}_i \mathbf{q}\|^2 \, d\boldsymbol{\sigma} = 0. \tag{4.1.30}$$

Therefore, $\mathbf{L}_i \mathbf{q} = \mathbf{0}$ on Γ_i, which yields

$$\mathbf{L}_i^t (\mathbf{L}_i \mathbf{q}) = (\mathbf{A}_i^+ - \mathbf{A}_i^-)\mathbf{q} = \mathbf{0}, \ \boldsymbol{\xi} \in \Gamma_i, \ 1 \le i \le d, \tag{4.1.31}$$

which combined with property (1.2.25b) leads to

$$\mathbf{A}_i^s \mathbf{v} = \mathbf{0}, \ \boldsymbol{\xi} \in \Gamma_i, \ s = +, -, \ 1 \le i \le d. \tag{4.1.32}$$

By (4.1.32), the boundary integral in (4.1.28) vanishes. Thus, (4.1.28) yields for $\mathbf{v} = \sum_{i=1}^{d} \mathbf{A}_i \frac{\partial \mathbf{q}}{\partial \xi_i}$

$$-\int_{\Delta} \left\| \sum_{i=1}^{d} \mathbf{A}_i \frac{\partial \mathbf{q}}{\partial \xi_i} \right\|^2 d\boldsymbol{\xi} = 0, \tag{4.1.33}$$

which in turn yields

$$\sum_{i=1}^{d} \mathbf{A}_i \frac{\partial \mathbf{q}}{\partial \xi_i} = \mathbf{0}, \ \boldsymbol{\xi} \in \Delta. \tag{4.1.34}$$

Combining (4.1.32) and (4.1.34) proves the lemma. □

4.2 Local Error Analysis

Now we are ready to state a theorem for the local error.

Theorem 4.2.1. *Let $\mathbf{u} \in [C^2([0,T], C^{p+2}(\bar{\omega}))]^m$ be the solution of (4.0.1) and let $\mathbf{u}_h \in \mathcal{P}_p$ satisfy (4.0.2). Then the local finite element error on ω, at $t = \mathcal{O}(1)$ and for $p \ge 1$, can be written as*

$$\mathbf{u}(t, h\boldsymbol{\xi}) - \mathbf{u}_h(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} \mathbf{r}_i(t, h\xi_i) + \mathcal{O}(h^{p+2}), \ \boldsymbol{\xi} \in \Delta, \tag{4.2.1a}$$

where

$$\mathbf{r}_i(t, h\xi_i) = L_{p+1}(\xi_i)\mathbf{c}_i(t) - L_p(\xi_i)(sgn(\mathbf{A}_i)\mathbf{c}_i(t) + \mathbf{d}_i(t)), \ 1 \le i \le d, \tag{4.2.1b}$$

with

$$\mathbf{c}_i(t) = \frac{1}{a_{p+1}} \frac{1}{(p+1)!} \frac{\partial^{p+1} \mathbf{u}(t, \mathbf{0})}{\partial x_i^{p+1}}, \quad \mathbf{d}_i(t) \in \mathcal{N}(\mathbf{A}_i) \cap \bigoplus_{k=1}^{d} \mathcal{R}(\mathbf{A}_k). \tag{4.2.1c}$$

Proof. First we derive the *orthogonality condition* for the error $\mathbf{e} = \mathbf{u} - \mathbf{u}_h$. By (4.0.1a), $\mathbf{u}$ satisfies

$$\int_\omega \mathbf{v}^t \left(\frac{\partial \mathbf{u}}{\partial t} - \mathbf{g} \right) d\mathbf{x} = \sum_{i=1}^{d} \left(\int_\omega \frac{\partial \mathbf{v}^t}{\partial x_i} \mathbf{A}_i \mathbf{u} \, d\mathbf{x} - \int_{\gamma_i} \mathbf{v}^t \nu_i \mathbf{A}_i \mathbf{u} \, d\mathbf{s} \right), \ \forall \ \mathbf{v} \in \mathcal{P}_p, \ 0 < t < T. \tag{4.2.2}$$

Subtracting (4.0.2) from (4.2.2) we obtain

$$\int_\omega \mathbf{v}^t \frac{\partial \mathbf{e}}{\partial t} \, d\mathbf{x} = \sum_{i=1}^{d} \left(\int_\omega \frac{\partial \mathbf{v}^t}{\partial x_i} \mathbf{A}_i \mathbf{e} \, d\mathbf{x} - \int_{\gamma_i} \mathbf{v}^t \nu_i (\mathbf{A}_i^{\mu_i} \mathbf{e} + \mathbf{A}_i^{\bar{\mu}_i} \mathbf{e}^-) \, d\mathbf{s} \right), \ \forall \ \mathbf{v} \in \mathcal{P}_p, \ 0 < t < T. \tag{4.2.3}$$

Apply the scalings $\tau = T^{-1}t$ and $\boldsymbol{\xi} = h^{-1}\mathbf{x}$ and write $\hat{\mathbf{e}}(\tau, \boldsymbol{\xi}) = \mathbf{e}(T\tau, h\boldsymbol{\xi})$ to obtain the orthogonality condition

$$\frac{h}{T} \int_\Delta \mathbf{v}^t \frac{\partial \hat{\mathbf{e}}}{\partial \tau} \, d\boldsymbol{\xi} = \sum_{i=1}^{d} \left(\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \hat{\mathbf{e}} \, d\boldsymbol{\xi} - \int_{\Gamma_i} \mathbf{v}^t \nu_i (\mathbf{A}_i^{\mu_i} \hat{\mathbf{e}} + \mathbf{A}_i^{\bar{\mu}_i} \hat{\mathbf{e}}^-) \, d\boldsymbol{\sigma} \right), \ \mathbf{v} \in \mathcal{P}_p, \ 0 < \tau < 1. \tag{4.2.4}$$

Now note that, since $\mathcal{P}_p$ is a subspace of $[L^2(\Delta)]^m$, we can split $\hat{\mathbf{e}}$ by

$$\hat{\mathbf{e}} = \bar{\mathbf{e}} + \tilde{\mathbf{e}}, \quad \bar{\mathbf{e}} \in \bar{\mathcal{P}}_p, \quad \tilde{\mathbf{e}} \in \bar{\mathcal{P}}_p{}^{\perp}, \tag{4.2.5}$$

as defined in (4.1.1).

We will first show that

$$\bar{\mathbf{e}}(\tau, \boldsymbol{\xi}) = \mathcal{O}(h^{p+2}), \quad \boldsymbol{\xi} \in \Delta, \ 0 \le \tau \le 1. \tag{4.2.6}$$

Since $\bar{\mathcal{P}}_p$ is a finite dimensional vector space and $\bar{\mathbf{e}} \in \bar{\mathcal{P}}_p$, we have

$$\frac{\partial \bar{\mathbf{e}}}{\partial \tau}(\tau, \boldsymbol{\xi}) = \lim_{h \to 0} \frac{\bar{\mathbf{e}}(\tau + h, \boldsymbol{\xi}) - \bar{\mathbf{e}}(\tau, \boldsymbol{\xi})}{h} \in \bar{\mathcal{P}}_p, \tag{4.2.7a}$$

$$\frac{\partial \tilde{\mathbf{e}}}{\partial \tau}(\tau, \boldsymbol{\xi}) = \lim_{h \to 0} \frac{\tilde{\mathbf{e}}(\tau + h, \boldsymbol{\xi}) - \tilde{\mathbf{e}}(\tau, \boldsymbol{\xi})}{h} \in \bar{\mathcal{P}}_p{}^{\perp}, \quad \boldsymbol{\xi} \in \Delta, \ 0 < \tau < 1. \tag{4.2.7b}$$

By the definition of $\bar{\mathcal{P}}_p$ in (4.1.1) and the symmetry of $\mathbf{A}_i$, $\mathbf{A}_i^+$, and $\mathbf{A}_i^-$, $1 \le i \le d$, (4.2.4) yields for $\mathbf{v} \in \bar{\mathcal{P}}_p$

$$\begin{aligned} \frac{h}{T} \int_\Delta \mathbf{v}^t \frac{\partial \hat{\mathbf{e}}}{\partial \tau} \, d\boldsymbol{\xi} &= \sum_{i=1}^{d} \left(\int_\Delta \left(\mathbf{A}_i \frac{\partial \mathbf{v}}{\partial \xi_i} \right)^t \hat{\mathbf{e}} \, d\boldsymbol{\xi} - \int_{\Gamma_i} \left(\mathbf{A}_i^{\mu_i} \mathbf{v} \right)^t \nu_i \hat{\mathbf{e}} + \left(\mathbf{A}_i^{\bar{\mu}_i} \mathbf{v} \right)^t \nu_i \hat{\mathbf{e}}^- \, d\boldsymbol{\sigma} \right) \\ &= 0, \quad \forall \ \mathbf{v} \in \bar{\mathcal{P}}_p, \quad 0 < \tau < 1. \end{aligned} \tag{4.2.8}$$

Thus, $\frac{\partial \hat{\mathbf{e}}}{\partial \tau} \in \bar{\mathcal{P}}_p{}^{\perp}$, which combined with (4.2.5) and (4.2.7b) yields

$$\frac{\partial \bar{\mathbf{e}}}{\partial \tau} = \frac{\partial \hat{\mathbf{e}}}{\partial \tau} - \frac{\partial \tilde{\mathbf{e}}}{\partial \tau} \in \bar{\mathcal{P}}_p{}^{\perp}, \quad 0 < \tau < 1. \tag{4.2.9}$$

Then (4.2.9) and (4.2.7a) together yield

$$\frac{\partial \bar{\mathbf{e}}}{\partial \tau}(\tau, \boldsymbol{\xi}) = \mathbf{0}, \quad \boldsymbol{\xi} \in \Delta, \quad 0 < \tau < 1. \tag{4.2.10}$$

By Lemma 2.2.4, the initial conditions satisfy either

$$\hat{\mathbf{e}}(0, \boldsymbol{\xi}) = \mathbf{u}_0(h\boldsymbol{\xi}) - \Pi \mathbf{u}_0(h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} L_{p+1}(\xi_i)\mathbf{c}_i + \mathcal{O}(h^{p+2}), \quad \boldsymbol{\xi} \in \Delta, \tag{4.2.11a}$$

or

$$\hat{\mathbf{e}}(0, \boldsymbol{\xi}) = \mathbf{u}_0(h\boldsymbol{\xi}) - \pi \mathbf{u}_0(h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} \mathbf{r}_i(0, \xi_i) + \mathcal{O}(h^{p+2}), \quad \boldsymbol{\xi} \in \Delta. \tag{4.2.11b}$$

By the orthogonality properties (2.2.6) and Lemma 4.1.3, we have

$$L_{p+1}(\xi_i)\mathbf{c}_i \in \bar{\mathcal{P}_p}^{\perp}, \quad \mathbf{r}_i(0, \xi_i) \in \bar{\mathcal{P}_p}^{\perp}, \quad 1 \leq i \leq d, \tag{4.2.12}$$

which, when combined with (4.2.11) yields

$$\bar{\mathbf{e}}(0, \boldsymbol{\xi}) = \mathcal{O}(h^{p+2}), \quad \boldsymbol{\xi} \in \Delta. \tag{4.2.13}$$

By the Fundamental Theorem of Calculus, (4.2.13) and (4.2.10) yields (4.2.6).

In the remainder of the proof, we will investigate the asymptotic behavior of $\tilde{\mathbf{e}}$. We write the Maclaurin series of $\hat{\mathbf{e}}$ with respect to the mesh parameter h as

$$\hat{\mathbf{e}}(\tau, \boldsymbol{\xi}) = \sum_{k=0}^{p+1} h^k \mathbf{q}_k(\tau, \boldsymbol{\xi}) + \mathcal{O}(h^{p+2}), \quad \boldsymbol{\xi} \in \Delta, \ 0 < \tau < 1, \tag{4.2.14}$$

where, since $\mathbf{u}_h$ is a function of $T\tau$, $h\boldsymbol{\xi}$, and h,

$$\mathbf{q}_k(\tau, \boldsymbol{\xi}) = \frac{1}{k!} \frac{d^k(\mathbf{u}(T\tau, h\boldsymbol{\xi}) - \mathbf{u}_h(T\tau, h\boldsymbol{\xi}, h)}{dh^k}\bigg|_{h=0}. \tag{4.2.15}$$

We write the Maclaurin series of $\tilde{\mathbf{e}} \in \bar{\mathcal{P}_p}^{\perp}$ with respect to the mesh parameter h as

$$\tilde{\mathbf{e}}(\tau, \boldsymbol{\xi}) = \sum_{k=0}^{\infty} h^k \tilde{\mathbf{q}}_k(\tau, \boldsymbol{\xi}), \quad \tilde{\mathbf{q}}_k \in \bar{\mathcal{P}_p}^{\perp}, \ \boldsymbol{\xi} \in \Delta, \ 0 < \tau < 1. \tag{4.2.16}$$

By (4.2.5) and (4.2.13), $\hat{\mathbf{e}} = \tilde{\mathbf{e}} + \mathcal{O}(h^{p+2})$, thus subtracting (4.2.14) from (4.2.16) and setting all terms having the same power of h equal yields

$$\mathbf{q}_k = \tilde{\mathbf{q}}_k \in \bar{\mathcal{P}_p}^{\perp}, \ 0 \leq k \leq p+1. \tag{4.2.17}$$

Let $\hat{\mathbf{r}}_i(\tau,\xi_i) = \mathbf{r}_i(T\tau, h\xi_i)$, $1 \le i \le d$. By Lemma 2.2.4, the boundary conditions satisfy

$$\begin{aligned}\hat{\mathbf{e}}^-(\tau,\boldsymbol{\xi}) &= \mathbf{u}(t,h\boldsymbol{\xi}) - \mathbf{u}_h^-(t,h\boldsymbol{\xi})\\ &= h^{p+1}\sum_{j\in D(i)} \hat{\mathbf{r}}_j(\tau,\xi_j) + \mathcal{O}(h^{p+2}),\ \boldsymbol{\xi}\in\Gamma_i,\ 1\le i\le d. \end{aligned} \tag{4.2.18}$$

Substituting (4.2.14) and (4.2.18) in (4.2.4) yields

$$\begin{aligned}&\sum_{k=0}^{p+1} h^k \left(\frac{h}{T}\int_\Delta \mathbf{v}^t \frac{\partial \mathbf{q_k}}{\partial\tau}\,d\boldsymbol{\xi} - \sum_{i=1}^d\left(\int_\Delta \frac{\partial \mathbf{v}^t}{\partial\xi_i}\mathbf{A}_i\mathbf{q}_k\,d\boldsymbol{\xi} + \int_{\Gamma_i}\mathbf{v}^t\nu_i\mathbf{A}_i^{\mu_i}\mathbf{q}_k\,d\boldsymbol{\sigma}\right)\right)\\ &\qquad = -h^{p+1}\sum_{i=1}^d\int_{\Gamma_i}\mathbf{v}^t\nu_i\mathbf{A}_i^{\bar{\mu}_i}\sum_{j\in D(i)}\hat{\mathbf{r}}_j\,d\boldsymbol{\sigma} + \mathcal{O}(h^{p+2}),\quad \mathbf{v}\in\mathcal{P}_p. \end{aligned}\tag{4.2.19}$$

Now assume that $T = \mathcal{O}(1)$ and set to zero all terms in (4.2.19) having the same power of h. The $\mathcal{O}(1)$ term $\mathbf{q}_0$ satisfies the orthogonality condition

$$\sum_{i=1}^d\left(\int_\Delta \frac{\partial\mathbf{v}^t}{\partial\xi_i}\mathbf{A}_i\mathbf{q}_0\,d\boldsymbol{\xi} - \int_{\Gamma_i}\mathbf{v}^t\nu_i\mathbf{A}_i^{\mu_i}\mathbf{q}_0\,d\boldsymbol{\sigma}\right) = 0,\ \mathbf{v}\in\mathcal{P}_p. \tag{4.2.20}$$

Lemma 4.1.4 yields $\mathbf{q}_0 \in \bar{\mathcal{P}}_p$, which combined with (4.2.17) shows that $\mathbf{q}_0 = \mathbf{0}$ on Δ.

Assume that $\mathbf{q}_j = \mathbf{0}$ for all $0 \le j \le k-1$, where $k \le p$. Thus, the $\mathcal{O}(h^k)$ term is written as

$$\sum_{i=1}^d\left(\int_\Delta \frac{\partial\mathbf{v}^t}{\partial\xi_i}\mathbf{A}_i\mathbf{q}_k\,d\boldsymbol{\xi} - \int_{\Gamma_i}\mathbf{v}^t\nu_i\mathbf{A}_i^{\mu_i}\mathbf{q}_k\,d\boldsymbol{\sigma}\right) = 0,\ \mathbf{v}\in\mathcal{P}_p. \tag{4.2.21}$$

Lemma 4.1.4 yields $\mathbf{q}_k \in \bar{\mathcal{P}}_p$, which combined with (4.2.17) shows that $\mathbf{q}_k = \mathbf{0}$ on Δ for $0 \le k \le p$.

The $\mathcal{O}(h^{p+1})$ term satisfies the orthogonality condition

$$\sum_{i=1}^d\left(\int_\Delta \frac{\partial\mathbf{v}^t}{\partial\xi_i}\mathbf{A}_i\mathbf{q}_{p+1}\,d\boldsymbol{\xi} - \int_{\Gamma_i}\mathbf{v}^t\nu_i\big(\mathbf{A}_i^{\mu_i}\mathbf{q}_{p+1} - \mathbf{A}_i^{\bar{\mu}_i}\sum_{j\in D(i)}\hat{\mathbf{r}}_j\big)\,d\boldsymbol{\sigma}\right) = 0,\ \forall\ \mathbf{v}\in\mathcal{P}_p. \tag{4.2.22}$$

We will first show that $\mathbf{q}_{p+1} = \sum_{i=1}^d \hat{\mathbf{r}}_i + \mathbf{p},\ \mathbf{p}\in\mathcal{P}_p$.

Since $\frac{\partial^{p+1}}{\partial x_i^{p+1}}\mathbf{u}_h = \mathbf{0}$ for $1 \le i \le d$, (4.2.15) yields for $k = p+1$

$$\begin{aligned}\mathbf{q}_{p+1}(\tau,\boldsymbol{\xi}) &= \frac{1}{k!}\frac{d^k(\mathbf{u}-\mathbf{u}_h)(T\tau,\boldsymbol{\xi}h,h)}{dh^k}\bigg|_{h=0} = \sum_{|\boldsymbol{\alpha}|\le p+1}\frac{1}{\boldsymbol{\alpha}!}D^{\boldsymbol{\alpha}}(\mathbf{u}-\mathbf{u}_h)(T\tau,\mathbf{0})\boldsymbol{\xi}^{\boldsymbol{\alpha}}\\ &= \sum_{i=1}^d\frac{1}{(p+1)!}\frac{\partial^{p+1}\mathbf{u}(T\tau,\mathbf{0})}{\partial x_i^{p+1}}\xi_i^{p+1} + \mathbf{p}_1(\tau,\boldsymbol{\xi}), \end{aligned}\tag{4.2.23}$$

where $\mathbf{p}_1(\tau,\boldsymbol{\xi}) \in \mathcal{P}_p$. By the definition of $\mathbf{c}_i$ in (4.2.1c),

$$\begin{aligned} L_{p+1}(\xi_i)\mathbf{c}_i(T\tau) &= \frac{1}{a_{p+1}}\frac{1}{(p+1)!}\frac{\partial^{p+1}\mathbf{u}(T\tau,\mathbf{0})}{\partial x_i^{p+1}}L_{p+1}(\xi_i) \\ &= \frac{1}{(p+1)!}\frac{\partial^{p+1}\mathbf{u}(T\tau,\mathbf{0})}{\partial x_i^{p+1}}\xi_i^{p+1} + \breve{\mathbf{p}}_i(\tau,\boldsymbol{\xi}), \end{aligned} \tag{4.2.24}$$

where $\breve{\mathbf{p}}_i \in \mathcal{P}_p$. Substituting (4.2.24) into (4.2.1b) yields

$$\begin{aligned} \sum_{i=1}^{d}\hat{\mathbf{r}}_i(\tau,\xi_i) &= \sum_{i=1}^{d}(L_{p+1}(\xi_i)\mathbf{c}_i(T\tau) - L_p(\xi_i)(\operatorname{sgn}(\mathbf{A}_i)\mathbf{c}_i(T\tau) + \mathbf{d}_i(T\tau))) \\ &= \sum_{i=1}^{d}\frac{1}{(p+1)!}\frac{\partial^{p+1}\mathbf{u}(T\tau,\mathbf{0})}{\partial x_i^{p+1}}\xi_i^{p+1} + \mathbf{p}_2(\tau,\boldsymbol{\xi}), \end{aligned} \tag{4.2.25}$$

where

$$\mathbf{p}_2 = \sum_{i=1}^{d}(\breve{\mathbf{p}}_i - L_p(\xi_i)(\operatorname{sgn}(\mathbf{A}_i)\mathbf{c}_i + \mathbf{d}_i)) \in \mathcal{P}_p. \tag{4.2.26}$$

Combining (4.2.25) and (4.2.23) yields for $\mathbf{p} = \mathbf{p}_1 - \mathbf{p}_2 \in \mathcal{P}_p$.

$$\mathbf{q}_{p+1}(\tau,\boldsymbol{\xi}) = \sum_{i=1}^{d}\hat{\mathbf{r}}_i(\tau,\xi_i) + \mathbf{p}(\tau,\boldsymbol{\xi}),\ \mathbf{p} \in \mathcal{P}_p. \tag{4.2.27}$$

Substituting (4.2.27) into (4.2.22) yields

$$\begin{aligned} &\sum_{i=1}^{d}\left(\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i}\mathbf{A}_i\mathbf{p}\,d\boldsymbol{\xi} - \int_{\Gamma_i}\mathbf{v}^t\nu_i\mathbf{A}_i^{\mu_i}\mathbf{p}\,d\boldsymbol{\sigma}\right) \\ &= \sum_{i=1}^{d}\left(-\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i}\mathbf{A}_i\sum_{j=1}^{d}\hat{\mathbf{r}}_j\,d\boldsymbol{\xi} + \int_{\Gamma_i}\mathbf{v}^t\nu_i\left(\mathbf{A}_i^{\mu_i}\sum_{j=1}^{d}\hat{\mathbf{r}}_j + \mathbf{A}_i^{\bar{\mu}_i}\sum_{j\in D(i)}\hat{\mathbf{r}}_j\right)d\boldsymbol{\sigma}\right) \\ &= \sum_{i=1}^{d}\left(T_1^i(\mathbf{v}) + \sum_{j\in D(i)}T_2^{i,j}(\mathbf{v}) + T_3^i(\mathbf{v})\right), \quad \forall\ \mathbf{v} \in \mathcal{P}_p, \end{aligned} \tag{4.2.28a}$$

where

$$T_1^i(\mathbf{v}) = -\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i}\mathbf{A}_i\hat{\mathbf{r}}_i\,d\boldsymbol{\xi}, \tag{4.2.28b}$$

$$T_2^{i,j}(\mathbf{v}) = -\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i}\mathbf{A}_i\hat{\mathbf{r}}_j\,d\boldsymbol{\xi} + \int_{\Gamma_i}\mathbf{v}^t\nu_i\mathbf{A}_i\hat{\mathbf{r}}_j\,d\boldsymbol{\sigma}, \tag{4.2.28c}$$

$$T_3^i(\mathbf{v}) = \int_{\Gamma_i}\mathbf{v}^t\nu_i\mathbf{A}_i^{\mu_i}\hat{\mathbf{r}}_i\,d\boldsymbol{\sigma}, \quad j \in D(i),\ 1 \le i \le d. \tag{4.2.28d}$$

We will now show that $T_1^i(\mathbf{v}) = T_2^{i,j}(\mathbf{v}) = T_3^i(\mathbf{v}) = 0$ for all $\mathbf{v} \in \mathcal{P}_p,\ j \in D(i),\ 1 \leq i \leq d$. By the orthogonality properties of Legendre polynomials, we have

$$T_1^i(\mathbf{v}) = -\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \hat{\mathbf{r}}_i \, d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \mathcal{P}_p,\ 1 \leq i \leq d. \tag{4.2.29}$$

Integration (4.2.28c) by parts w.r.t. ξ_i yields

$$T_2^{i,j}(\mathbf{v}) = \int_\Delta \mathbf{v}^t \mathbf{A}_i \frac{\partial \hat{\mathbf{r}}_j}{\partial \xi_i} \, d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \mathcal{P}_p,\ 1 \leq i \leq d,\ j \in D(i), \tag{4.2.30}$$

since $\hat{\mathbf{r}}_j(t, \xi_j)$ is independent of ξ_i for $j \in D(i)$.

Finally, applying $\mathbf{A}_i^{\mu_i}$ to $\hat{\mathbf{r}}_i$ on Γ_i yields

$$\mathbf{A}_i^{\mu_i} \hat{\mathbf{r}}_i\big|_{\Gamma_i^+} = \mathbf{A}_i^+ \hat{\mathbf{r}}_i(\tau, 1), \quad \mathbf{A}_i^{\mu_i} \hat{\mathbf{r}}_i\big|_{\Gamma_i^-} = \mathbf{A}_i^- \hat{\mathbf{r}}_i(\tau, 0), \quad 1 \leq i \leq d. \tag{4.2.31}$$

Using property (1.2.25d), $L_p(0) = (-1)^p$ and $L_p(1) = 1$, we obtain

$$\mathbf{A}_i^+ \hat{\mathbf{r}}_i(\tau, 1) = \mathbf{A}_i^+ \mathbf{c}_i(T\tau) - \mathbf{A}_i^+ \mathrm{sgn}(\mathbf{A}_i) \mathbf{c}_i(T\tau) = \mathbf{0}, \tag{4.2.32a}$$

$$\mathbf{A}_i^- \hat{\mathbf{r}}_i(\tau, 0) = (-1)^{p+1} \mathbf{A}_i^- \mathbf{c}_i(T\tau) - (-1)^p \mathbf{A}_i^- \mathrm{sgn}(\mathbf{A}_i) \mathbf{c}_i(T\tau)) = \mathbf{0},\ 1 \leq i \leq d. \tag{4.2.32b}$$

Thus, we have established that $\mathbf{A}_i^{\mu_i} \hat{\mathbf{r}}_i\big|_{\Gamma_i} = \mathbf{0}$, thus

$$T_3^i(\mathbf{v}) = \int_{\Gamma_i} \mathbf{v}^t \nu_i \mathbf{A}_i^{\mu_i} \hat{\mathbf{r}}_i \, d\boldsymbol{\sigma} = 0, \quad \forall\ \mathbf{v} \in \mathcal{P}_p,\ 1 \leq i \leq d. \tag{4.2.33}$$

Substituting (4.2.29), (4.2.30), and (4.2.33) into (4.2.28a) leads to

$$\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \mathbf{p} \, d\boldsymbol{\xi} - \int_{\Gamma_i} \mathbf{v}^t \nu_i \mathbf{A}_i^{\mu_i} \mathbf{p} \, d\boldsymbol{\sigma} = 0,\ \mathbf{v} \in \mathcal{P}_p, \tag{4.2.34}$$

which combined with Lemma 4.1.4 yields

$$\mathbf{p} \in \bar{\mathcal{P}}_p. \tag{4.2.35}$$

On the other hand, by Lemma 4.1.3, (4.2.17) and (4.2.27) we obtain

$$\mathbf{p} = \mathbf{q}_{p+1} - \sum_{i=1}^d \hat{\mathbf{r}}_i \in \bar{\mathcal{P}}_p{}^\perp. \tag{4.2.36}$$

Combining (4.2.35) and (4.2.36) yields $\mathbf{p} = \mathbf{0}$ on Δ, which by (4.2.27) leads to

$$\mathbf{q}_{p+1}(\tau, \xi) = \sum_{i=1}^d \hat{\mathbf{r}}_i(\tau, \xi_i), \quad \boldsymbol{\xi} \in \Delta. \tag{4.2.37}$$

Substituting $\mathbf{q}_k = \mathbf{0},\ 0 \leq k \leq p$, and (4.2.37) into (4.3.80) yields (4.2.1a). This completes the proof. □

Corollary 4.2.2. *Under the conditions of Theorem 4.2.1, if all matrices* $\mathbf{A}_i$, $1 \leq i \leq d$ *are invertible, then the local DG error can be written as*

$$\mathbf{e}(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{j=1}^{d} \left(\mathbf{M}_j R^+_{p+1}(\xi_j) + (\mathbf{I} - \mathbf{M}_j) R^-_{p+1}(\xi_j) \right) \mathbf{c}_j + \mathcal{O}(h^{p+2}), \tag{4.2.38}$$

where

$$\mathbf{M}_j = \frac{1}{2}(\mathbf{I} + sgn(\mathbf{A}_j)),\ j = 1, 2. \tag{4.2.39}$$

Moreover, if, for instance, only $\mathbf{A}_1$ *is invertible, then the error can be written as*

$$\begin{aligned} \mathbf{e}(t, h\boldsymbol{\xi}) = {} & h^{p+1} \left(\mathbf{M}_1 R^+_{p+1}(\xi_1) + (\mathbf{I} - \mathbf{M}_1) R^-_{p+1}(\xi_1) \right) \mathbf{c}_1 \\ & + h^{p+1} \sum_{j=2}^{d} L_{p+1}(\xi_j)\mathbf{c}_j + L_p(\xi_j)(sgn(\mathbf{A}_j)\mathbf{c}_j + \mathbf{d}_j) + \mathcal{O}(h^{p+2}), \end{aligned} \tag{4.2.40}$$

where $\mathbf{c}_j$ *and* $\mathbf{d}_j$ *satisfy (4.2.1c).*

Proof. We prove the theorem when $\mathbf{A}_i$, $1 \leq i \leq d$ are invertible. Then $\mathbf{d}_i = \mathbf{0}$, $1 \leq i \leq d$ and we obtain

$$\begin{aligned} \mathbf{e}(t, h\boldsymbol{\xi}) &= h^{p+1} \sum_{i=1}^{d} L_{p+1}(\xi_i)\mathbf{c}_i + L_p(\xi_i)\mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i + \mathcal{O}(h^{p+2}) \\ &= h^{p+1} \sum_{i=1}^{d} (L_{p+1}(\xi_i) + L_p(\xi_i)) \frac{\mathbf{I} + \mathrm{sgn}(\mathbf{A}_i)}{2} \mathbf{c}_i \\ &\qquad + (L_{p+1}(\xi_i) - L_p(\xi_i)) \frac{\mathbf{I} - \mathrm{sgn}(\mathbf{A}_i)}{2} \mathbf{c}_i + \mathcal{O}(h^{p+2}). \end{aligned} \tag{4.2.41}$$

The proof for the other cases is similar and we will be omitted. □

4.3 Superconvergence and *A Posteriori* Error Estimation

In this section we investigate pointwise superconvergence for DG solutions and describe procedures to compute *a posteriori* DG error estimates that are asymptotically correct under mesh refinement.

4.3.1 Superconvergence

In order for the DG solution $\mathbf{u}_h$ to be $\mathcal{O}(h^{p+2})$-superconvergent at few points in element ω, the leading error term shown in Theorem 4.2.1 has to be zero at these points. This

pointwise superconvergence happens only for special hyperbolic problems as shown in the following theorem.

Theorem 4.3.1. *We let* $\bar{\xi}^s_k,\ 1 \le k \le p+1$, *denote the roots of* $R^s_{p+1}(\xi),\ s = +,-$, *shifted to* $[0,1]$. *Thus, under the conditions of Theorem 4.2.1 with* $p \ge 1$ *and* $t = \mathcal{O}(1)$,

i) If $\mathbf{z}$ *is a unit vector in the union of the spaces* $\bigcap_{i=1}^d \mathcal{R}(\mathbf{A}_i^{s_i}),\ s_i = +,-$, *then the projection* $\mathbf{z}^t\mathbf{e}(t,\mathbf{x})$ *of the local error onto span*$\{\mathbf{z}\}$ *is* $\mathcal{O}(h^{p+2})$ *superconvergent at the points* $(t, h\bar{\boldsymbol{\xi}}),\ \bar{\boldsymbol{\xi}} = (\bar{\xi}^{s_1}_{k_1}, \ldots, \bar{\xi}^{s_d}_{k_d}),\ 1 \le k_i \le p+1,\ s_i = +,-,\ 1 \le i \le d$, *i.e.,*

$$\mathbf{z}^t\mathbf{e}(t, h\bar{\boldsymbol{\xi}}) = \mathcal{O}(h^{p+2}). \tag{4.3.1}$$

ii) Moreover, if $\gamma_i(a) = \{\mathbf{x} \in (0,h)^d : x_i = a\}$, $0 \le a \le h$, *and if* $\mathbf{v} \in \mathcal{P}_{p-1}$ *is a unit vector with respect to the* C^∞ *norm, then, at* $a = h\bar{\xi}^s_k$, *we have the superconvergence of the following error averages*

$$\frac{1}{h^{d-1}} \int_{\gamma_i(h\bar{\xi}^s_k)} \mathbf{v}^t \mathbf{A}_i^s \mathbf{e}\, d\mathbf{s} = \mathcal{O}(h^{p+2}),\ 1 \le k \le p+1,\ s = +,-,\ 1 \le i \le d, \tag{4.3.2}$$

and

$$\frac{1}{h^{d-1}} \int_{\gamma_i^s} \mathbf{v}^t (\mathbf{A}_i^{\mu_i}\mathbf{e} + \mathbf{A}_i^{\bar{\mu}_i}\mathbf{e}^-)\, d\mathbf{s} = \mathcal{O}(h^{p+2}),\ s = +,-,\quad 1 \le i \le d. \tag{4.3.3}$$

Proof. We will prove (4.3.1) for the case $s_i = +$, $1 \le i \le d$.
Thus, assume that there exists a unit vector $\mathbf{z} \in \bigcap_{i=1}^d \mathcal{R}(\mathbf{A}_i^+)$, *i.e.*, there exists $\mathbf{v}_i$ such that

$$\mathbf{A}_i^+ \mathbf{v}_i = \mathbf{z},\ 1 \le i \le d. \tag{4.3.4}$$

Left pre-multiplying $\mathbf{e}$ in (4.2.1a) by $\mathbf{z}^t$ and evaluating the resulting function at the points $(t, h\bar{\boldsymbol{\xi}}),\ \bar{\boldsymbol{\xi}} = (\bar{\xi}^+_{k_1}, \ldots, \bar{\xi}^+_{k_d}),\ 1 \le k_i \le p+1,\ 1 \le i \le d$, we obtain

$$\mathbf{z}^t\mathbf{e}(t, h\bar{\boldsymbol{\xi}}) = h^{p+1} \sum_{i=1}^d \left(L_{p+1}(\bar{\xi}^+_{k_i}) \mathbf{z}^t\mathbf{c}_i - L_p(\bar{\xi}^+_{k_i})(\mathbf{z}^t \mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i + \mathbf{z}^t\mathbf{d}_i) \right) + \mathcal{O}(h^{p+2}). \tag{4.3.5}$$

By the property (1.2.25b) and (4.2.1c) we have $\mathbf{d}_i \in \mathcal{N}(\mathbf{A}_i) \subseteq \mathcal{N}(\mathbf{A}_i^+)$, which yields by (4.3.4)

$$\mathbf{z}^t\mathbf{d}_i = \mathbf{v}_i^t \mathbf{A}_i^+ \mathbf{d}_i = \mathbf{0}. \tag{4.3.6}$$

Applying (4.3.4) and the property (1.2.25d) yields

$$\mathbf{z}^t \mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i = \mathbf{v}_i^t \mathbf{A}_i^+ \mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i = \mathbf{v}_i^t \mathbf{A}_i^+ \mathbf{c}_i = \mathbf{z}^t\mathbf{c}_i. \tag{4.3.7}$$

Substituting (4.3.6) and (4.3.7) into (4.3.5), we prove that

$$\mathbf{z}^t\mathbf{e}(t, h\bar{\boldsymbol{\xi}}) = h^{p+1} \sum_{i=1}^d R^+_{p+1}(\bar{\xi}^+_{k_i}) \mathbf{z}^t\mathbf{c}_i + \mathcal{O}(h^{p+2}) = \mathcal{O}(h^{p+2}). \tag{4.3.8}$$

Following the same line of reasoning we establish (4.3.1) for all other cases.

We will establish (4.3.2) for the case $i = 1$ and $s = +$.
Let $\mathbf{v} \in \mathcal{P}_{p-1}$ be a unit vector in $[C^\infty]^m$ norm. We apply the scaling $\boldsymbol{\xi} = h^{-1}\mathbf{x}$, write $\hat{\mathbf{e}}(t, \boldsymbol{\xi}) = \mathbf{e}(t, h\boldsymbol{\xi})$ and use the definition of $\mathbf{e}$ in (4.2.1a) to obtain

$$\frac{1}{h^{d-1}} \int_{\gamma_1(h\bar{\xi}_k^+)} \mathbf{v}^t \mathbf{A}_1^+ \mathbf{e} \, d\mathbf{s} = \int_{\Gamma_1(\bar{\xi}_k^+)} \mathbf{v}^t \mathbf{A}_1^+ \hat{\mathbf{e}} \, d\boldsymbol{\sigma} \tag{4.3.9a}$$

$$= h^{p+1} \int_{\Gamma_1(\bar{\xi}_k^+)} \mathbf{v}^t \mathbf{A}_1^+ \Big(L_{p+1}(\bar{\xi}_k^+)\mathbf{c}_1 - L_p(\bar{\xi}_k^+)(\operatorname{sgn}(\mathbf{A}_1)\mathbf{c}_1 + \mathbf{d}_1) \tag{4.3.9b}$$

$$+ \sum_{j=2}^{d} L_{p+1}(\xi_j)\mathbf{c}_j - L_p(\xi_j)(\operatorname{sgn}(\mathbf{A}_j)\mathbf{c}_j + \mathbf{d}_j) \Big) \, d\boldsymbol{\sigma} + \mathcal{O}(h^{p+2}). \tag{4.3.9c}$$

By the flux properties (1.2.25b) and (1.2.25d), we have

$$\mathbf{A}_1^+(\operatorname{sgn}(\mathbf{A}_1)\mathbf{c}_1 + \mathbf{d}_1) = \mathbf{A}_1^+ \mathbf{c}_1. \tag{4.3.10}$$

Substituting (4.3.10) into (4.3.9b) and applying the orthogonality of Legendre polynomials together with $\mathbf{v} \in \mathcal{P}_{p-1}$ in (4.3.9c), we obtain

$$\begin{aligned} \frac{1}{h^{d-1}} \int_{\gamma_1(h\bar{\xi}_k^+)} \mathbf{v}^t \mathbf{A}_1^+ \mathbf{e} \, d\mathbf{s} &= h^{p+1} \int_{\Gamma_1(\bar{\xi}_k^+)} \mathbf{v}^t \mathbf{A}_1^+ R_{p+1}^+(\bar{\xi}_k^+)\mathbf{c}_1 \, d\boldsymbol{\sigma} + \mathcal{O}(h^{p+2}) \\ &= \mathcal{O}(h^{p+2}). \end{aligned} \tag{4.3.11}$$

Following the same line of reasoning, we establish (4.3.2) for all other cases.

To establish (4.3.3), assume $s = +$. Then (4.3.2) infers for $\gamma_i^+ = \gamma_i(h)$

$$\frac{1}{h^{d-1}} \int_{\gamma_i^+} \mathbf{v}^t \mathbf{A}_i^+ \mathbf{e} \, d\mathbf{s} = \mathcal{O}(h^{p+2}). \tag{4.3.12}$$

Further, (2.2.11c) yields

$$\mathbf{e}^-(t, \mathbf{x}) = h^{p+1} \sum_{j \in D(i)} L_{p+1}\left(\frac{x_j}{h}\right) \mathbf{c}_j + L_p\left(\frac{x_j}{h}\right) \operatorname{sgn}(\mathbf{A}_j)\mathbf{c}_j + \mathcal{O}(h^{p+2}), \tag{4.3.13}$$

which, by the orthogonality property (2.2.6) of Legendre polynomials, yields

$$\frac{1}{h^{d-1}} \int_{\gamma_i^+} \mathbf{v}^t \mathbf{A}_i^- \mathbf{e}^- \, d\mathbf{s} = \mathcal{O}(h^{p+2}), \ \forall \ \mathbf{v} \in \mathcal{P}_{p-1}. \tag{4.3.14}$$

Adding (4.3.12) and (4.3.14) yields (4.3.3). The case $s = -$ can be treated using the same line of reasoning and is omitted. □

4.3.2 *A Posteriori* Error Estimation

In this section we present an *a posteriori* error estimation procedure which consists of computing asymptotically exact local and global error estimates of the DG error. In Theorem 4.2.1 we showed that the local discretization error for the DG method on a physical element $\omega = (0,h)^d$ can be written as

$$\mathbf{e}(t,h\boldsymbol{\xi}) = h^{p+1}\sum_{i=1}^{d} L_{p+1}(\xi_i)\mathbf{c}_i(t) - L_p(\xi_i)(\mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i(t) + \mathbf{d}_i(t)) + \mathcal{O}(h^{p+2}), \tag{4.3.15}$$

where

$$\mathbf{d}_i(t) \in \mathcal{N}(\mathbf{A}_i) \cap \bigoplus_{k=1}^{d} \mathcal{R}(\mathbf{A}_k), \quad 1 \le i \le d. \tag{4.3.16}$$

We apply the pseudoinverse $\mathbf{A}_i^\dagger$ of $\mathbf{A}_i$ to split $\mathbf{c}_i$ into

$$\mathbf{c}_i = \mathbf{c}_i^\perp + \mathbf{c}_i^\Phi, \text{ where } \mathbf{c}_i^\perp = \mathbf{A}_i^\dagger\mathbf{A}_i\mathbf{c}_i, \quad 1 \le i \le d. \tag{4.3.17}$$

We note that by Lemma 1.2.13, $\mathbf{A}_i^\dagger\mathbf{A}_i$ is the projection onto $\mathcal{N}(\mathbf{A}_i)^\perp$, thus

$$\mathbf{c}_i^\perp = \mathbf{A}_i^\dagger\mathbf{A}_i\mathbf{c}_i \in \mathcal{N}(\mathbf{A}_i)^\perp, \ \mathbf{c}_i^\Phi \in \mathcal{N}(\mathbf{A}_i), \ 1 \le i \le d. \tag{4.3.18}$$

By (1.2.25b), $\mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i^\Phi = \mathbf{0}$, which yields $\mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i = \mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i^\perp \in \mathcal{R}(\mathbf{A}_i) = \mathcal{N}(\mathbf{A}_i)^\perp$.

Hence, the leading term of the spatial discretization error can be split into two parts as

$$\mathbf{e} = \mathbf{e}^\perp + \mathbf{e}^\Phi + \mathcal{O}(h^{p+2}), \tag{4.3.19}$$

where

$$\mathbf{e}^\perp(t,h\boldsymbol{\xi}) = h^{p+1}\sum_{i=1}^{d} L_{p+1}(\xi_i)\mathbf{c}_i^\perp(t) - L_p(\xi_i)\mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i^\perp(t), \tag{4.3.20}$$

and

$$\mathbf{e}^\Phi(t,h\boldsymbol{\xi}) = h^{p+1}\sum_{i=1}^{d} L_{p+1}(\xi_i)\mathbf{c}_i^\Phi(t) - L_p(\xi_i)\mathbf{d}_i(t). \tag{4.3.21}$$

We note that for invertible matrices $\mathbf{A}_i$, $1 \le i \le d$, the error component $\mathbf{e}^\Phi(t,\mathbf{x})$ is zero.

Next, we develop an *a posteriori* error estimation procedure for estimating both $\mathbf{e}^\perp$ and $\mathbf{e}^\Phi$ (if needed). We prove that, for smooth solutions, our local error estimates converge to the true error under mesh refinement. Up to this point we are not able to prove the asymptotic exactness of our global *a posteriori* error estimates. However, computational results for several hyperbolic systems shown in § 4.4 suggest that our global *a posteriori* error estimates are asymptotically exact under mesh refinement for smooth solutions.

4.3.3 The Stationary Component of the Error Estimate

The *a posteriori* error estimation procedure to compute estimates for $\mathbf{e}^\perp$ consists of determining

$$\mathbf{E}^\perp(t, h\boldsymbol{\xi}) = \sum_{j=1}^{d} \left(L_{p+1}(\xi_j)\boldsymbol{\gamma}_j^\perp(t) - L_p(\xi_j)\mathrm{sgn}(\mathbf{A}_j)\boldsymbol{\gamma}_j^\perp(t) \right),\ \boldsymbol{\gamma}_j^\perp \in \mathcal{N}(\mathbf{A}_j)^\perp, \tag{4.3.22a}$$

such that

$$\int_\omega L_p\left(\frac{x_i}{h}\right) \mathbf{v}^t \left(\frac{\partial \mathbf{u}_h}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial(\mathbf{u}_h + \mathbf{E}^\perp)}{\partial x_j} - \mathbf{g} \right) d\mathbf{x} = 0,\ \forall\ \mathbf{v} \in \mathcal{N}(\mathbf{A}_i)^\perp,\ 1 \le i \le d. \tag{4.3.22b}$$

We note that, by Lemma 1.2.13, $\mathsf{P}_i = \mathbf{A}_i\mathbf{A}_i^\dagger$ is symmetric and projects any vector in $\mathbb{R}^m$ into $\mathcal{R}(\mathbf{A}_i) = \mathcal{N}(\mathbf{A}_i)^\perp$, thus the columns of P_i span $\mathcal{N}(\mathbf{A}_i)^\perp$.

Substituting $\mathbf{v}$ by $\mathsf{P}_i = \mathsf{P}_i^t$ in (4.3.22b) yields

$$\int_\omega L_p\left(\frac{x_i}{h}\right) \mathsf{P}_i \left(\frac{\partial \mathbf{u}_h}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial(\mathbf{u}_h + \mathbf{E}^\perp)}{\partial x_j} - \mathbf{g} \right) d\mathbf{x} = 0,\ 1 \le i \le d. \tag{4.3.23}$$

Substituting (4.3.22a) into (4.3.23) and applying the orthogonality properties (2.2.6), we obtain

$$\mathbf{A}_i\boldsymbol{\gamma}_i^\perp \int_\omega L_p\left(\frac{x_i}{h}\right) L'_{p+1}\left(\frac{x_i}{h}\right) d\mathbf{x} = \mathfrak{r}_{p,i}^\perp, \tag{4.3.24a}$$

where $\mathfrak{r}_{p,i}^\perp$ is the projection of the residual defined as

$$\mathfrak{r}_{p,i}^\perp = \mathsf{P}_i \int_\omega L_p\left(\frac{x_i}{h}\right) \left(\mathbf{g} - \frac{\partial \mathbf{u}_h}{\partial t} - \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{u}_h}{\partial x_j} \right) d\mathbf{x}, \quad 1 \le i \le d. \tag{4.3.24b}$$

Using (2.2.6) we further reduce (4.3.24a), obtaining

$$2h^{d-1}\mathbf{A}_i\boldsymbol{\gamma}_i^\perp = \mathfrak{r}_{p,i}^\perp, \quad 1 \le i \le d. \tag{4.3.25}$$

Since $\mathfrak{r}_{p,i}^\perp \in \mathcal{N}(\mathbf{A}_i)^\perp = \mathcal{R}(\mathbf{A}_i)$, we can solve (4.3.25) to find the unique solution $\boldsymbol{\gamma}_i^\perp \in \mathcal{N}(\mathbf{A}_i)^\perp$,

$$\boldsymbol{\gamma}_i^\perp = \frac{h^{1-d}}{2}\mathbf{A}_i^\dagger \mathfrak{r}_{p,i}^\perp, \quad 1 \le i \le d. \tag{4.3.26}$$

We will now show that this static error estimate is asymptotically exact.

Theorem 4.3.2. *Under the assumptions of Theorem 4.2.1, let us consider the error estimate*

$$\mathbf{E}^{\perp}(t,h\boldsymbol{\xi}) = \sum_{i=1}^{d} \left(L_{p+1}(\xi_i) - L_p(\xi_i) sgn(\mathbf{A}_i)\right) \frac{h^{1-d}}{2} \mathbf{A}_i^{\dagger} \mathfrak{r}_{p,i}^{\perp}, \tag{4.3.27}$$

where $\mathfrak{r}_{p,i}^{\perp},\ 1 \le i \le d$, *are defined in (4.3.24b).*
Then, for $p \ge 1$ *and* $t = \mathcal{O}(1)$,

$$\mathbf{e}^{\perp}(t,\mathbf{x}) = \mathbf{E}^{\perp}(t,\mathbf{x}) + \mathcal{O}(h^{p+2}), \quad \mathbf{x} \in \omega. \tag{4.3.28}$$

Proof. Since the true solution $\mathbf{u}$ satisfies equation (4.0.1), we have

$$\int_{\omega} L_p\left(\frac{x_i}{h}\right) \mathsf{P}_i \left(\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{u}}{\partial x_j} - \mathbf{g}\right) d\mathbf{x} = 0, \quad 1 \le i \le d. \tag{4.3.29}$$

Subtracting (4.3.23) from (4.3.29) yields

$$\int_{\omega} L_p\left(\frac{x_i}{h}\right) \mathsf{P}_i \left(\frac{\partial \mathbf{e}}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial (\mathbf{e} - \mathbf{E}^{\perp})}{\partial x_j}\right) d\mathbf{x} = 0. \tag{4.3.30}$$

Applying $\mathbf{A}_i \mathbf{e}_{,x_i}^{\Phi} = \mathbf{0}$ and the orthogonality of Legendre polynomials (2.2.6), (4.3.30) infers that

$$\int_{\omega} L_p\left(\frac{x_i}{h}\right) \mathsf{P}_i \left(\frac{\partial \mathbf{e}}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial (\mathbf{e}^{\perp} - \mathbf{E}^{\perp})}{\partial x_j}\right) d\mathbf{x} = 0. \tag{4.3.31}$$

Applying the linear transformations $t = T\tau$ and $\mathbf{x} = h\boldsymbol{\xi}$, (4.3.31) becomes

$$\int_{\Delta} L_p(\xi_i) \mathsf{P}_i \left(\frac{h}{T}\frac{\partial \hat{\mathbf{e}}}{\partial \tau} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial (\mathbf{e}^{\perp} - \mathbf{E}^{\perp})}{\partial \xi_j}\right) d\boldsymbol{\xi} = 0. \tag{4.3.32}$$

Substituting the definitions of $\mathbf{e}^{\perp}$ (4.3.20) and $\mathbf{E}^{\perp}$ (4.3.22a) into (4.3.32), noting that used that $\operatorname{sgn}(\mathbf{A}_i)\frac{\partial}{\partial \tau}\hat{\mathbf{d}}_i = \mathbf{0}$ by the property (1.2.25b) and applying the orthogonality properties (2.2.6), we obtain

$$\mathsf{P}_i \int_{\Delta} \frac{h^{p+2}}{T} L_p^2(\xi_i) \operatorname{sgn}(\mathbf{A}_i) \frac{\partial \hat{\mathbf{c}}_i^{\perp}}{\partial \tau} + L_p(\xi_i) L_{p+1}'(\xi_i) \mathbf{A}_i (h^{p+1} \hat{\mathbf{c}}_i^{\perp} - \hat{\boldsymbol{\gamma}}_i^{\perp})\, d\boldsymbol{\xi} = \mathcal{O}(h^{p+2}). \tag{4.3.33}$$

Using (2.2.6), and the fact that $\mathsf{P}_i \operatorname{sgn}(\mathbf{A}_i) = \operatorname{sgn}(\mathbf{A}_i)$ by the property (1.2.25c), (4.3.33) can be further simplified to

$$\frac{h^{p+2}}{T(2p+1)} \operatorname{sgn}(\mathbf{A}_i) \frac{\partial \hat{\mathbf{c}}^{\perp}}{\partial \tau} + 2\mathbf{A}_i (h^{p+1} \hat{\mathbf{c}}_i^{\perp} - \hat{\boldsymbol{\gamma}}_i^{\perp}) = \mathcal{O}(h^{p+2}). \tag{4.3.34}$$

Thus, at $T = \mathcal{O}(1)$, we have

$$2\mathbf{A}_i(h^{p+1}\mathbf{c}_i^\perp - \boldsymbol{\gamma}_i^\perp) = \mathcal{O}(h^{p+2}). \tag{4.3.35}$$

Since $\mathbf{c}_i^\perp, \boldsymbol{\gamma}_i^\perp \in \mathcal{N}(\mathbf{A}_i)^\perp$, equation (4.3.35) has the unique solution

$$\boldsymbol{\gamma}_i^\perp = h^{p+1}\mathbf{c}_i^\perp + \mathcal{O}(h^{p+2}), \quad 1 \le i \le d. \tag{4.3.36}$$

This establishes (4.3.28). □

4.3.4 The Transient Component of the Error Estimate

We will first present an *a posteriori* error estimation procedure to compute estimates for $\mathbf{e}^{\Psi}$. Then we will show the asymptotic exactness of this error estimate.

Note that $\mathbf{e}^{\Psi} = \mathbf{0}$ by definition in (4.3.21), if all $\mathbf{A}_i$, $1 \le i \le d$, are invertible.

By Lemma 2.2.4, the approximations $\pi\mathbf{u}_0$ on ω and $\pi_i\mathbf{u}$ on the boundary $\partial\omega$ satisfy

$$\begin{aligned}\mathbf{e}(0,\mathbf{x}) &= \mathbf{u}_0(\mathbf{x}) - \pi\mathbf{u}_0(\mathbf{x}) \qquad (4.3.37)\\ &= h^{p+1}\sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j(0) - L_p\left(\frac{x_j}{h}\right)\operatorname{sgn}(\mathbf{A}_j)\mathbf{c}_j(0) + \mathcal{O}(h^{p+2}),\ \mathbf{x}\in\omega,\\ \mathbf{e}^-(t,\mathbf{x}) &= \mathbf{u}(t,\mathbf{x}) - \pi_i\mathbf{u}(t,\mathbf{x}) \qquad (4.3.38)\\ &= h^{p+1}\sum_{j\in D(i)} L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j(t) - L_p\left(\frac{x_j}{h}\right)\operatorname{sgn}(\mathbf{A}_j)\mathbf{c}_j(t) + \mathcal{O}(h^{p+2}),\ \mathbf{x}\in\gamma_i,\ 1\le i\le d.\end{aligned}$$

We split the error at $t = 0$ into $\mathbf{e} = \mathbf{e}^\perp + \mathbf{e}^{\Psi} + \mathcal{O}(h^{p+2})$ as in (4.3.19) and define $\mathbf{E}^{\Psi}(0,\mathbf{x})$ by

$$\mathbf{E}^{\Psi}(0,\mathbf{x}) = \mathbf{e}^{\Psi}(0,\mathbf{x}) = h^{p+1}\sum_{i=1}^{d} L_{p+1}\left(\frac{x_i}{h}\right)(\mathbf{I}-\mathsf{P}_i)\mathbf{c}_i(0), \tag{4.3.39}$$

where $(\mathbf{I}-\mathsf{P}_i)\mathbf{c}_i(0)$ is the projection of $\mathbf{c}_i(0)$ into $\mathcal{N}(\mathbf{A}_i)$.

On the boundary, we define $\mathbf{E}^-$ by the leading term of (4.3.38),

$$\mathbf{E}^-(t,\mathbf{x}) = h^{p+1}\sum_{j\in D(i)} L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j(t) - L_p\left(\frac{x_j}{h}\right)\operatorname{sgn}(\mathbf{A}_j)\mathbf{c}_j(t),\ \mathbf{x}\in\gamma_i,\ 1\le i\le d. \tag{4.3.40}$$

Then we define the error estimate for $\mathbf{e}^{\Psi}$ by determining the coefficients of

$$\mathbf{E}^{\Psi}(t,\mathbf{x}) = \sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right)\boldsymbol{\gamma}_j^{\Psi}(t) - L_p\left(\frac{x_j}{h}\right)\boldsymbol{\delta}_j^{\Psi}(t), \quad \boldsymbol{\gamma}_j^{\Psi}, \boldsymbol{\delta}_j^{\Psi} \in \mathcal{N}(\mathbf{A}_j),\ 1\le j\le d, \tag{4.3.41a}$$

such that

$$\int_\omega \mathbf{v}^t \left(\frac{\partial(\mathbf{u}_h + \mathbf{E}^{\mathbf{\Phi}})}{\partial t} + \sum_{j=1}^d \mathbf{A}_j \frac{\partial \mathbf{u}_h}{\partial x_j} - \mathbf{g} \right) d\mathbf{x}$$
$$= \sum_{j=1}^d \int_{\gamma_j} \mathbf{v}^t \nu_j \mathbf{A}_j^{\bar{\mu}_j} \left(\mathbf{u}_h + \mathbf{E}^\perp + \mathbf{E}^{\mathbf{\Phi}} - \mathbf{u}_h^- - \mathbf{E}^- \right) d\mathbf{s}, \ \forall \ \mathbf{v} \in \mathcal{E}_p, \tag{4.3.41b}$$

where $\mathbf{E}^\perp$ equals the static component defined by (4.3.22) and

$$\mathcal{E}_p = \left\{ \mathbf{v}(\mathbf{x}) = \sum_{i=1}^d \left(L_{p+1}\left(\frac{x_i}{h}\right) \mathbf{a}_i - L_p\left(\frac{x_i}{h}\right) \mathbf{b}_i \right) : \ \mathbf{a}_i, \mathbf{b}_i \in \mathcal{N}(\mathbf{A}_i) \right\}. \tag{4.3.41c}$$

The reason for choosing equation (4.3.41b) to estimate $\mathbf{E}^{\mathbf{\Phi}}$ will become clear when we prove the asymptotic exactness of the $\mathbf{E}^{\mathbf{\Phi}}$ in Theorem 4.3.4.

By Lemma 1.2.14, $(\mathbf{I} - \mathsf{P}_i)$ projects any vector in $\mathbb{R}^m$ into $\mathcal{N}(\mathbf{A}_i)$ and the columns of $(\mathbf{I} - \mathsf{P}_i)$ span $\mathcal{N}(\mathbf{A}_i)$. Hence the columns of $L_{p+1}(\xi_i)(\mathbf{I} - \mathsf{P}_i)$ and $L_p(\xi_i)(\mathbf{I} - \mathsf{P}_i),\ 1 \le i \le d$, span $\mathcal{E}_p$.

Replacing $\mathbf{v}$ in (4.3.41b) by $L_m(\xi_i)(\mathbf{I} - \mathsf{P}_i),\ m = p, p+1,\ 1 \le i \le d$, yields

$$\int_\omega L_m\left(\frac{x_i}{h}\right)(\mathbf{I} - \mathsf{P}_i) \left(\frac{\partial(\mathbf{u}_h + \mathbf{E}^{\mathbf{\Phi}})}{\partial t} + \sum_{j=1}^d \mathbf{A}_j \frac{\partial \mathbf{u}_h}{\partial x_j} - \mathbf{g} \right) d\mathbf{x}$$
$$= \sum_{j=1}^d \int_{\gamma_j} L_m\left(\frac{x_i}{h}\right)(\mathbf{I} - \mathsf{P}_i) \nu_j \mathbf{A}_j^{\bar{\mu}_j} \left(\mathbf{u}_h + \mathbf{E}^\perp + \mathbf{E}^{\mathbf{\Phi}} - \mathbf{u}_h^- - \mathbf{E}^- \right) d\mathbf{s},$$
$$\forall\ m = p, p+1,\ 1 \le i \le d. \tag{4.3.42}$$

By Lemma 1.2.14, $(\mathbf{I} - \mathsf{P}_i)\mathbf{A}_i^{\bar{\mu}_i} = \mathbf{0}$, thus (4.3.42) can be written as

$$\int_\omega L_m\left(\frac{x_i}{h}\right)(\mathbf{I} - \mathsf{P}_i) \frac{\partial \mathbf{E}^{\mathbf{\Phi}}}{\partial t} d\mathbf{x} - \sum_{j \in D(i)} \int_{\gamma_j} L_m\left(\frac{x_i}{h}\right)(\mathbf{I} - \mathsf{P}_i) \nu_j \mathbf{A}_j^{\bar{\mu}_j} \mathbf{E}^{\mathbf{\Phi}} d\mathbf{s} = \mathfrak{r}_{m,i}^{\mathbf{\Phi}}, \tag{4.3.43a}$$

where $\mathfrak{r}_{m,i}^{\mathbf{\Phi}}$ is the projection of the residual given by

$$\mathfrak{r}_{m,i}^{\mathbf{\Phi}} = (\mathbf{I} - \mathsf{P}_i) \int_\omega L_m\left(\frac{x_i}{h}\right) \left(\mathbf{g} - \frac{\partial \mathbf{u}_h}{\partial t} - \sum_{j=1}^d \mathbf{A}_j \frac{\partial \mathbf{u}_h}{\partial x_j} \right) d\mathbf{x}$$
$$+ (\mathbf{I} - \mathsf{P}_i) \sum_{j=1}^d \int_{\gamma_j} L_m\left(\frac{x_i}{h}\right) \nu_j \mathbf{A}_j^{\bar{\mu}_j} \left(\mathbf{u}_h + \mathbf{E}^\perp - \mathbf{u}_h^- - \mathbf{E}^- \right) d\mathbf{s},$$
$$m = p, p+1,\ 1 \le i \le d. \tag{4.3.43b}$$

For $m = p+1$, we use the orthogonality properties (2.2.6) to reduce (4.3.43a) to

$$\int_\omega L_{p+1}^2\left(\frac{x_i}{h}\right)\dot{\boldsymbol{\gamma}}_i^{\boldsymbol{\Phi}}\,d\mathbf{x} - \sum_{j\in D(i)}\int_{\gamma_j} L_{p+1}^2\left(\frac{x_i}{h}\right)\nu_j(\mathbf{I}-\mathsf{P}_i)\mathbf{A}_j^{\bar{\mu}_j}\boldsymbol{\gamma}_i^{\boldsymbol{\Phi}}\,d\mathbf{s} = \mathfrak{r}_{p+1,i}^{\boldsymbol{\Phi}}, \tag{4.3.44}$$

which by (2.2.6) is equal to

$$\dot{\boldsymbol{\gamma}}_i^{\boldsymbol{\Phi}} = \frac{1}{h}(\mathbf{I}-\mathsf{P}_i)\sum_{j\in D(i)}(\mathbf{A}_j^- - \mathbf{A}_j^+)\boldsymbol{\gamma}_i^{\boldsymbol{\Phi}} + \frac{2p+3}{h^d}\mathfrak{r}_{p+1,i}^{\boldsymbol{\Phi}}. \tag{4.3.45a}$$

For $m = p$, we get similarly

$$\dot{\boldsymbol{\delta}}_i^{\boldsymbol{\Phi}} = \frac{1}{h}(\mathbf{I}-\mathsf{P}_i)\sum_{j\in D(i)}(\mathbf{A}_j^- - \mathbf{A}_j^+)\boldsymbol{\delta}_i^{\boldsymbol{\Phi}} + \frac{2p+1}{h^d}\mathfrak{r}_{p,i}^{\boldsymbol{\Phi}}, \tag{4.3.45b}$$

subject to the initial conditions

$$\boldsymbol{\gamma}_i^{\boldsymbol{\Phi}}(0) = h^{p+1}(\mathbf{I}-\mathsf{P}_i)\mathbf{c}_i(0),\ \boldsymbol{\delta}_i^{\boldsymbol{\Phi}}(0) = \mathbf{0}. \tag{4.3.45c}$$

Note that (4.3.45) and (4.3.43b) ensures that $\boldsymbol{\gamma}_i^{\boldsymbol{\Phi}}, \boldsymbol{\delta}_i^{\boldsymbol{\Phi}} \in \mathcal{N}(\mathbf{A}_i),\ 1 \le i \le d$.

Then (4.3.45) and (4.3.43b) together describe the procedure to obtain the coefficients of $\mathbf{E}^{\boldsymbol{\Phi}}$.

4.3.5 Asymptotic Exactness of the Transient Component of the Error Estimate

In this subsection we will show the asymptotic exactness of the error estimate.

We define

$$\bar{\mathcal{E}}_p = \left\{\mathbf{v}(\mathbf{x}) = \sum_{i=1}^d \left(L_{p+1}\left(\frac{x_i}{h}\right)\mathbf{a}_i - L_p\left(\frac{x_i}{h}\right)\mathbf{b}_i\right) :\ \mathbf{a}_i,\mathbf{b}_i \in \mathcal{N}(\mathbf{A}_i)\cap\bigoplus_{k=1}^d \mathcal{R}(\mathbf{A}_k)\right\}. \tag{4.3.46a}$$

Then we can split $\mathcal{E}_p = \bar{\mathcal{E}}_p \oplus \bar{\mathcal{E}}_p{}^\perp$, where

$$\bar{\mathcal{E}}_p{}^\perp = \left\{\mathbf{v}(\mathbf{x}) \in \mathcal{E}_p :\ \int_\omega \mathbf{w}^t\mathbf{v}\,d\mathbf{x} = 0,\ \forall\ \mathbf{w} \in \bar{\mathcal{E}}_p\right\} \tag{4.3.46b}$$

$$= \left\{\mathbf{v}(\mathbf{x}) = \sum_{i=1}^d \left(L_{p+1}\left(\frac{x_i}{h}\right)\mathbf{a}_i - L_p\left(\frac{x_i}{h}\right)\mathbf{b}_i\right) :\ \mathbf{a}_i,\mathbf{b}_i \in \bigcap_{k=1}^d \mathcal{N}(\mathbf{A}_k)\right\}, \tag{4.3.46c}$$

where we used the fact that, by Lemma 1.2.3,

$$\bigcap_{k=1}^d \mathcal{N}(\mathbf{A}_k) = \left(\bigoplus_{k=1}^d \mathcal{R}(\mathbf{A}_k)\right)^\perp. \tag{4.3.47}$$

Lemma 4.3.3. *If $\mathbf{q} \in \bar{\mathcal{E}}_p$ satisfies the orthogonality condition*

$$\sum_{i=1}^{d} \left(\int_{\Delta} \mathbf{v}^t \mathbf{A}_i \frac{\partial \mathbf{q}}{\partial \xi_i} \, d\boldsymbol{\xi} - \int_{\Gamma_i} \mathbf{v}^t \nu_i \mathbf{A}_i^{\bar{\mu}_i} \mathbf{q} \, d\boldsymbol{\sigma} \right) = 0, \ \forall \ \mathbf{v} \in \mathcal{E}_p, \tag{4.3.48}$$

then $\mathbf{q} = \mathbf{0}$.

Proof. First we integrate equation (4.3.48) by parts to write

$$\sum_{i=1}^{d} \left(- \int_{\Delta} \frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \mathbf{q} \, d\boldsymbol{\xi} + \int_{\Gamma_i} \mathbf{v}^t \nu_i \mathbf{A}_i^{\mu_i} \mathbf{q} \, d\boldsymbol{\sigma} \right) = 0, \ \forall \ \mathbf{v} \in \mathcal{E}_p. \tag{4.3.49}$$

Adding (4.3.48) and (4.3.49) and setting $\mathbf{v} = \mathbf{q}$, the integral on Δ vanishes because of the symmetry of $\mathbf{A}_i$, $1 \leq i \leq d$, and we get

$$\sum_{i=1}^{d} \int_{\Gamma_i} \mathbf{q}^t (\mathbf{A}_i^+ - \mathbf{A}_i^-) \mathbf{q} \, d\boldsymbol{\sigma} = 0. \tag{4.3.50}$$

Since $(\mathbf{A}_i^+ - \mathbf{A}_i^-)$ is positive semi-definite by (1.2.25f), there exists a matrix $\mathbf{L}_i$ such that $\mathbf{L}_i^t \mathbf{L}_i = (\mathbf{A}_i^+ - \mathbf{A}_i^-)$, and (4.3.50) yields

$$\sum_{i=1}^{d} \int_{\Gamma_i} \|\mathbf{L}_i \mathbf{q}\|^2 \, d\boldsymbol{\sigma} = 0, \tag{4.3.51}$$

which yields

$$\mathbf{L}_i \mathbf{q}|_{\Gamma_i} = \mathbf{0}, \tag{4.3.52}$$

and therefore

$$(\mathbf{A}_i^+ - \mathbf{A}_i^-) \mathbf{q}|_{\Gamma_i} = \mathbf{L}_i^t \mathbf{L}_i \mathbf{q}|_{\Gamma_i} = \mathbf{0}, \ 1 \leq i \leq d. \tag{4.3.53}$$

By property (1.2.25b) and (4.3.53) we obtain

$$\mathbf{A}_i \mathbf{q}|_{\Gamma_i} = \mathbf{0}, \quad 1 \leq i \leq d. \tag{4.3.54}$$

Since $\mathbf{q} \in \bar{\mathcal{E}}_p$,

$$\mathbf{q}(\boldsymbol{\xi}) = \sum_{j=1}^{d} L_{p+1}(\xi_j) \mathbf{a}_j - L_p(\xi_j) \mathbf{b}_j, \ \mathbf{a}_j, \mathbf{b}_j \in \mathcal{N}(\mathbf{A}_j) \cap \bigoplus_{i=1}^{d} \mathcal{R}(\mathbf{A}_k). \tag{4.3.55}$$

Substituting (4.3.55) into (4.3.54) yields

$$\sum_{j=1}^{d} L_{p+1}(\xi_j)(\mathbf{A}_i \mathbf{a}_j) - L_p(\xi_j)(\mathbf{A}_i \mathbf{b}_j) = \mathbf{0}, \quad \boldsymbol{\xi} \in \Gamma_i, \ 1 \leq i \leq d. \tag{4.3.56}$$

Since $L_p(\xi_j),\ L_{p+1}(\xi_j),\ j \in D(i)$ are pairwise orthogonal functions on Γ_i and therefore linearly independent, (4.3.56) yields

$$\mathbf{A}_i\mathbf{a}_j = \mathbf{0},\ \mathbf{A}_i\mathbf{b}_j = \mathbf{0}, \quad 1 \le i,j \le d. \tag{4.3.57}$$

Thus, $\mathbf{a}_j, \mathbf{b}_j \in \bigcup_{i=1}^d \mathcal{N}(\mathbf{A}_i)$, which, when combined with $\mathbf{a}_j, \mathbf{b}_j \in \bigoplus_{i=1}^d \mathcal{R}(\mathbf{A}_k)$ and (4.3.47), yields

$$\mathbf{a}_j = \mathbf{b}_j = \mathbf{0},\ 1 \le j \le d, \tag{4.3.58}$$

or equivalently $\mathbf{q} = \mathbf{0}$. □

Theorem 4.3.4. *Under the assumptions of Theorem 4.2.1, assume further that $\mathbf{u}_h$ is computed by approximating the initial conditions by $\pi\mathbf{u}_0$ and let*

$$\mathbf{E}^{\mathbf{\Phi}}(t, h\boldsymbol{\xi}) = \sum_{j=1}^{d} (L_{p+1}(\xi_j)\boldsymbol{\gamma}_j^{\mathbf{\Phi}}(t) - L_p(\xi_j)\boldsymbol{\delta}_j^{\mathbf{\Phi}}(t)), \tag{4.3.59}$$

where $\boldsymbol{\gamma}_i^{\mathbf{\Phi}}, \boldsymbol{\delta}_i^{\mathbf{\Phi}},\ 1 \le i \le d$, are solutions of (4.3.45) and (4.3.43b).
Then, at $t = \mathcal{O}(1)$ and for $p \ge 1$,

$$\mathbf{e}^{\mathbf{\Phi}}(t, \mathbf{x}) = \mathbf{E}^{\mathbf{\Phi}}(t, \mathbf{x}) + \mathcal{O}(h^{p+2}), \quad \mathbf{x} \in \omega. \tag{4.3.60}$$

Proof. Since the true solution $\mathbf{u}$ is continuous and $\mathbf{u} = \mathbf{u}^-$ on $\partial\omega$, $\mathbf{u}$ satisfies

$$\int_\omega \mathbf{v}^t \left(\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{u}}{\partial x_j} - \mathbf{g} \right) d\mathbf{x} = \sum_{j=1}^{d} \int_{\gamma_j} \mathbf{v}^t \nu_j \mathbf{A}_j^{\bar{\mu}_j} \left(\mathbf{u} - \mathbf{u}^- \right) d\mathbf{s},\ \forall\ \mathbf{v} \in \mathcal{E}_p. \tag{4.3.61}$$

Subtracting (4.3.41b) from (4.3.61) gives

$$\begin{aligned} &\int_\omega \mathbf{v}^t \left(\frac{\partial (\mathbf{e} - \mathbf{E}^{\mathbf{\Phi}})}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{e}}{\partial x_j} \right) d\mathbf{x} \\ &= \sum_{j=1}^{d} \int_{\gamma_j} \mathbf{v}^t \nu_j \mathbf{A}_j^{\bar{\mu}_j} \left(\mathbf{e} - \mathbf{E}^\perp - \mathbf{E}^{\mathbf{\Phi}} - \mathbf{e}^- + \mathbf{E}^- \right) d\mathbf{s},\ \forall\ \mathbf{v} \in \mathcal{E}_p. \end{aligned} \tag{4.3.62}$$

Since $\mathbf{v} \in \mathcal{E}_p$, we can write

$$\mathbf{v}(\mathbf{x}) = \sum_{i=1}^{d} L_{p+1}\left(\frac{x_i}{h}\right) \mathbf{a}_i - L_p\left(\frac{x_i}{h}\right) \mathbf{b}_i, \quad \mathbf{a}_i, \mathbf{b}_i \in \mathcal{N}(\mathbf{A}_i), \tag{4.3.63}$$

while $\mathbf{E}^\perp$ is defined in (4.3.22a) as

$$\mathbf{E}^\perp(t, h\boldsymbol{\xi}) = \sum_{j=1}^{d} L_{p+1}(\xi_j)\boldsymbol{\gamma}_j^\perp(t) - L_p(\xi_j)\mathrm{sgn}(\mathbf{A}_j)\boldsymbol{\gamma}_j^\perp(t), \quad \boldsymbol{\gamma}_j^\perp \in \mathcal{N}(\mathbf{A}_j)^\perp. \tag{4.3.64}$$

By (1.2.25c) and (1.2.12), $\operatorname{sgn}(\mathbf{A}_j)\dot{\boldsymbol{\gamma}}_j^\perp \in \mathcal{R}(\mathbf{A}_j) = \mathcal{N}(\mathbf{A}_j)^\perp$, which, together with (4.3.64), yields

$$\langle \mathbf{a}_i, \dot{\boldsymbol{\gamma}}_i^\perp(t)\rangle = 0,\ \langle \mathbf{b}_i, \operatorname{sgn}(\mathbf{A}_j)\dot{\boldsymbol{\gamma}}_j^\perp(t)\rangle = 0, \quad 1 \le i \le d. \tag{4.3.65}$$

By substituting $\mathbf{v}$ and $\mathbf{E}^\perp$, as defined in (4.3.63) and (4.3.64), into $\int_\omega \mathbf{v}^t \frac{\partial \mathbf{E}^\perp}{\partial t}\, d\mathbf{x}$ and applying the orthogonality property (2.2.6), we obtain

$$\begin{aligned}
\int_\omega \mathbf{v}^t \frac{\partial \mathbf{E}^\perp}{\partial t}\, d\mathbf{x} &= \sum_{i=1}^d \sum_{j=1}^d \int_\omega \left(L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{a}_i - L_p\left(\frac{x_j}{h}\right)\mathbf{b}_i\right)^t \\
&\qquad \left(L_{p+1}\left(\frac{x_i}{h}\right)\dot{\boldsymbol{\gamma}}_i^\perp(t) - L_p\left(\frac{x_i}{h}\right)\operatorname{sgn}(\mathbf{A}_i)\dot{\boldsymbol{\gamma}}_i^\perp(t)\right) d\mathbf{x} \\
&= \sum_{i=1}^d \int_\omega L_{p+1}^2\left(\frac{x_i}{h}\right)\langle \mathbf{a}_i, \dot{\boldsymbol{\gamma}}_i^\perp(t)\rangle + L_p^2\left(\frac{x_i}{h}\right)\langle \mathbf{b}_i, \operatorname{sgn}(\mathbf{A}_j)\dot{\boldsymbol{\gamma}}_j^\perp(t)\rangle\, d\mathbf{x} \\
&= 0, \quad \forall\ \mathbf{v} \in \mathcal{E}_p.
\end{aligned} \tag{4.3.66}$$

Furthermore, by substituting $\mathbf{v}$, $\mathbf{E}^\perp$ and $\mathbf{E}^\Phi$, as defined in (4.3.63), (4.3.64) and (4.3.41a), into $\int_\omega \mathbf{v}^t \mathbf{A}_i \frac{\partial(\mathbf{E}^\perp + \mathbf{E}^\Phi)}{\partial x_i}\, d\mathbf{x}$ and applying the orthogonality property (2.2.6), we obtain

$$\begin{aligned}
\int_\omega \mathbf{v}^t \mathbf{A}_i \frac{\partial(\mathbf{E}^\perp + \mathbf{E}^\Phi)}{\partial x_i}\, d\mathbf{x} &= \frac{1}{h}\sum_{j=1}^d \int_\omega \left(L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{a}_i - L_p\left(\frac{x_j}{h}\right)\mathbf{b}_i\right)^t \mathbf{A}_i \\
&\qquad \left(L'_{p+1}\left(\frac{x_i}{h}\right)(\boldsymbol{\gamma}_i^\perp + \boldsymbol{\gamma}_i^\Phi) - L'_p\left(\frac{x_i}{h}\right)(\operatorname{sgn}(\mathbf{A}_i)\boldsymbol{\gamma}_i^\perp + \boldsymbol{\delta}_i^\Phi)\right) d\mathbf{x} \\
&= \frac{1}{h}\int_\omega L_p\left(\frac{x_i}{h}\right) L'_{p+1}\left(\frac{x_i}{h}\right)(\mathbf{A}_i\mathbf{b}_i)^t(\boldsymbol{\gamma}_i^\perp + \boldsymbol{\gamma}_i^\Phi)\, d\mathbf{x} \\
&= 0, \quad \forall\ \mathbf{v} \in \mathcal{E}_p,\ 1 \le i \le d,
\end{aligned} \tag{4.3.67}$$

where we used the fact that $\mathbf{b}_i \in \mathcal{N}(\mathbf{A}_i)$.

Subtracting (4.3.66) and (4.3.67) from (4.3.62) yields for $\boldsymbol{\epsilon} = \mathbf{e} - \mathbf{E}^\perp - \mathbf{E}^\Phi$ and $\boldsymbol{\epsilon}^- = \mathbf{e}^- - \mathbf{E}^-$

$$\int_\omega \mathbf{v}^t \left(\frac{\partial \boldsymbol{\epsilon}}{\partial t} + \sum_{j=1}^d \mathbf{A}_j \frac{\partial \boldsymbol{\epsilon}}{\partial x_j}\right) d\mathbf{x} = \sum_{j=1}^d \int_{\gamma_j} \mathbf{v}^t \nu_j \mathbf{A}_j^{\bar{\mu}_j}(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^-)\, d\mathbf{s},\ \forall\ \mathbf{v} \in \mathcal{E}_p. \tag{4.3.68}$$

By (4.3.19) we can write

$$\boldsymbol{\epsilon} = (\mathbf{e}^\perp - \mathbf{E}^\perp) + (\mathbf{e}^\Phi - \mathbf{E}^\Phi) + \mathcal{O}(h^{p+2}), \tag{4.3.69}$$

which, since $\mathbf{e}^\perp - \mathbf{E}^\perp = \mathcal{O}(h^{p+2})$ by Theorem 4.3.2, infers

$$\boldsymbol{\epsilon} = (\mathbf{e}^\Phi - \mathbf{E}^\Phi) + \mathcal{O}(h^{p+2}). \tag{4.3.70}$$

We will now show that $\boldsymbol{\epsilon} = \mathcal{O}(h^{p+2})$.

Since, by definition, $\mathbf{e}^{\Phi} - \mathbf{E}^{\Phi} \in \mathcal{E}_p = \bar{\mathcal{E}}_p \oplus \bar{\mathcal{E}}_p{}^{\perp}$, we can split $\boldsymbol{\epsilon}$ into

$$\boldsymbol{\epsilon} = \bar{\boldsymbol{\epsilon}} + \bar{\boldsymbol{\epsilon}}^{\perp} + \mathcal{O}(h^{p+2}),\ \bar{\boldsymbol{\epsilon}} \in \bar{\mathcal{E}}_p,\ \bar{\boldsymbol{\epsilon}}^{\perp} \in \bar{\mathcal{E}}_p{}^{\perp}, \tag{4.3.71}$$

where $\bar{\boldsymbol{\epsilon}}, \bar{\boldsymbol{\epsilon}}^{\perp}$ are the projections of $(\mathbf{e}^{\Phi} - \mathbf{E}^{\Phi})$ into $\bar{\mathcal{E}}_p$ and $\bar{\mathcal{E}}_p{}^{\perp}$.

First, we show that $\bar{\boldsymbol{\epsilon}}^{\perp} = \mathcal{O}(h^{p+2})$.

By property (1.2.25b) and the fact that $\frac{\partial}{\partial\tau}\bar{\boldsymbol{\epsilon}}^{\perp} \in \bar{\mathcal{E}}_p{}^{\perp}$, we have $\mathbf{A}_i^s \frac{\partial}{\partial\tau}\bar{\boldsymbol{\epsilon}}^{\perp} = \mathbf{0},\ s = +, -,$ $1 \leq i \leq d$. Thus, substituting $\frac{\partial}{\partial\tau}\bar{\boldsymbol{\epsilon}}^{\perp}$ for $\mathbf{v}$ in (4.3.68), the right side vanishes, and we obtain

$$\int_\omega \left(\frac{\partial\bar{\boldsymbol{\epsilon}}^{\perp}}{\partial t}\right)^t \frac{\partial\boldsymbol{\epsilon}}{\partial t}\, d\mathbf{x} = \int_\omega \left(\frac{\partial\bar{\boldsymbol{\epsilon}}^{\perp}}{\partial t}\right)^t \left(\frac{\partial\bar{\boldsymbol{\epsilon}}^{\perp}}{\partial t} + \mathcal{O}(h^{p+2})\right) d\mathbf{x} = 0, \tag{4.3.72}$$

which, by applying the Cauchy-Schwarz inequality, infers

$$\left\|\frac{\partial\bar{\boldsymbol{\epsilon}}^{\perp}}{\partial t}\right\|_{2,\omega}^2 = -\int_\omega \left(\frac{\partial\bar{\boldsymbol{\epsilon}}^{\perp}}{\partial t}\right)^t \mathcal{O}(h^{p+2})\, d\mathbf{x} \leq Ch^{p+2}|\omega|^{1/2} \left\|\frac{\partial\bar{\boldsymbol{\epsilon}}^{\perp}}{\partial t}\right\|_{2,\omega}. \tag{4.3.73}$$

Dividing (4.3.73) by $\|\frac{\partial\bar{\boldsymbol{\epsilon}}^{\perp}}{\partial t}\|_{2,\omega}$ yields

$$\left\|\frac{\partial\bar{\boldsymbol{\epsilon}}^{\perp}}{\partial t}\right\|_{2,\omega} \leq Ch^{p+2}|\omega|^{1/2}. \tag{4.3.74}$$

Applying inverse inequality (1.2.65) to (4.3.74), we obtain

$$\left\|\frac{\partial\bar{\boldsymbol{\epsilon}}^{\perp}}{\partial t}\right\|_{\infty,\omega} \leq C'|\omega|^{-1/2} \left\|\frac{\partial\bar{\boldsymbol{\epsilon}}^{\perp}}{\partial t}\right\|_{2,\omega} \leq C''h^{p+2}. \tag{4.3.75}$$

By initial conditions (4.3.39), $\mathbf{E}^{\Phi}(0,\mathbf{x}) = \mathbf{e}^{\Phi}(0,\mathbf{x}),\ \mathbf{x} \in \omega$, thus

$$(\bar{\boldsymbol{\epsilon}} + \bar{\boldsymbol{\epsilon}}^{\perp})(0,\mathbf{x}) = (\mathbf{e}^{\Phi} - \mathbf{E}^{\Phi})(0,\mathbf{x}) = \mathbf{0},\ \mathbf{x} \in \omega. \tag{4.3.76}$$

Thus, $\bar{\boldsymbol{\epsilon}}^{\perp}(0,\mathbf{x}) = \mathbf{0},\ \mathbf{x} \in \omega$, which together with (4.3.75) and the Fundamental Theorem of Calculus yields $\bar{\boldsymbol{\epsilon}}^{\perp} = \mathcal{O}(h^{p+2})$, and therefore

$$\boldsymbol{\epsilon} = \bar{\boldsymbol{\epsilon}} + \mathcal{O}(h^{p+2}), \quad \bar{\boldsymbol{\epsilon}} \in \bar{\mathcal{E}}_p. \tag{4.3.77}$$

Applying the linear transformations $t = T\tau,\ T > 0$, and $\mathbf{x} = h\boldsymbol{\xi}$, (4.3.68) becomes

$$\int_\Delta \mathbf{v}^t \left(\frac{h}{T}\frac{\partial\hat{\boldsymbol{\epsilon}}}{\partial\tau} + \sum_{j=1}^d \mathbf{A}_j \frac{\partial\hat{\boldsymbol{\epsilon}}}{\partial\xi_j}\right) d\boldsymbol{\xi} = \sum_{j=1}^d \int_{\Gamma_j} \mathbf{v}^t \nu_j \mathbf{A}_j^{\bar{\mu}_j} (\hat{\boldsymbol{\epsilon}} - \hat{\boldsymbol{\epsilon}}^-)\, d\boldsymbol{\sigma},\ \forall\ \mathbf{v} \in \mathcal{E}_p, \tag{4.3.78}$$

where $\hat{\boldsymbol{\epsilon}}(\tau, \boldsymbol{\xi}) = \boldsymbol{\epsilon}(T\tau, h\boldsymbol{\xi})$.

The Maclaurin series of $\bar{\boldsymbol{\epsilon}} \in \bar{\mathcal{E}}_p$ with respect to h is

$$\bar{\boldsymbol{\epsilon}}(t, h\boldsymbol{\xi}) = \sum_{k=0}^{\infty} h^k \mathbf{q}_k(t, \boldsymbol{\xi}),\ \mathbf{q}_k \in \bar{\mathcal{E}}_p,\ k \geq 0, \tag{4.3.79}$$

which together with (4.3.77) yields

$$\hat{\boldsymbol{\epsilon}}(t, \boldsymbol{\xi}) = \sum_{k=0}^{p+1} h^k \mathbf{q}_k(t, \boldsymbol{\xi}) + \mathcal{O}(h^{p+2}). \tag{4.3.80}$$

By (4.3.38) and (4.3.40),

$$\hat{\boldsymbol{\epsilon}}^-(t, \boldsymbol{\xi}) = \mathcal{O}(h^{p+2}). \tag{4.3.81}$$

Substituting (4.3.80) and (4.3.81) into (4.3.68) yields

$$\sum_{k=0}^{p+1} h^k \left(\int_\Delta \mathbf{v}^t \left(\frac{h}{T} \frac{\partial \mathbf{q}_k}{\partial \tau} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{q}_k}{\partial \xi_j} \right) d\boldsymbol{\xi} - \sum_{j=1}^{d} \int_{\Gamma_j} \mathbf{v}^t \nu_j \mathbf{A}_j^{\bar{\mu}_j} \mathbf{q}_k \, d\boldsymbol{\sigma} \right) = \mathcal{O}(h^{p+2}),\ \forall\ \mathbf{v} \in \mathcal{E}_p, \tag{4.3.82}$$

which infers that all terms of the same power in h are zero.

The $\mathcal{O}(1)$ term leads to the orthogonality condition for $\mathbf{q}_0$,

$$\sum_{j=1}^{d} \left(\int_\Delta \mathbf{v}^t \mathbf{A}_j \frac{\partial \mathbf{q}_0}{\partial \xi_j} \, d\boldsymbol{\xi} - \int_{\Gamma_j} \mathbf{v}^t \nu_j \mathbf{A}_j^{\bar{\mu}_j} \mathbf{q}_0 \, d\boldsymbol{\sigma} \right) = 0,\ \forall\ \mathbf{v} \in \mathcal{E}_p. \tag{4.3.83}$$

Since $\mathbf{q}_0 \in \bar{\mathcal{E}}_p$ by (4.3.79) and satisfies (4.3.83), Lemma 4.3.3 infers $\mathbf{q}_0 = \mathbf{0}$.

Using induction, we assume that $\mathbf{q}_l = \mathbf{0},\ 0 \leq l \leq k-1,\ k \leq p+1$, and apply the $\mathcal{O}(h^k)$ term to obtain the orthogonality condition

$$\sum_{j=1}^{d} \left(\int_\Delta \mathbf{v}^t \mathbf{A}_j \frac{\partial \mathbf{q}_k}{\partial \xi_j} \, d\boldsymbol{\xi} - \int_{\Gamma_j} \mathbf{v}^t \nu_j \mathbf{A}_j^{\bar{\mu}_j} \mathbf{q}_k \, d\boldsymbol{\sigma} \right) = 0,\ \forall\ \mathbf{v} \in \mathcal{E}_p, \tag{4.3.84}$$

which, by Lemma 4.3.3 and (4.3.79), infers $\mathbf{q}_k = \mathbf{0},\ k \leq p+1$.

Substituting $\mathbf{q}_k = \mathbf{0},\ k \leq p+1$ into (4.3.80) yields $\hat{\boldsymbol{\epsilon}} = \mathcal{O}(h^{p+2})$, which, when substituted into (4.3.70), yields (4.3.60). This completes the proof. □

4.4 Computational Examples

We start with an example that validates the superconvergence results of Theorem 4.3.1 for $d = 3$. Then, we apply our error estimation procedure to several problems to show that computations and theory are in full agreement. We use the L^2-projection $\Pi \mathbf{u}_0$ to approximate the initial conditions and π_i, $1 \leq i \leq d$ to approximate the boundary conditions.

4.4.1 Example for Superconvergence

Example 4.4.1. *Let us consider the three-dimensional system*

$$\mathbf{u}_{,t} + \mathbf{A}_1\mathbf{u}_{,x} + \mathbf{A}_2\mathbf{u}_{,y} + \mathbf{A}_3\mathbf{u}_{,z} = \mathbf{g}(t,x,y,z),\ (x,y,z) \in (0,1)^3,\ 0 < t \leq 1, \tag{4.4.1a}$$

where

$$\mathbf{A}_1 = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 2 \end{pmatrix},\ \mathbf{A}_2 = \begin{pmatrix} -1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 3 \end{pmatrix},\ \mathbf{A}_3 = \begin{pmatrix} -1 & 0 & 2 \\ 0 & -1 & 0 \\ 2 & 0 & -1 \end{pmatrix}, \tag{4.4.1b}$$

and select $\mathbf{g}(t,x,y,z)$*, the initial and boundary conditions such that the true solution is*

$$\mathbf{u} = \begin{pmatrix} \exp(t+x-y-z) \\ \exp(t-x+y-z) \\ \exp(t-x-y+z) \end{pmatrix}. \tag{4.4.1c}$$

Basic linear algebra leads to the following results

$$\mathcal{R}(\mathbf{A}_1^+) = \mathbb{R}^3,\ \mathcal{R}(\mathbf{A}_1^-) = \{\mathbf{0}\}, \tag{4.4.2}$$

$$\mathcal{R}(\mathbf{A}_2^-) = \text{span}\{\mathbf{z}_1\},\ \mathcal{R}(\mathbf{A}_2^+) = \mathcal{R}(\mathbf{A}_2^-)^\perp,\ \mathbf{z}_1 = \begin{pmatrix} -0.9260656482554513 \\ 0.1480943063089058 \\ 0.3470885932439939 \end{pmatrix}, \tag{4.4.3}$$

and

$$\mathcal{R}(\mathbf{A}_3^+) = \text{span}\{\mathbf{z}_2\},\ \mathcal{R}(\mathbf{A}_3^-) = \mathcal{R}(\mathbf{A}_3^+)^\perp,\ \mathbf{z}_2 = (1,0,1)^t. \tag{4.4.4}$$

The spaces $\mathcal{R}(\mathbf{A}_2^-)$ and $\mathcal{R}(\mathbf{A}_3^+)$ are two planes whose intersection is the line defined as $\text{span}\{\mathbf{z}\}$, where

$$\mathbf{z} = \frac{\mathbf{z}_1 \times \mathbf{z}_2}{\|\mathbf{z}_1 \times \mathbf{z}_2\|} = \begin{pmatrix} 0.1147781473323876 \\ 0.9867380370644933 \\ -0.1147781473323876 \end{pmatrix} \in \mathcal{R}(A_1^+) \cap \mathcal{R}(A_2^+) \cap \mathcal{R}(A_3^-). \tag{4.4.5}$$

Applying the superconvergence result (3.2.1) we show that $\mathbf{z}^t\mathbf{e}$ is $\mathcal{O}(h^{p+2})$ at the shifted Radau points $(\xi_i^+, \xi_j^+, \xi_k^-)$.

Next, we solve (4.4.1) on uniform meshes having $4^3, 6^3, 8^3, 10^3$ square elements and $p = 0, 1, 2, 3$, and present in Table 4.4.1 the maximum errors $|\mathbf{z}^t\mathbf{e}|$ at shifted Radau points over all elements. These results validate the superconvergence theory of Theorem 3.2.1.

	$p=0$		$p=1$		$p=2$		$p=3$	
N	$\lvert\mathbf{z}^t\mathbf{e}\rvert$	*order*	$\lvert\mathbf{z}^t\mathbf{e}\rvert$	*order*	$\lvert\mathbf{z}^t\mathbf{e}\rvert$	*order*	$\lvert\mathbf{z}^t\mathbf{e}\rvert$	*order*
4^3	0.1320	–	1.566*e*-2	–	6.086*e*-4	–	5.147*e*-5	–
6^3	0.1067	0.5253	5.259*e*-3	2.6904	1.403*e*-4	3.6189	7.646*e*-6	4.7030
8^3	8.419*e*-2	0.8225	2.287*e*-3	2.8948	4.808*e*-5	3.7223	1.929*e*-6	4.7878
10^3	7.472*e*-2	0.5350	1.256*e*-3	2.6839	2.066*e*-5	3.7863	6.568*e*-7	4.8271

Table 4.4.1: Maximum errors $|\mathbf{z}^t\mathbf{e}|$, $\mathbf{z}$ given by (4.4.5), at shifted Radau points $(\xi_i^+, \xi_j^+, \xi_k^-)$ and $t = 1$ over all elements for Example 4.4.1.

4.4.2 Examples for *A Posteriori* Error Estimation

Example 4.4.2. *Let us consider three-dimensional wave equation*

$$\frac{\partial^2 v}{\partial t^2} = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2},\ (x,y,z) \in (0,1)^3,\ 0 < t \leq 1, \tag{4.4.6}$$

which can be written as the first-order linear hyperbolic system

$$\mathbf{u}_{,t} + \mathbf{A}_1\mathbf{u}_{,x} + \mathbf{A}_2\mathbf{u}_{,y} + \mathbf{A}_3\mathbf{u}_{,z} = 0,\ (x,y,z) \in (0,1)^3,\ 0 < t \leq 1, \tag{4.4.7a}$$

where

$$\mathbf{u} = \begin{pmatrix} v_{,t} + v_{,x} \\ v_{,y} \\ v_{,z} \end{pmatrix},\ \mathbf{A}_1 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ \mathbf{A}_2 = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},\ \mathbf{A}_3 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \tag{4.4.7b}$$

and select initial and boundary conditions such that the true solution is

$$\mathbf{u} = [2, (\sqrt{3}-1), (\sqrt{3}-1)]^t \sin(\sqrt{3}t + x + y + z). \tag{4.4.7c}$$

While the matrix $\mathbf{A}_1$ is invertible and admits the eigenvalues $\{-1, 1, 1\}$, both $\mathbf{A}_2, \mathbf{A}_3$ are singular and admit the eigenvalues $\{-1, 0, 1\}$. Moreover, the eigenvectors vectors $(0,0,1)^t$ and $(0,1,0)^t$ are associated with the zero eigenvalue for $\mathbf{A}_2$ and $\mathbf{A}_3$, respectively. Applying our theory, the stationary error estimate $\mathbf{E}^\perp$ can only accurately estimate the component of error lying in $(\mathcal{N}(\mathbf{A}_1) \bigoplus \mathcal{N}(A_2) \bigoplus \mathcal{N}(A_3))^\perp = \text{span}\{(1,0,0)^t\}$, *i.e.,* only $E_1^\perp$ is an accurate estimate of e_1.

In order to validate our theory we solve (4.4.7) on uniform meshes having $N = 10^3, 15^3, 20^3, 30^3$ elements with $p = 1, 2, 3$ and show the stationary error estimates at $t = 1$ in Table 4.4.3. These results confirm our theory, *i.e.,* only $E_1^\perp$ is an accurate estimate of e_1. On the other-hand we observe that the transient error estimates $\mathbf{E}^\perp + \mathbf{E}^\Psi$ shown in Table 4.4.2 are accurate for all components of $\mathbf{e}$ and converge to the true error with mesh refinement.

Finally, we plot the global effectivity indices versus time in Figure 4.4.1 where we observe that the global effectivity indices for the transient *a posteriori* error estimate $\mathbf{E}^{\perp} + \mathbf{E}^{\Phi}$ oscillate near $t = 0$ before approaching unity with increasing time for $\Pi\mathbf{u}_0$. However, effectivity indices stay close to unity at all times when using $\pi\mathbf{u}_0$.

p	N	$\|\mathbf{e}\|_{2,\Omega}$	*order*	$\|\mathbf{e} - \mathbf{E}^{\perp} - \mathbf{E}^{\Phi}\|_{2,\Omega}$	*order*	θ
1	10^3	1.0690*e*-3	–	1.7744*e*-4	–	0.9897
	15^3	4.7450*e*-4	2.003	5.4945*e*-5	2.891	0.9949
	20^3	2.6690*e*-4	2	2.4042*e*-5	2.873	0.9969
	30^3	1.1867*e*-4	1.999	7.5900*e*-6	2.844	0.9985
2	10^3	1.6675*e*-5	–	1.0085*e*-6	–	0.998
	15^3	4.9355*e*-6	3.003	2.0831*e*-7	3.89	0.9988
	20^3	2.0813*e*-6	3.001	6.8713*e*-8	3.855	0.9992
	30^3	6.1650*e*-7	3.001	1.4723*e*-8	3.799	0.9994
3	10^3	5.9039*e*-8	–	2.7592*e*-8	–	0.8923
	15^3	1.0998*e*-8	4.145	3.6731*e*-9	4.973	0.9463
	20^3	3.4026*e*-9	4.078	8.7744*e*-10	4.977	0.9684

Table 4.4.2: L^2-errors $\|\mathbf{e}\|_{2,\Omega}$, $\|\mathbf{e} - \mathbf{E}^{\perp} - \mathbf{E}^{\Phi}\|_{2,\Omega}$ and their order of convergence. Global effectivity indices corresponding to transient estimates for Example 4.4.2 at $t = 1$ using $\Pi\mathbf{u}_0$.

Example 4.4.3. *Let us consider Maxwell's equations of electromagnetism,*

$$\varepsilon_0 \frac{\partial \boldsymbol{\mathcal{E}}}{\partial t} = \nabla \times \boldsymbol{\mathcal{H}}, \quad \nabla \cdot \boldsymbol{\mathcal{E}} = 0, \tag{4.4.8a}$$

$$\mu_0 \frac{\partial \boldsymbol{\mathcal{H}}}{\partial t} = \nabla \times \boldsymbol{\mathcal{E}}, \quad \nabla \cdot \boldsymbol{\mathcal{H}} = 0, \tag{4.4.8b}$$

where $\boldsymbol{\mathcal{E}}(t, \mathbf{x}) = (\mathcal{E}_x, \mathcal{E}_y, \mathcal{E}_z)^t$ *and* $\boldsymbol{\mathcal{H}}(t, \mathbf{x}) = (\mathcal{H}_x, \mathcal{H}_y, \mathcal{H}_z)^t$ *denote the* electric and magnetic field *and* $\mu_0 = 4\pi \cdot 10^{-7}\, NA^{-2}$ *and* $\varepsilon_0 = c_0^{-2}\mu_0^{-1}$ *denote the* magnetic and electric permittivity in free space*, respectively, with* $c_0 = 299,792,458\, ms^{-2}$ *being the* speed of light*.*

For a transverse electric wave traveling in the x_1x_2*-plane,* $\mathcal{E}_z = \mathcal{H}_x = \mathcal{H}_y = 0$*. If we choose space and time units such that* $c_0 = 1$*, (4.4.8) yields the symmetrizable hyperbolic system*

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}_1 \frac{\partial \mathbf{u}}{\partial x_1} + \mathbf{A}_2 \frac{\partial \mathbf{u}}{\partial x_2} = \mathbf{g}, \ \mathbf{x} \in \Omega = (0,1)^2, 0 < t < 1, \tag{4.4.9a}$$

where

$$\mathbf{u} = \begin{pmatrix} \sqrt{\varepsilon_0}\mathcal{E}_x \\ \sqrt{\varepsilon_0}\mathcal{E}_y \\ \sqrt{\mu_0}\mathcal{H}_z \end{pmatrix}, \ \mathbf{A}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \ \mathbf{A}_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}. \tag{4.4.9b}$$

p	N	$\|\mathbf{e}\|^*$	*order*	$\|\mathbf{e}-\mathbf{E}^\perp\|^*$	*order*	θ^*
1	10^3	7.7279*e*-4		1.4936*e*-4		0.9854
		5.2229*e*-4	–	1.7895*e*-4	–	0.9414
		5.2229*e*-4		1.7895*e*-4		0.9414
	15^3	3.4175*e*-4	2.0124	4.5616*e*-5	2.9253	0.9930
		2.3277*e*-4	1.9932	7.7992*e*-5	2.0483	0.9430
		2.3277*e*-4	1.9932	7.7992*e*-5	2.0483	0.9430
	20^3	1.9191*e*-4	2.0058	1.9691*e*-5	2.9202	0.9958
		1.3116*e*-4	1.9939	4.3639*e*-5	2.0184	0.9435
		1.3116*e*-4	1.9939	4.3639*e*-5	2.0184	0.9435
	30^3	8.5201*e*-5	2.0027	6.0422*e*-6	2.9137	0.9980
		5.8411*e*-5	1.9951	1.9367*e*-5	2.0036	0.9436
		5.8411*e*-5	1.9951	1.9367*e*-5	2.0036	0.9436
2	10^3	1.2286*e*-5		7.6622*e*-7		0.9987
		7.9715*e*-6	–	2.2178*e*-6	–	0.9599
		7.9715*e*-6		2.2178*e*-6		0.9599
	15^3	3.6364*e*-6	3.0027	1.5291*e*-7	3.9747	0.9993
		2.3596*e*-6	3.0024	6.5430*e*-7	3.0106	0.9604
		2.3596*e*-6	3.0024	6.5430*e*-7	3.0106	0.9604
	20^3	1.5335*e*-6	3.0014	4.8655*e*-8	3.9804	0.9995
		9.9509*e*-7	3.0013	2.7577*e*-7	3.0034	0.9606
		9.9509*e*-7	3.0013	2.7577*e*-7	3.0034	0.9606
	30^3	4.5422*e*-7	3.0008	9.6651*e*-9	3.9861	0.9997
		2.9475*e*-7	3.0008	8.1706*e*-8	3.0001	0.9607
		2.9475*e*-7	3.0008	8.1706*e*-8	3.0001	0.9607
3	10^3	4.3963*e*-8		2.1158*e*-8		0.8856
		2.7864*e*-8	–	1.5094*e*-8	–	0.8500
		2.7864*e*-8		1.5094*e*-8		0.8500
	15^3	8.1367*e*-9	4.1606	2.7924*e*-9	4.9945	0.9441
		5.2316*e*-9	4.1252	2.3889*e*-9	4.5465	0.8943
		5.2316*e*-9	4.1252	2.3889*e*-9	4.5465	0.8943
	20^3	2.5112*e*-9	4.0865	6.6341*e*-10	4.9960	0.9678
		1.6234*e*-9	4.0676	6.7499*e*-10	4.3934	0.9123
		1.6234*e*-9	4.0676	6.7499*e*-10	4.3934	0.9123

Table 4.4.3: Componentwise $L^2(\Omega)$-errors $\|\mathbf{e}\|^*$, $\|\mathbf{e}-\mathbf{E}^\perp\|^*$ at $t = 1$, their order of convergence and global effectivity indices θ^* corresponding to stationary estimates for Example 4.4.2 using $\Pi\mathbf{u}_0$.

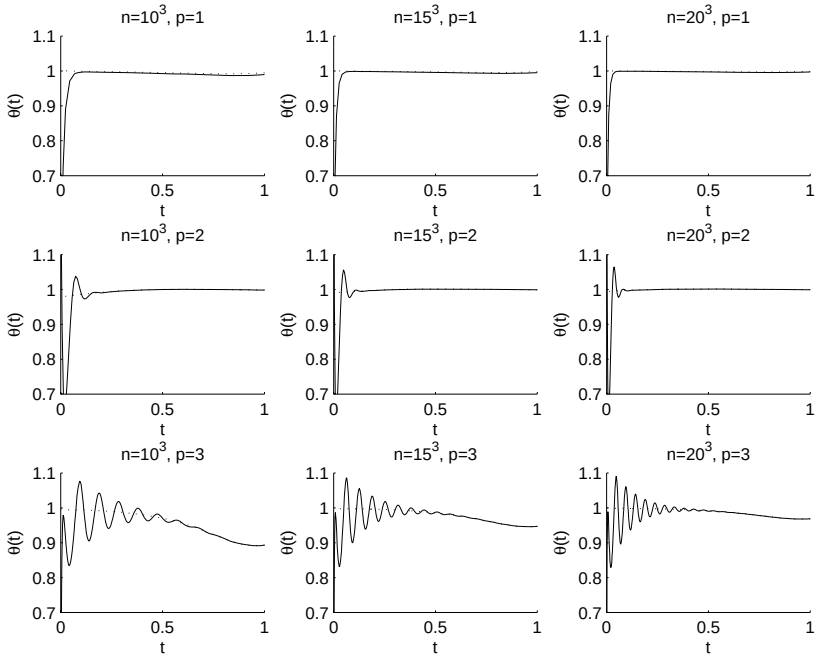

Figure 4.4.1: Global effectivity indices versus time using $\pi\mathbf{u}_0$ (dotted) and $\Pi\mathbf{u}_0$ (solid) for Example 4.4.2.

We will select initial and boundary conditions such that the true solution is

$$\mathbf{u}(t,\mathbf{x}) = (1,-1,\sqrt{2})^t \exp(t+\frac{x+y}{\sqrt{2}}),\ \forall\ 0\leq t\leq 1,\ \mathbf{x}\in\Omega. \tag{4.4.9c}$$

Both matrices $\mathbf{A}_1,\mathbf{A}_2$ are singular and admit the eigenvalues $\{-1,0,1\}$. Moreover, the eigenvectors $(1,0,0)^t$ and $(0,1,0)^t$ are associated with the zero eigenvalue for $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively. Applying our theory, the stationary error estimate $\mathbf{E}^\perp$ can only accurately approximate the component of the error lying in $(\mathcal{N}(\mathbf{A}_1)\bigoplus\mathcal{N}(A_2))^\perp = \text{span}\{(0,0,1)^t\}$, *i.e.,* only $E_3^\perp$ is an accurate estimate of e_3.

We further note, that (4.4.9) does not satisfy the conditions of Lemma 3.1.1, since both matrices are singular and $\mathbf{P}_{1,2}^t\mathbf{A}_2\mathbf{P}_{1,2} = \mathbf{P}_{2,2}^t\mathbf{A}_1\mathbf{P}_{2,2} = 0$.

To validate our theory, we solve (4.4.9) on uniform meshes having $N = 20^2, 30^2, 40^2$ elements for $p = 1,2,3$ using $\Pi\mathbf{u}_0$. We present the componentwise L^2-errors and effectivity indices corresponding to the stationary error estimate $\mathbf{E}^\perp$ at $t = 1$ in Table 4.4.4. In Table 4.4.5, we present the L^2-errors and effectivity indices for the transient error estimate $\mathbf{E}^\perp + \mathbf{E}^{\Pi}$ at $t = 1$. We observe that the effectivity indices for the transient error estimate and for the third component of the static estimate converge to unity under mesh refinement. Furthermore, we plot the effectivity indices for the transient error estimate versus time in Figure 4.4.2 to show that $\mathbf{E}^\perp + \mathbf{E}^{\Pi}$ is asymptotically accurate for $t = \mathcal{O}(1)$, which is in full agreement with

Theorem 4.3.2. We further note that the effectivity indices stay close to unity at all times when using $\pi\mathbf{u}_0$. For $\Pi\mathbf{u}_0$, the effectivity indices oscillates about unity near $t = 0$ before approaching unity.

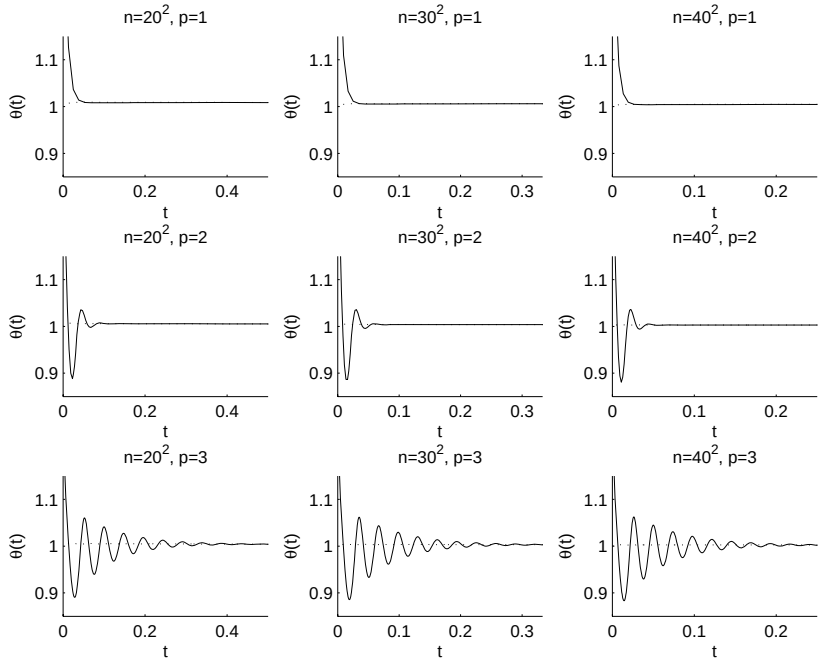

Figure 4.4.2: Global effectivity indices versus $0 \leq t \leq \frac{10}{n}$ using $\pi\mathbf{u}_0$ (dotted) and $\Pi\mathbf{u}_0$ (solid) for Example 4.4.3.

Example 4.4.4. *For a general electromagnetic wave, if we choose space and time units such that $c_0 = 1$, we can write* (4.4.8) *as*

$$\mathbf{u}_t + \mathbf{A}_1\mathbf{u}_{x_1} + \mathbf{A}_2\mathbf{u}_{x_2} + \mathbf{A}_3\mathbf{u}_{x_3} = \mathbf{0} \tag{4.4.10a}$$

where

$$\mathbf{u} = \begin{pmatrix} \sqrt{\varepsilon_0}\mathcal{E}_x \\ \sqrt{\varepsilon_0}\mathcal{E}_y \\ \sqrt{\varepsilon_0}\mathcal{E}_z \\ \sqrt{\mu_0}\mathcal{H}_x \\ \sqrt{\mu_0}\mathcal{H}_y \\ \sqrt{\mu_0}\mathcal{H}_z \end{pmatrix}, \quad \mathbf{A}_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{4.4.10b}$$

$$\mathbf{A}_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{A}_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{4.4.10c}$$

p	N	$\|\|\mathbf{e}\|\|^*$	*order*	$\|\|\mathbf{e}-\mathbf{E}^\perp\|\|^*$	*order*	θ^*
1	20^2	6.5522*e*-4	–	3.1339*e*-4	–	0.8869
		6.5522*e*-4		3.1339*e*-4		0.8869
		7.4984*e*-4		1.6742*e*-5		1.0098
	30^2	2.9219*e*-4	1.9917	1.3995*e*-4	1.9883	0.8835
		2.9219*e*-4	1.9917	1.3995*e*-4	1.9883	0.8835
		3.3401*e*-4	1.9944	5.1471*e*-6	2.9090	1.0064
	40^2	1.6464*e*-4	1.9941	7.8949*e*-5	1.9899	0.8818
		1.6464*e*-4	1.9941	7.8949*e*-5	1.9899	0.8818
		1.8808*e*-4	1.9963	2.2516*e*-6	2.8739	1.0049
2	20^2	1.8615*e*-6	–	9.2510*e*-7	–	0.8738
		1.8615*e*-6		9.2510*e*-7		0.8738
		2.1636*e*-6		4.3112*e*-8		1.0068
	30^2	5.5275*e*-7	2.9946	2.7502*e*-7	2.9918	0.8714
		5.5275*e*-7	2.9946	2.7502*e*-7	2.9918	0.8714
		6.4190*e*-7	2.9968	8.5587*e*-9	3.9877	1.0045
	40^2	2.3346*e*-7	2.9960	1.1626*e*-7	2.9929	0.8702
		2.3346*e*-7	2.9960	1.1626*e*-7	2.9929	0.8702
		2.7097*e*-7	2.9978	2.7180*e*-9	3.9872	1.0034
3	20^2	4.0497*e*-9	–	2.0511*e*-9	–	0.8674
		4.0497*e*-9		2.0511*e*-9		0.8674
		4.7423*e*-9		1.2019*e*-10		1.0056
	30^2	8.0124*e*-10	3.9960	4.0607*e*-10	3.9945	0.8655
		8.0124*e*-10	3.9960	4.0607*e*-10	3.9945	0.8655
		9.3760*e*-10	3.9978	1.5854*e*-11	4.9960	1.0038
	40^2	2.5374*e*-10	3.9969	1.2867*e*-10	3.9949	0.8644
		2.5374*e*-10	3.9969	1.2867*e*-10	3.9948	0.8644
		2.9680*e*-10	3.9984	3.7661*e*-12	4.9964	1.0029

Table 4.4.4: Componentwise $L^2(\Omega)$-errors $\|\mathbf{e}\|^*$, $\|\mathbf{e}-\mathbf{E}^\perp\|^*$ and their order of convergence. Global effectivity indices corresponding to static estimates for Example 4.4.3 at $t = 1$ using $\Pi\mathbf{u}_0$.

p	N	$\|\|\mathbf{e}\|\|$	*order*	$\|\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Pi}\|\|$	*order*	θ
1	10^2	4.7312*e*-3	–	2.1562*e*-4	–	1.014
	20^2	1.1920*e*-3	1.989	3.7742*e*-5	2.514	1.007
	30^2	5.3133*e*-4	1.993	1.4545*e*-5	2.352	1.005
	40^2	2.9931*e*-4	1.995	7.5490*e*-6	2.28	1.004
2	10^2	2.7124*e*-5	–	1.3917*e*-6	–	1.007
	20^2	3.4076*e*-6	2.993	1.2010*e*-7	3.535	1.003
	30^2	1.0115*e*-6	2.996	3.0819*e*-8	3.355	1.002
	40^2	4.2712*e*-7	2.997	1.1987*e*-8	3.283	1.001
3	10^2	1.1855*e*-7	–	7.7486*e*-9	–	1.005
	20^2	7.4357*e*-9	3.995	3.0928*e*-10	4.647	1.002
	30^2	1.4707*e*-9	3.997	5.0845*e*-11	4.453	1.001
	40^2	4.6568*e*-10	3.998	1.4522*e*-11	4.356	1

Table 4.4.5: $L^2(\Omega)$-errors $\|\mathbf{e}\|_{2,\Omega}$, $\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Pi}\|_{2,\Omega}$ and their order of convergence. Global effectivity indices corresponding to transient estimates for Example 4.4.3 at $t=1$ using $\Pi\mathbf{u}_0$.

We will choose the initial and boundary conditions, such that

$$\mathbf{u} = \begin{pmatrix} 1 & 4 & 1 & 3 & 0 & 3 \end{pmatrix}^t \cos\left(t + \frac{x_1+x_2+x_3}{\sqrt{3}}\right). \tag{4.4.10d}$$

The matrices $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$ each have eigenvalues $\{-1,-1,0,0,1,1\}$. Since $\bigcap_{i=1}^{3}\mathcal{N}(\mathbf{A}_i)^{\perp} = \{\mathbf{0}\}$, we expect that the stationary error estimate $\mathbf{E}^{\perp}$ will not accurately estimate any component of the error.

To show that the global effectivity index approaches 1 as $h \to 0$ for the transient error estimate, we plot the L_2 norms $\|\mathbf{e}\|_{2,\Omega}$, $\|\mathbf{e}-\mathbf{E}^{\perp}\|_{2,\Omega}$ and $\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Pi}\|_{2,\Omega}$, and the order of convergence and effectivity indices for both the static and transient error estimate of the solution at $t=1$ in Table 4.4.6. We observe that the effectivity indices for the transient error estimate converge to unity under mesh refinement. Furthermore, we plot the effectivity indices for the transient error estimate versus time in Figure 4.4.3 to note that $\mathbf{E}^{\perp}+\mathbf{E}^{\Pi}$ is asymptotically accurate, which is in full agreement with Theorem 4.3.2.

Example 4.4.5. *Let us consider the acoustic wave equation,*

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} - \nabla^2 p = 0, \tag{4.4.11}$$

where p is the acoustic pressure*, measured as the local deviation from the ambient pressure, and where c is the* speed of sound*. In two space dimensions,if we choose space and time*

p	N	$\|\mathbf{e}\|_{2,\Omega}$	*order*	$\|\mathbf{e}-\mathbf{E}^{\perp}\|_{2,\Omega}$	*order*	θ	$\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Psi}\|$	*order*	θ
1	5^3	2.7204e-3	–	1.5911e-3	–	0.8247	6.4114e-4	–	0.9373
	10^3	7.0738e-4	1.943	4.2000e-4	1.922	0.8099	1.1585e-4	2.468	0.9468
	15^3	3.2008e-4	1.956	1.9281e-4	1.92	0.8013	4.4673e-5	2.35	0.9531
	20^3	1.8197e-4	1.963	1.1072e-4	1.928	0.7957	2.3132e-5	2.288	0.9578
2	5^3	5.7423e-5	–	2.4109e-5	–	0.9042	4.0141e-6	–	0.9911
	10^3	7.1593e-6	3.004	3.0060e-6	3.004	0.9059	3.3614e-7	3.578	0.9943
	15^3	2.1202e-6	3.001	8.9145e-7	2.998	0.9062	8.3428e-8	3.437	0.9955
	20^3	8.9430e-7	3.001	3.7643e-7	2.997	0.9062	3.1751e-8	3.358	0.9961
3	5^3	2.2997e-7	–	1.7503e-7	–	0.6743	1.4469e-7	–	0.7627
	10^3	1.2713e-8	4.177	8.4473e-9	4.373	0.7588	4.8993e-9	4.884	0.8877
	15^3	2.4573e-9	4.054	1.5742e-9	4.144	0.7751	6.9336e-10	4.822	0.9228
	20^3	7.7323e-10	4.019	4.8982e-10	4.058	0.7788	1.7721e-10	4.742	0.9378

Table 4.4.6: Example 4.4.4: L^2 errors $\|\mathbf{e}\|_{2,\Omega}$, $\|\mathbf{e}-\mathbf{E}^{\perp}\|_{2,\Omega}$ and $\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Psi}\|_{2,\Omega}$ at $t=1$, their order of convergence and global effectivity indices θ for Example 4.4.4 using $\Pi\mathbf{u}_0$.

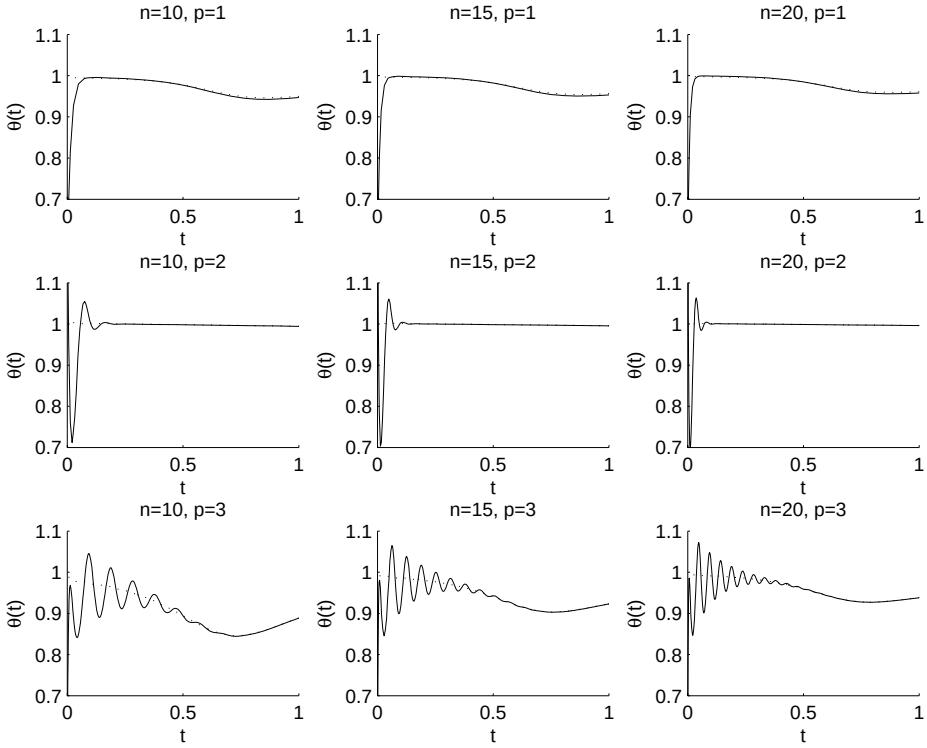

Figure 4.4.3: Global effectivity indices versus time using $\pi\mathbf{u}_0$ (dotted) and $\Pi\mathbf{u}_0$ (solid) for Example 4.4.4.

units such that $c = 1$, (4.4.11) can be written as the symmetric hyperbolic system

$$\mathbf{u}_{,t} + \mathbf{A}_1\mathbf{u}_{,x} + \mathbf{A}_2\mathbf{u}_{,y} = 0,\ (x,y) \in (0,1)^2,\ 0 < t \leq 1, \tag{4.4.12a}$$

where

$$\mathbf{u} = \begin{pmatrix} p_{,t} \\ -p_{,x} \\ -p_{,y} \end{pmatrix},\ \mathbf{A}_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},\ \mathbf{A}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{4.4.12b}$$

We select initial and boundary conditions such that the true solution is

$$p = \sin(t - x - y). \tag{4.4.12c}$$

Both matrices $\mathbf{A}_1, \mathbf{A}_2$ are singular and admit the eigenvalues $\{-1, 0, 1\}$. Moreover, the eigenvectors $(0,1,0)^t$ and $(0,0,1)^t$ are associated with the zero eigenvalue for $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively. Applying our theory, the stationary error estimate $\mathbf{E}^\perp$ can only accurately approximate the component of the error lying in $(\mathcal{N}(\mathbf{A}_1) \bigoplus \mathcal{N}(A_2))^\perp = \text{span}\{(1,0,0)^t\}$, *i.e.,* only $E_1^\perp$ is an accurate estimate of e_1. We solve (4.4.12) on uniform meshes having $N = 10^2, 20^2, 30^2$ elements for $p = 1, 2, 3$ using $\Pi\mathbf{u}_0$ and present the L^2 errors and effectivity indices corresponding to the transient error estimate $\mathbf{E}^\perp + \mathbf{E}^{\maltese}$ at $t = 1$ in Table 4.4.8. We observe that the effectivity indices converge to unity under mesh refinement.

We plot the L^2 errors and effectivity indices corresponding to the static error estimate $\mathbf{E}^\perp$ for each component in Table 4.4.7. We observe that $\mathbf{E}^\perp$ is fully approximates the error only in the third component, which is in full agreement with Theorem 4.3.4.

In Figure 4.4.4 we plot the global effectivity indices for the transient error estimate versus time. We note that the error estimate $\mathbf{E}^\perp + \mathbf{E}^{\maltese}$ is asymptotically accurate at $t = \mathcal{O}(1)$ under mesh refinement, and that the effectivity indices stay close to unity at all times when using $\pi\mathbf{u}_0$. For $\Pi\mathbf{u}_0$ the effectivity indices oscillates about unity near $t = 0$ before approaching unity.

p	N	$\|\|\mathbf{e}\|\|^*$	*order*	$\|\|\mathbf{e}-\mathbf{E}^\perp\|\|^*$	*order*	θ^*
1	10^2	1.3359*e*-4		1.6607*e*-5		1.0262
		1.4760*e*-4	–	1.0632*e*-4	–	0.7140
		1.4760*e*-4		1.0632*e*-4		0.7140
	20^2	3.3812*e*-5	1.9822	2.1585*e*-6	2.9437	1.0140
		3.7917*e*-5	1.9607	2.7574*e*-5	1.9470	0.6956
		3.7917*e*-5	1.9607	2.7574*e*-5	1.9470	0.6956
	30^2	1.5099*e*-5	1.9883	6.5571*e*-7	2.9385	1.0096
		1.7038*e*-5	1.9730	1.2455*e*-5	1.9601	0.6883
		1.7038*e*-5	1.9730	1.2455*e*-5	1.9601	0.6883
2	10^2	1.8992*e*-6		3.5861*e*-8		0.9959
		1.6059*e*-6	–	7.4800*e*-7	–	0.8810
		1.6059*e*-6		7.4800*e*-7		0.8810
	20^2	2.3711*e*-7	3.0017	2.3493*e*-9	3.9321	0.9980
		2.0061*e*-7	3.0009	9.3787*e*-8	2.9956	0.8821
		2.0061*e*-7	3.0009	9.3787*e*-8	2.9956	0.8821
	30^2	7.0230*e*-8	3.0009	4.7920*e*-10	3.9208	0.9987
		5.9436*e*-8	3.0002	2.7831*e*-8	2.9962	0.8824
		5.9436*e*-8	3.0002	2.7831*e*-8	2.9962	0.8824
3	10^2	3.4333*e*-9		4.1789*e*-10		1.0110
		3.6786*e*-9	–	2.6819*e*-9	–	0.6947
		3.6786*e*-9		2.6819*e*-9		0.6947
	20^2	2.1520*e*-10	3.9958	1.3074*e*-11	4.9984	1.0071
		2.3313*e*-10	3.9799	1.7088*e*-10	3.9722	0.6850
		2.3313*e*-10	3.9799	1.7088*e*-10	3.9722	0.6850
	30^2	4.2576*e*-11	3.9961	1.7230*e*-12	4.9981	1.0051
		4.6335*e*-11	3.9848	3.4077*e*-11	3.9765	0.6806
		4.6335*e*-11	3.9848	3.4077*e*-11	3.9765	0.6806

Table 4.4.7: Componentwise $L^2(\Omega)$-errors $\|\mathbf{e}\|^*$, $\|\mathbf{e}-\mathbf{E}^\perp\|^*$ and their order of convergence. Global effectivity indices corresponding to static estimates for Example 4.4.5 at $t = 1$ using $\Pi\mathbf{u}_0$.

p	N	$\|\|\mathbf{e}\|\|$	*order*	$\|\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Pi}\|\|$	*order*	θ
1	10^2	2.4782e-4	–	3.5072e-5	–	0.9685
	20^2	6.3394e-5	1.967	6.5200e-6	2.427	0.9717
	30^2	2.8435e-5	1.977	2.5723e-6	2.294	0.974
2	10^2	2.9606e-6	–	1.1041e-7	–	0.9927
	20^2	3.6975e-7	3.001	1.0393e-8	3.409	0.9956
	30^2	1.0953e-7	3	2.7126e-9	3.313	0.9967
3	10^2	6.2331e-9	–	1.0456e-9	–	0.9562
	20^2	3.9372e-10	3.985	4.5621e-11	4.518	0.9695
	30^2	7.8144e-11	3.988	7.8266e-12	4.348	0.9748

Table 4.4.8: L^2-errors $\|\mathbf{e}\|_{2,\Omega}$, $\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Pi}\|_{2,\Omega}$ and their order of convergence. Global effectivity indices corresponding to transient estimates for Example 4.4.5 at $t=1$ using $\Pi\mathbf{u}_0$.

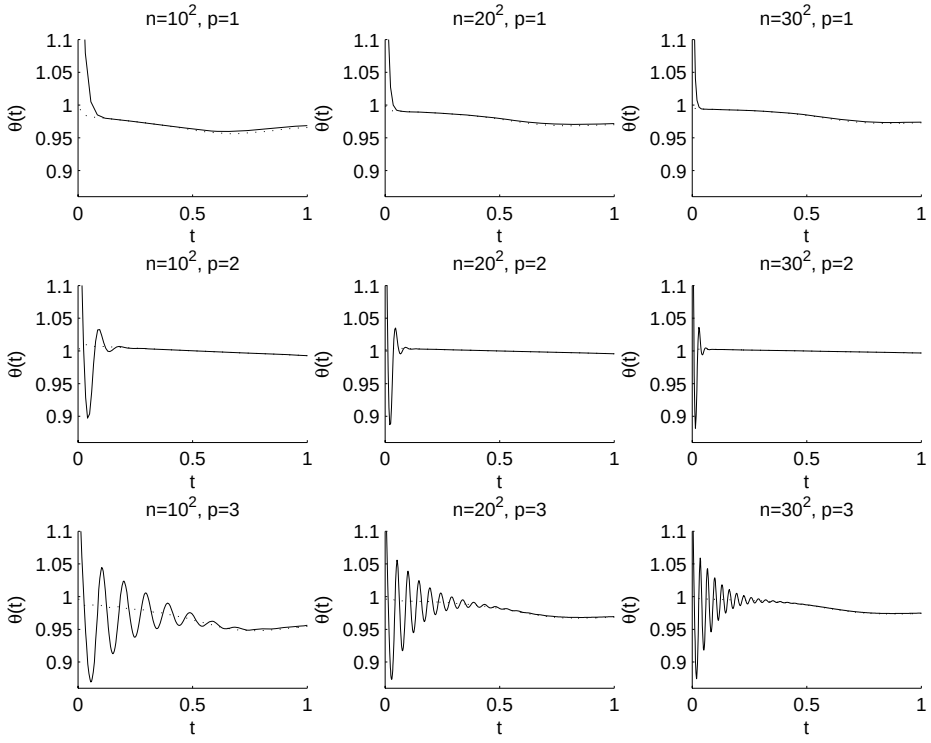

Figure 4.4.4: Global effectivity indices versus time using $\pi\mathbf{u}_0$ (dotted) and $\Pi\mathbf{u}_0$ (solid) for Example 4.4.5.

Chapter 5

Error Analysis for Linear Symmetrizable Hyperbolic Systems

In this chapter we will investigate the asymptotic error behavior for the DG Method on one element $\omega = (0,h)^d$ applied to symmetrizable hyperbolic systems with constant coefficients.

Thus, assume that there is a symmetric positive definite matrix $\mathbf{S}_0$ for which the matrices

$$\mathbf{S}_i = \mathbf{S}_0\mathbf{A}_i, \quad 1 \le i \le d, \tag{5.0.1}$$

are symmetric, and let $\mathbf{u} \in [C^2([0,T], C^{p+2}(\bar{\Omega}))]^m$ be the true solution of the symmetrizable hyperbolic system (1.3.1).

The DG method on ω consists of finding $\mathbf{u}_h \in \mathcal{P}_p$ that satisfies

$$\int_\omega \mathbf{v}^t \left(\frac{\partial \mathbf{u}_h}{\partial t} - \mathbf{g} \right) d\mathbf{x} = \sum_{j=1}^{d} \left(\int_\omega \frac{\partial \mathbf{v}^t}{\partial x_j} \mathbf{A}_j \mathbf{u}_h \, d\mathbf{x} - \int_{\partial\omega} \mathbf{v}^t \nu_j (\mathbf{A}_j^{\mu_j} \mathbf{u}_h^+ + \mathbf{A}_j^{\bar{\mu}_j} \mathbf{u}_h^-) \, d\mathbf{s} \right),$$

$$\forall\ \mathbf{v} \in \mathcal{P}_p,\ 0 < t < T, \tag{5.0.2a}$$

subject to the initial and boundary conditions

$$\mathbf{u}_h(0,\mathbf{x}) = \pi\mathbf{u}_0(\mathbf{x}) \text{ or } \mathbf{u}_h(0,\mathbf{x}) = \Pi\mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \omega, \tag{5.0.2b}$$

$$(\nu_i \mathbf{A}_i^{\bar{\mu}_i})\mathbf{u}_h^-(t,\mathbf{x}) = (\nu_i \mathbf{A}_i^{\bar{\mu}_i})\pi_i\mathbf{u}(t,\mathbf{x}), \quad \mathbf{x} \in \gamma_i,\ 1 \le i \le d,\ 0 < t < T. \tag{5.0.2c}$$

We will show that (1.3.1) and (5.0.2a) can be reduced to a symmetric hyperbolic system, for which the results of Chapter 4 hold. Thus, the leading term can again be expressed as a linear combination of Legendre polynomials of degree p and $p+1$. The superconvergence results hold also for the symmetrizable case. We then apply these asymptotic results and generalize the error estimation procedure of the previous chapters to obtain efficient and asymptotically exact estimates of the discretization error. Finally, we present some computational results for multi-dimensional systems such as Maxwell's equations and the acoustic equation.

5.1 Local Error Analysis

We will not show that (1.3.1) and (5.0.2a) can be reduced to a symmetric hyperbolic problem.

Since $\mathbf{S}_0$ is symmetric positive definite, we can use Cholesky factorization, which states that there exists a positive definite matrix $\mathbf{R}$ such that

$$\mathbf{S}_0 = \mathbf{R}^t\mathbf{R}. \tag{5.1.1}$$

We define

$$\mathbf{B}_i = \mathbf{R}\mathbf{A}_i\mathbf{R}^{-1},\ 1 \leq i \leq d. \tag{5.1.2a}$$

Since $\mathbf{B}_i$ is similar to $\mathbf{A}_i$, we can use Lemma 1.2.17 to show that

$$\mathbf{B}_i \text{ is symmetric,} \tag{5.1.2b}$$

$$\mathbf{B}_i^+ = \mathbf{R}\mathbf{A}_i^+\mathbf{R}^{-1}, \quad \mathbf{B}_i^- = \mathbf{R}\mathbf{A}_i^-\mathbf{R}^{-1}, \tag{5.1.2c}$$

and

$$\operatorname{sgn}(\mathbf{B}_i) = \mathbf{R}\operatorname{sgn}(\mathbf{A}_i)\mathbf{R}^{-1}, \quad 1 \leq i \leq d. \tag{5.1.2d}$$

Next, we define the approximation operators for the initial and boundary values in the following definition and lemma.

Definition 5.1.1. *Let $\tilde{\pi}\mathbf{v}$ denote the* corrected L^2-projection *of $\mathbf{v} \in [L^2(\omega)]^m$ onto $\mathcal{P}_p$, defined by*

$$\tilde{\pi}\mathbf{v}(\mathbf{x}) = \mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) - \sum_{i=1}^{d} \left(L_{p+1}\left(\frac{x_i}{h}\right)\bar{\mathbf{c}}_i - L_p\left(\frac{x_i}{h}\right) sgn(\mathbf{B}_i)\bar{\mathbf{c}}_i \right), \tag{5.1.3a}$$

and let $\tilde{\pi}_i^s\mathbf{v}$ be the corrected L^2-projection *of $\mathbf{v} \in [L^2(\gamma_i^s)]^m$ onto $\mathcal{P}_p$, defined by*

$$\tilde{\pi}_i^s\mathbf{v}(\mathbf{x}) = \mathcal{L}_{p+1}^{i,s}\mathbf{v}(\mathbf{x}) - \sum_{j\in D(i)} \left(L_{p+1}\left(\frac{x_j}{h}\right)\bar{\mathbf{c}}_{ij}^s - L_p\left(\frac{x_j}{h}\right) sgn(\mathbf{B}_j)\bar{\mathbf{c}}_{ij}^s \right), \quad 1 \leq i \leq d,\ s = +,-, \tag{5.1.3b}$$

where $\mathcal{L}_{p+1}$, $\bar{\mathbf{c}}_i$, $\mathcal{L}_{p+1}^{i,s}$ and $\bar{\mathbf{c}}_{ij}^s$ are defined in (2.2.15) and (2.2.16).

Lemma 5.1.2. *Let $\mathbf{R}$ be defined in (5.1.1), Π, π in (2.2.15), π_i^s in (2.2.16) and $\tilde{\pi}$, $\tilde{\pi}_i^s$ in Definition 5.1.1, $s = +,-$, $1 \leq i \leq d$. Then, for any function $\mathbf{v} \in [C^{p+2}(\bar{\omega})]^m$, we have*

$$\mathbf{R}(\Pi\mathbf{v})(\mathbf{x}) = \Pi(\mathbf{R}\mathbf{v})(\mathbf{x}), \quad \mathbf{x} \in \omega, \tag{5.1.4a}$$

$$\mathbf{R}(\pi\mathbf{v})(\mathbf{x}) = \tilde{\pi}(\mathbf{R}\mathbf{v})(\mathbf{x}), \quad \mathbf{x} \in \omega, \tag{5.1.4b}$$

and

$$\mathbf{R}(\pi_i^s\mathbf{v})(\mathbf{x}) = \tilde{\pi}_i^s(\mathbf{R}\mathbf{v})(\mathbf{x}), \quad \mathbf{x} \in \gamma_i,\ s = +,-,\ 1 \leq i \leq d. \tag{5.1.4c}$$

Proof. By (2.2.15),

$$\begin{aligned}\mathbf{R}(\mathcal{L}_{p+1}\mathbf{v})(\mathbf{x}) &= \mathbf{R}\sum_{|\boldsymbol{\alpha}|\leq p+1}\frac{\int_\omega \mathbf{v}(\mathbf{x})\psi_{\boldsymbol{\alpha}}\left(\frac{x_i}{h}\right)\,d\mathbf{x}}{\int_\omega \psi^2_{\boldsymbol{\alpha}}\left(\frac{x_i}{h}\right)\,d\mathbf{x}}\psi_{\boldsymbol{\alpha}}\left(\frac{x_i}{h}\right)\\ &= \sum_{|\boldsymbol{\alpha}|\leq p+1}\frac{\int_\omega \mathbf{R}\mathbf{v}(\mathbf{x})\psi_{\boldsymbol{\alpha}}\left(\frac{x_i}{h}\right)\,d\mathbf{x}}{\int_\omega \psi^2_{\boldsymbol{\alpha}}\left(\frac{x_i}{h}\right)\,d\mathbf{x}}\psi_{\boldsymbol{\alpha}}\left(\frac{x_i}{h}\right) = \mathcal{L}_{p+1}(\mathbf{R}\mathbf{v})(\mathbf{x}),\end{aligned} \tag{5.1.5}$$

and

$$\mathbf{R}\bar{\mathbf{c}}_i = \frac{\int_\omega \mathbf{R}\mathbf{v}(\mathbf{x})L_{p+1}\left(\frac{x_i}{h}\right)\,d\mathbf{x}}{\int_\omega L^2_{p+1}\left(\frac{x_i}{h}\right)\,d\mathbf{x}},\quad 1\leq i\leq d. \tag{5.1.6}$$

By substituting (5.1.5) and (5.1.6) into the definition of $\Pi(\mathbf{R}\mathbf{v})$ in (2.2.15b), we get

$$\begin{aligned}\Pi(\mathbf{R}\mathbf{v})(\mathbf{x}) &= \mathcal{L}_{p+1}(\mathbf{R}\mathbf{v})(\mathbf{x}) - \sum_{i=1}^d L_{p+1}\left(\frac{x_i}{h}\right)\left(\frac{\int_\omega \mathbf{R}\mathbf{v}(\mathbf{x})L_{p+1}\left(\frac{x_i}{h}\right)\,d\mathbf{x}}{\int_\omega L^2_{p+1}\left(\frac{x_i}{h}\right)\,d\mathbf{x}}\right)\\ &= \mathbf{R}\left(\mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) - \sum_{i=1}^d L_{p+1}\left(\frac{x_i}{h}\right)\bar{\mathbf{c}}_i\right) = \mathbf{R}(\Pi\mathbf{v})(\mathbf{x}).\end{aligned} \tag{5.1.7}$$

By substituting (5.1.5) and (5.1.6) into the definition of $\tilde{\pi}(\mathbf{R}\mathbf{v})$ in (5.1.3a), we get

$$\begin{aligned}\tilde{\pi}(\mathbf{R}\mathbf{v})\mathbf{x} &= \mathcal{L}_{p+1}(\mathbf{R}\mathbf{v})(\mathbf{x}) - \sum_{i=1}^d\left(L_{p+1}\left(\frac{x_i}{h}\right) - L_p\left(\frac{x_i}{h}\right)\mathrm{sgn}(\mathbf{B}_i)\right)\left(\frac{\int_\omega \mathbf{R}\mathbf{v}(\mathbf{x})L_{p+1}\left(\frac{x_i}{h}\right)\,d\mathbf{x}}{\int_\omega L^2_{p+1}\left(\frac{x_i}{h}\right)\,d\mathbf{x}}\right)\\ &= \mathbf{R}\left(\mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) - \sum_{i=1}^d\left(L_{p+1}\left(\frac{x_i}{h}\right)\bar{\mathbf{c}}_i - L_p\left(\frac{x_i}{h}\right)\mathrm{sgn}(\mathbf{A}_i)\bar{\mathbf{c}}_i\right)\right)\\ &= \mathbf{R}(\pi\mathbf{v})(\mathbf{x}),\end{aligned} \tag{5.1.8}$$

where we used the fact that $\mathrm{sgn}(\mathbf{B}_i) = \mathbf{R}\mathrm{sgn}(\mathbf{A}_i)\mathbf{R}^{-1}$ by (5.1.2d).

The proof of (5.1.4c) follows the same reasoning and is therefore omitted. □

Left-multiplying (1.3.1) by $\mathbf{R}$ and substituting $\mathbf{u} = \mathbf{R}^{-1}\mathbf{U}$ yields the symmetric system

$$\frac{\partial \mathbf{U}}{\partial t} + \sum_{i=1}^d \mathbf{B}_i\frac{\partial \mathbf{U}}{\partial x_i} = \mathbf{R}\mathbf{g}(t,\mathbf{x}),\quad \mathbf{x}\in\Omega,\ 0<t<T, \tag{5.1.9a}$$

subject to the initial and boundary conditions

$$\mathbf{U}(0,\mathbf{x}) = \mathbf{R}\mathbf{u}_0(\mathbf{x}),\quad \mathbf{x}\in\Omega, \tag{5.1.9b}$$

$$\left(\sum_{i=1}^d \nu_i\mathbf{B}_i^{\bar{\mu}_i}\right)\mathbf{U}(t,\mathbf{x}) = \left(\sum_{i=1}^d \nu_i\mathbf{B}_i^{\bar{\mu}_i}\right)\mathbf{R}\mathbf{u}(t,\mathbf{x}),\quad \mathbf{x}\in\partial\Omega,\ 0<t<T. \tag{5.1.9c}$$

Substituting $\mathbf{v} = \mathbf{R}^t\mathbf{w}$ and $\mathbf{u}_h = \mathbf{R}^{-1}\mathbf{U}_h$ in (5.0.2a) and using (5.1.2c) yields

$$\int_\omega \mathbf{w}^t \left(\frac{\partial \mathbf{U}_h}{\partial t} - \mathbf{R}\mathbf{g} \right) d\mathbf{x} = \sum_{j=1}^{d} \Bigg(\int_\omega \frac{\partial \mathbf{w}^t}{\partial x_j} \mathbf{B}_j \mathbf{U}_h \, d\mathbf{x} - \int_{\partial\omega} \mathbf{w}^t \nu_j (\mathbf{B}_j^{\mu_j} \mathbf{U}_h^+ + \mathbf{B}_j^{\bar{\mu}_j} \mathbf{U}_h^-) \, d\mathbf{s} \Bigg), \ \forall\ \mathbf{v} \in \mathcal{P}_p, \ 0 < t < T. \tag{5.1.10a}$$

Combining Definition 5.1.1 with (5.0.2b-c) yields the initial and boundary conditions

$$\mathbf{U}_h(0,\mathbf{x}) = \tilde{\pi}\mathbf{R}\mathbf{u}_0(\mathbf{x}) \text{ or } \mathbf{U}_h(0,\mathbf{x}) = \Pi\mathbf{R}\mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \omega, \tag{5.1.10b}$$

$$(\nu_i \mathbf{B}_i^{\bar{\mu}_i})\mathbf{U}_h^-(t,\mathbf{x}) = (\nu_i \mathbf{B}_i^{\bar{\mu}_i})\tilde{\pi}_i^s \mathbf{R}\mathbf{u}(t,\mathbf{x}), \quad \mathbf{x} \in \gamma_i^s, \ s = +,-, \ 1 \le i \le d, \ 0 < t < T. \tag{5.1.10c}$$

Now we can state the main theorem for the spatial discretization error.

Theorem 5.1.3. *Let $\mathbf{u} \in [C^2([0,T], C^{p+2}(\bar{\omega}))]^m$ be the solution of (1.3.1) and let $\mathbf{u}_h \in \mathcal{P}_p$ satisfy (5.0.2). Then the local finite element error on ω for $t = \mathcal{O}(1)$ and $p \ge 1$ can be written as*

$$\mathbf{u}(t, h\boldsymbol{\xi}) - \mathbf{u}_h(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} \mathbf{r}_i(t, h\xi_i) + \mathcal{O}(h^{p+2}), \ \boldsymbol{\xi} \in \Delta, \tag{5.1.11a}$$

where

$$\mathbf{r}_i(t, h\xi_i) = L_{p+1}(\xi_i)\mathbf{c}_i(t) - L_p(\xi_i)(sgn(\mathbf{A}_i)\mathbf{c}_i(t) + \mathbf{d}_i(t)), \ 1 \le i \le d, \tag{5.1.11b}$$

with

$$\mathbf{c}_i(t) = \frac{1}{a_{p+1}} \frac{1}{(p+1)!} \frac{\partial^{p+1}\mathbf{u}(t,\mathbf{0})}{\partial x_i^{p+1}}, \quad \mathbf{d}_i(t) \in \mathcal{N}(\mathbf{A}_i) \cap \bigoplus_{k=1}^{d} \mathcal{R}(\mathbf{A}_k). \tag{5.1.11c}$$

Proof. Since $\mathbf{u} \in [C^2([0,T], C^{p+2}(\bar{\omega}))]^m$ and $\mathbf{u}_h \in \mathcal{P}_p$, we have $\mathbf{U} \in [C^2([0,T], C^{p+2}(\bar{\omega}))]^m$ and $\mathbf{U}_h \in \mathcal{P}_p$. By (5.1.2b), $\mathbf{B}_i$ is symmetric for $1 \le i \le d$. Thus, $\mathbf{U}$ and $\mathbf{U}_h$, as defined in (5.1.9) and (5.1.10), satisfy the conditions of Theorem 4.2.1 and

$$\mathbf{U}(t, h\boldsymbol{\xi}) - \mathbf{U}_h(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} \tilde{\mathbf{r}}_i(t, h\xi_i) + \mathcal{O}(h^{p+2}), \ \boldsymbol{\xi} \in \Delta, \tag{5.1.12a}$$

where

$$\tilde{\mathbf{r}}_i(t, h\xi_i) = L_{p+1}(\xi_i)\tilde{\mathbf{c}}_i(t) - L_p(\xi_i)(\mathrm{sgn}(\mathbf{B}_i)\tilde{\mathbf{c}}_i(t) + \tilde{\mathbf{d}}_i(t)), \ 1 \le i \le d, \tag{5.1.12b}$$

with

$$\tilde{\mathbf{c}}_i(t) = \frac{1}{a_{p+1}} \frac{1}{(p+1)!} \frac{\partial^{p+1}\mathbf{U}(t,\mathbf{0})}{\partial x_i^{p+1}}, \quad \tilde{\mathbf{d}}_i(t) \in \mathcal{N}(\mathbf{B}_i) \cap \bigoplus_{k=1}^{d} \mathcal{R}(\mathbf{B}_k). \tag{5.1.12c}$$

Lemma 1.2.16i,ii yields that $\mathbf{d}_i = \mathbf{R}^{-1}\tilde{\mathbf{d}}_i \in \mathcal{N}(\mathbf{A}_i) \cap \bigoplus_{k=1}^{d} \mathcal{R}(\mathbf{A}_k)$. Then we obtain (5.1.11) by substituting $\mathbf{U} = \mathbf{R}\mathbf{u}$, $\mathbf{U}_h = \mathbf{R}\mathbf{u}_h$, $\tilde{\mathbf{r}}_i = \mathbf{R}\mathbf{r}_i$, $\tilde{\mathbf{c}}_i = \mathbf{R}\mathbf{c}_i$ and $\tilde{\mathbf{d}}_i = \mathbf{R}\mathbf{d}_i$, $1 \le i \le d$, into (5.1.12). ∎

5.2 Superconvergence and *A Posteriori* Error Estimation

The next theorem states superconvergence results for symmetrizable systems.

Theorem 5.2.1. *Under the conditions of Theorem 5.1.3 with $p \geq 1$ we let $\bar{\xi}_j^s,\ 1 \leq j \leq p+1$, denote the roots of $R_{p+1}^s(\xi),\ s = +, -$, shifted to $[0,1]$. Thus,*

i) If $\mathbf{z}$ is a unit vector in the union of the spaces $\bigcap_{i=1}^d \mathcal{N}(\mathbf{A}_i^{s_i})^\perp,\ s_i = +, -$, then the projection $\mathbf{z}^t\mathbf{e}(t,\mathbf{x})$ of the local error onto span$\{\mathbf{z}\}$ is $\mathcal{O}(h^{p+2})$ superconvergent at the points $(t, h\bar{\boldsymbol{\xi}})$, $\bar{\xi}_i = \bar{\xi}_{k_i}^{s_i},\ 1 \leq k_i \leq p+1,\ 1 \leq i \leq d$, $t = \mathcal{O}(1)$, i.e.,

$$\mathbf{z}^t\mathbf{e}(t, h\bar{\boldsymbol{\xi}}) = \mathcal{O}(h^{p+2}). \tag{5.2.1}$$

ii) Moreover, if $\gamma_i(a) = \{\mathbf{x} \in (0,h)^d : x_i = a\}$, $0 \leq a \leq h$, and if $\mathbf{v} \in \mathcal{P}_{p-1}$ is a unit vector with respect to the C^∞ norm, then, at $a = h\bar{\xi}_k^s$, we have the superconvergence of the following error averages

$$\frac{1}{h^{d-1}} \int_{\gamma_i(h\bar{\xi}_k^s)} \mathbf{v}^t \mathbf{A}_i^s \mathbf{e}\, d\mathbf{s} = \mathcal{O}(h^{p+2}),\ 1 \leq k \leq p+1,\ s = +, -,\ 1 \leq i \leq d, \tag{5.2.2}$$

and

$$\frac{1}{h^{d-1}} \int_{\gamma_i^s} \mathbf{v}^t (\mathbf{A}_i^{\mu_i}\mathbf{e} + \mathbf{A}_i^{\bar{\mu}_i}\mathbf{e}^-)\, d\mathbf{s} = \mathcal{O}(h^{p+2}),\ s = +, -,\quad 1 \leq i \leq d. \tag{5.2.3}$$

Proof. We will prove (5.2.1) for the case $s_i = +$, $1 \leq i \leq d$.
Thus, assume that there exists a unit vector $\mathbf{z} \in \bigcap_{i=1}^d \mathcal{N}(\mathbf{A}_i^+)^\perp$, which by Lemma 1.2.2 yields $\mathbf{z} \in \bigcap_{i=1}^d \mathcal{R}(\mathbf{A}_i^{+t})$, *i.e.*, there exists $\mathbf{v}_i$ such that

$$\mathbf{A}_i^{+t}\mathbf{v}_i = \mathbf{z},\ 1 \leq i \leq d. \tag{5.2.4}$$

Left pre-multiplying $\mathbf{e}$ in (5.1.11a) by $\mathbf{z}^t$ and evaluating the resulting function at the points $(t, h\bar{\boldsymbol{\xi}}),\ \bar{\boldsymbol{\xi}} = (\bar{\xi}_{k_1}^+, \ldots, \bar{\xi}_{k_d}^+),\ 1 \leq k_i \leq p+1,\ 1 \leq i \leq d$, we obtain

$$\mathbf{z}^t\mathbf{e}(t, h\bar{\boldsymbol{\xi}}) = h^{p+1} \sum_{i=1}^d \left(L_{p+1}(\bar{\xi}_{k_i}^+)\mathbf{z}^t\mathbf{c}_i - L_p(\bar{\xi}_{k_i}^+)(\mathbf{z}^t \mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i + \mathbf{z}^t\mathbf{d}_i) \right) + \mathcal{O}(h^{p+2}). \tag{5.2.5}$$

By the property (1.2.25b) and (5.1.11c) we have $\mathbf{d}_i \in \mathcal{N}(\mathbf{A}_i) \subseteq \mathcal{N}(\mathbf{A}_i^+)$, which yields by (5.2.4)

$$\mathbf{z}^t\mathbf{d}_i = \mathbf{v}_i^t \mathbf{A}_i^+ \mathbf{d}_i = \mathbf{0}. \tag{5.2.6}$$

Applying (5.2.4) and the property (1.2.25d) yields

$$\mathbf{z}^t \mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i = \mathbf{v}_i^t \mathbf{A}_i^+ \mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i = \mathbf{v}_i^t \mathbf{A}_i^+ \mathbf{c}_i = \mathbf{z}^t \mathbf{c}_i. \tag{5.2.7}$$

Substituting (5.2.6) and (5.2.7) into (5.2.5), we prove that

$$\mathbf{z}^t \mathbf{e}(t, h\bar{\boldsymbol{\xi}}) = h^{p+1} \sum_{i=1}^{d} R_{p+1}^+(\bar{\xi}_{k_i}^+) \mathbf{z}^t \mathbf{c}_i + \mathcal{O}(h^{p+2}) = \mathcal{O}(h^{p+2}). \tag{5.2.8}$$

Following the same line of reasoning we establish (5.2.1) for all other cases.

The proof of (5.2.2) and (5.2.2) is equal to the proof of Theorem 4.3.1, and will not be restated here. □

We will now present the *a posteriori* error estimation procedure for symmetrizable matrices, which differs in some ways from the symmetric case.

In Theorem 5.1.3 we showed that the local discretization error for the DG method on a physical element $\omega = (0, h)^d$ can be written as

$$\mathbf{e}(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} L_{p+1}(\xi_i)\mathbf{c}_i(t) - L_p(\xi_i)(\mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i(t) + \mathbf{d}_i(t)) + \mathcal{O}(h^{p+2}), \tag{5.2.9a}$$

where

$$\mathbf{d}_i(t) \in \mathcal{N}(\mathbf{A}_i) \cap \bigoplus_{k=1}^{d} \mathcal{R}(\mathbf{A}_k), \quad 1 \leq i \leq d. \tag{5.2.9b}$$

By Lemma 1.2.20, $\mathbf{c}_i^\perp = \mathbf{A}_i^\ddagger \mathbf{A}_i \mathbf{c}_i$ is the projection into $\mathcal{R}(\mathbf{A}_i)$ and $\mathbf{c}_i^\Phi = (\mathbf{I} - \mathbf{A}_i^\ddagger \mathbf{A}_i)\mathbf{c}_i$ is the projection into $\mathcal{N}(\mathbf{A}_i)$, thus

$$\mathbf{c}_i = \mathbf{c}_i^\perp + \mathbf{c}_i^\Phi, \quad \mathbf{c}_i^\perp \in \mathcal{R}(\mathbf{A}_i),\ \mathbf{c}_i^\Phi \in \mathcal{N}(\mathbf{A}_i). \tag{5.2.10}$$

Furthermore, by Lemma 1.2.20, $\mathcal{R}(\mathbf{A}_i) = \mathcal{R}(\mathrm{sgn}(\mathbf{A}_i))$ and $\mathcal{N}(\mathbf{A}_i) = \mathcal{N}(\mathrm{sgn}(\mathbf{A}_i))$, which yields $\mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i = \mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i^\perp \in \mathcal{R}(\mathbf{A}_i)$.

Hence, the leading term of the spatial discretization error can be split into two parts as

$$\mathbf{e} = \mathbf{e}^\perp + \mathbf{e}^\Phi + \mathcal{O}(h^{p+2}), \tag{5.2.11a}$$

where

$$\mathbf{e}^\perp(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} L_{p+1}(\xi_i)\mathbf{c}_i^\perp(t) - L_p(\xi_i)\mathrm{sgn}(\mathbf{A}_i)\mathbf{c}_i^\perp(t), \tag{5.2.11b}$$

$$\mathbf{e}^\Phi(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} L_{p+1}(\xi_i)\mathbf{c}_i^\Phi(t) - L_p(\xi_i)\mathbf{d}_i(t), \tag{5.2.11c}$$

and

$$\mathbf{c}_i^{\perp}, \operatorname{sgn}(\mathbf{A}_i)\mathbf{c}_i^{\perp} \in \mathcal{R}(\mathbf{A}_i), \quad \mathbf{c}_i^{\Psi}, \mathbf{d}_i^{\Psi} \in \mathcal{N}(\mathbf{A}_i),\ 1 \leq i \leq d. \tag{5.2.11d}$$

We note that $\mathbf{c}_i^{\perp}$ is now defined as a vector in $\mathcal{R}(\mathbf{A}_i)$ and not in $\mathcal{N}(\mathbf{A}_i)$, as for symmetric matrices.

Also note that for invertible matrices $\mathbf{A}_i,\ 1 \leq i \leq d$, the error component $\mathbf{e}^{\Psi}(t, \mathbf{x})$ is zero.

Substituting $\tilde{\mathbf{e}} = \mathbf{R}\mathbf{e}$, $\tilde{\mathbf{e}}^{\Psi} = \mathbf{R}\mathbf{e}^{\Psi}$, $\tilde{\mathbf{e}}^{\perp} = \mathbf{R}\mathbf{e}^{\perp}$, $\tilde{\mathbf{c}}_i^{\Psi} = \mathbf{R}\mathbf{c}_i^{\Psi}$, $\tilde{\mathbf{c}}_i^{\perp} = \mathbf{R}\mathbf{c}_i^{\perp}$ and $\tilde{\mathbf{d}}_i = \mathbf{R}\mathbf{d}_i$, $1 \leq i \leq d$, into (5.2.11), we obtain the equivalent formulation

$$\mathbf{U} - \mathbf{U}_h = \tilde{\mathbf{e}} = \tilde{\mathbf{e}}^{\perp} + \tilde{\mathbf{e}}^{\Psi} + \mathcal{O}(h^{p+2}), \tag{5.2.12a}$$

where, applying (5.1.2d),

$$\tilde{\mathbf{e}}^{\perp}(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} L_{p+1}(\xi_i)\tilde{\mathbf{c}}_i^{\perp}(t) - L_p(\xi_i)\operatorname{sgn}(\mathbf{B}_i)\tilde{\mathbf{c}}_i^{\perp}(t), \tag{5.2.12b}$$

$$\tilde{\mathbf{e}}^{\Psi}(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} L_{p+1}(\xi_i)\tilde{\mathbf{c}}_i^{\Psi}(t) - L_p(\xi_i)\tilde{\mathbf{d}}_i(t). \tag{5.2.12c}$$

By (5.1.2d) and (1.2.12) we have $\mathcal{R}(\operatorname{sgn}(\mathbf{B}_i)) = \mathcal{R}(\mathbf{B}_i) = \mathcal{N}(\mathbf{B}_i)^{\perp}$, thus

$$\operatorname{sgn}(\mathbf{B}_i)\tilde{\mathbf{c}}_i^{\perp} \in \mathcal{N}(\mathbf{B}_i)^{\perp}, \quad 1 \leq i \leq d. \tag{5.2.12d}$$

Further, (5.2.11d) and Lemma 1.2.16i,ii yields

$$\tilde{\mathbf{c}}_i^{\perp} \in \mathcal{N}(\mathbf{B}_i)^{\perp}, \quad \tilde{\mathbf{c}}_i^{\Psi}, \tilde{\mathbf{d}}_i \in \mathcal{N}(\mathbf{B}_i), \quad 1 \leq i \leq d. \tag{5.2.12e}$$

Next, we develop an *a posteriori* error estimation procedure for estimating both $\mathbf{e}^{\perp}$ and $\mathbf{e}^{\Psi}$ (if needed) and proving that, for smooth solutions, our local error estimates converge to the true error under mesh refinement. Up to this point we are not able to prove the asymptotic exactness of our global *a posteriori* error estimates. However, computational results for several hyperbolic systems shown in § 5.3 suggest that our global *a posteriori* error estimates are asymptotically exact under mesh refinement for smooth solutions.

5.2.1 The Stationary Component of the Error Estimate

The *a posteriori* error estimation procedure to compute estimates for $\mathbf{e}^{\perp}$ consists of determining

$$\mathbf{E}^{\perp}(t, h\boldsymbol{\xi}) = \sum_{j=1}^{d} \left(L_{p+1}(\xi_j)\boldsymbol{\gamma}_j^{\perp}(t) - L_p(\xi_j)\operatorname{sgn}(\mathbf{A}_j)\boldsymbol{\gamma}_j^{\perp}(t) \right), \quad \boldsymbol{\gamma}_j^{\perp} \in \mathcal{R}(\mathbf{A}_j), \tag{5.2.13a}$$

such that

$$\int_\omega \left(L_p\left(\frac{x_i}{h}\right)\mathbf{v}\right)^t \left(\frac{\partial \mathbf{u}_h}{\partial t} + \sum_{j=1}^d \mathbf{A}_j \frac{\partial(\mathbf{u}_h + \mathbf{E}^\perp)}{\partial x_j} - \mathbf{g}\right) d\mathbf{x} = 0,\ \forall\ \mathbf{v} \in \mathcal{N}(\mathbf{A}_i)^\perp,\ 1 \le i \le d. \tag{5.2.13b}$$

By (1.2.43a), (1.2.43b) and (1.2.12), $\mathcal{R}((\mathbf{A}_i^\ddagger)^t) = \mathcal{R}((\mathbf{A}_i^t)^\ddagger) = \mathcal{R}(\mathbf{A}_i^t) = \mathcal{N}(\mathbf{A}_i)^\perp$, thus the columns of $(\mathbf{A}_i^\ddagger)^t$ span $\mathcal{N}(\mathbf{A}_i)^\perp$.

Substituting $\mathbf{v}$ by $(\mathbf{A}_i^\ddagger)^t$ in (5.2.13b) yields

$$\int_\omega L_p\left(\frac{x_i}{h}\right)\mathbf{A}_i^\ddagger \left(\frac{\partial \mathbf{u}_h}{\partial t} + \sum_{j=1}^d \mathbf{A}_j \frac{\partial(\mathbf{u}_h + \mathbf{E}^\perp)}{\partial x_j} - \mathbf{g}\right) d\mathbf{x} = 0, \quad 1 \le i \le d. \tag{5.2.14}$$

Substituting (5.2.13a) into (5.2.14) and applying the orthogonality properties (2.2.6), we obtain

$$\mathbf{A}_i^\ddagger \mathbf{A}_i \boldsymbol{\gamma}_i^\perp \int_\omega L_p\left(\frac{x_i}{h}\right) L'_{p+1}\left(\frac{x_i}{h}\right) d\mathbf{x} = \mathfrak{r}_{p,i}^\perp, \tag{5.2.15a}$$

where $\mathfrak{r}_{p,i}^\perp$ is the residual in $\mathcal{R}(\mathbf{A}_i)$ defined as

$$\mathfrak{r}_{p,i}^\perp = \mathbf{A}_i^\ddagger \int_\omega L_p\left(\frac{x_i}{h}\right)\left(\mathbf{g} - \frac{\partial \mathbf{u}_h}{\partial t} - \sum_{j=1}^d \mathbf{A}_j \frac{\partial \mathbf{u}_h}{\partial x_j}\right) d\mathbf{x}, \quad 1 \le i \le d. \tag{5.2.15b}$$

Since $\boldsymbol{\gamma}_i^\perp \in \mathcal{R}(\mathbf{A}_i)$ and $\mathbf{A}_i^\ddagger \mathbf{A}_i$ is the projection onto $\mathcal{R}(\mathbf{A}_i)$, we further reduce (5.2.15a) to

$$\boldsymbol{\gamma}_i^\perp = \frac{h^{1-d}}{2}\mathfrak{r}_{p,i}^\perp, \quad 1 \le i \le d, \tag{5.2.16}$$

where we used (2.2.6) to show $\int_\omega L_p\left(\frac{x_i}{h}\right) L'_{p+1}\left(\frac{x_i}{h}\right) d\mathbf{x} = \frac{h^{1-d}}{2}$.

Since $\mathfrak{r}_{p,i}^\perp \in \mathcal{R}(\mathbf{A}_i)$, (5.2.16) has a unique solution in $\boldsymbol{\gamma}_i^\perp \in \mathcal{R}(\mathbf{A}_i)$.

Theorem 5.2.2. *Under the assumptions of Theorem 5.1.3, let us consider the error estimate*

$$\mathbf{E}^\perp(t, h\boldsymbol{\xi}) = \sum_{i=1}^d \left(L_{p+1}(\xi_i) - L_p(\xi_i) sgn(\mathbf{A}_i)\right) \frac{h^{1-d}}{2}\mathfrak{r}_{p,i}^\perp, \tag{5.2.17}$$

where $\mathfrak{r}_{p,i}^\perp,\ 1 \le i \le d$, are defined in (5.2.15b). Then, for $p \ge 1$ and $t = \mathcal{O}(1)$,

$$\mathbf{e}^\perp(t, \mathbf{x}) = \mathbf{E}^\perp(t, \mathbf{x}) + \mathcal{O}(h^{p+2}), \quad \mathbf{x} \in \omega. \tag{5.2.18}$$

Proof. Left-multiplying (5.2.13) by $\mathbf{R}$ and substituting $\tilde{\mathbf{E}}^\perp = \mathbf{R}\mathbf{E}^\perp$ and $\tilde{\boldsymbol{\gamma}}_i^\perp = \mathbf{R}\boldsymbol{\gamma}_i^\perp$ yields

$$\tilde{\mathbf{E}}^\perp(t, h\boldsymbol{\xi}) = \sum_{j=1}^d \left(L_{p+1}(\xi_j)\tilde{\boldsymbol{\gamma}}_j^\perp(t) - L_p(\xi_j)\mathrm{sgn}(\mathbf{B}_j)\tilde{\boldsymbol{\gamma}}_j^\perp(t)\right). \tag{5.2.19a}$$

Since $\boldsymbol{\gamma}_i^\perp \in \mathcal{R}(\mathbf{A}_i)$, Lemma 1.2.16ii yields

$$\tilde{\boldsymbol{\gamma}}_i^\perp \in \mathcal{N}(\mathbf{B}_i)^\perp, \quad 1 \leq i \leq d. \tag{5.2.19b}$$

Further, by (5.1.2d) and (1.2.12) we have $\mathcal{R}(\operatorname{sgn}(\mathbf{B}_i)) = \mathcal{R}(\mathbf{B}_i) = \mathcal{N}(\mathbf{B}_i)^\perp$, thus

$$\operatorname{sgn}(\mathbf{B}_i)\tilde{\boldsymbol{\gamma}}_i^\perp \in \mathcal{N}(\mathbf{B}_i)^\perp, \quad 1 \leq i \leq d. \tag{5.2.19c}$$

By Lemma 1.2.16iv, $\mathbf{w} = \mathbf{R}^{-t}\mathbf{v} \in \mathcal{N}(\mathbf{B}_i)^\perp$ for all $\mathbf{v} \in \mathcal{N}(\mathbf{A}_i)^\perp$.
Thus, substituting $\mathbf{v} = \mathbf{R}^t\mathbf{w}$, $\mathbf{u}_h = \mathbf{R}^{-1}\mathbf{U}_h$ and $\mathbf{E}^\perp = \mathbf{R}^{-1}\tilde{\mathbf{E}}^\perp$ into (5.2.13), we obtain

$$\int_\omega \left(L_p\left(\frac{x_i}{h}\right)\mathbf{w}\right)^t \left(\frac{\partial \mathbf{U}_h}{\partial t} + \sum_{j=1}^d \mathbf{B}_j \frac{\partial(\mathbf{U}_h + \tilde{\mathbf{E}}^\perp)}{\partial x_j} - \mathbf{R}\mathbf{g}\right) d\mathbf{x} = 0,$$
$$\forall\ \mathbf{w} \in \mathcal{N}(\mathbf{B}_i)^\perp,\ 1 \leq i \leq d. \tag{5.2.19d}$$

Then, system (5.2.19) satisfies Theorem 4.3.2, which yields, for $p \geq 1$ and $t \in \mathcal{O}(1)$,

$$\tilde{\mathbf{e}}^\perp(t,\mathbf{x}) = \tilde{\mathbf{E}}^\perp(t,\mathbf{x}) + \mathcal{O}(h^{p+2}), \quad \mathbf{x} \in \omega. \tag{5.2.20}$$

Left-multiplying (5.2.20) with $\mathbf{R}^{-1}$ and substituting $\mathbf{e}^\perp = \mathbf{R}^{-1}\tilde{\mathbf{e}}^\perp$ and $\mathbf{E}^\perp = \mathbf{R}^{-1}\tilde{\mathbf{E}}^\perp$, yields (5.2.18). □

5.2.2 The Transient Component of the Error Estimate

We will first present an *a posteriori* error estimation procedure to compute estimates for $\mathbf{e}^{\boldsymbol{\Pi}}$. Then we will show the asymptotic exactness of this error estimate.

By Lemma 2.2.4, the approximations $\pi\mathbf{u}_0$ on ω and $\pi_i\mathbf{u}$ on the boundary $\partial\omega$ satisfy

$$\begin{aligned}\mathbf{e}(0,\mathbf{x}) &= \mathbf{u}_0(\mathbf{x}) - \pi\mathbf{u}_0(\mathbf{x}) \qquad (5.2.21)\\ &= h^{p+1}\sum_{j=1}^d L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j(0) - L_p\left(\frac{x_j}{h}\right)\operatorname{sgn}(\mathbf{A}_j)\mathbf{c}_j(0) + \mathcal{O}(h^{p+2}),\ \mathbf{x} \in \omega,\\ \mathbf{e}^-(t,\mathbf{x}) &= \mathbf{u}(t,\mathbf{x}) - \pi_i\mathbf{u}(t,\mathbf{x}) \qquad (5.2.22)\\ &= h^{p+1}\sum_{j\in D(i)} L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j(t) - L_p\left(\frac{x_j}{h}\right)\operatorname{sgn}(\mathbf{A}_j)\mathbf{c}_j(t) + \mathcal{O}(h^{p+2}),\ \mathbf{x} \in \gamma_i,\ 1 \leq i \leq d.\end{aligned}$$

We split the error at $t = 0$ into $\mathbf{e} = \mathbf{e}^\perp + \mathbf{e}^{\boldsymbol{\Pi}} + \mathcal{O}(h^{p+2})$ as in (5.2.11a) and define $\mathbf{E}^{\boldsymbol{\Pi}}(0,\mathbf{x})$ by

$$\mathbf{E}^{\boldsymbol{\Pi}}(0,\mathbf{x}) = \mathbf{e}^{\boldsymbol{\Pi}}(0,\mathbf{x}) = h^{p+1}\sum_{i=1}^d L_{p+1}\left(\frac{x_i}{h}\right)\mathbf{c}_i^{\boldsymbol{\Pi}}(0), \tag{5.2.23}$$

where $\mathbf{c}_i^{\mathbf{\Phi}}(0)$ is the projection of $\mathbf{c}_i(0)$ into $\mathcal{N}(\mathbf{A}_i)$.

On the boundary, we define $\mathbf{E}^-$ by the leading term of (5.2.22),

$$\mathbf{E}^-(t,\mathbf{x}) = h^{p+1} \sum_{j\in D(i)} L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j(t) - L_p\left(\frac{x_j}{h}\right)\operatorname{sgn}(\mathbf{A}_j)\mathbf{c}_j(t),\ \mathbf{x}\in\gamma_i,\ 1\le i\le d. \quad (5.2.24)$$

Now let us approximate $\mathbf{e}^{\mathbf{\Phi}}$ by determining

$$\mathbf{E}^{\mathbf{\Phi}}(t,h\boldsymbol{\xi}) = \sum_{j=1}^{d} L_{p+1}(\xi_j)\boldsymbol{\gamma}_j^{\mathbf{\Phi}} - L_p(\xi_j)\boldsymbol{\delta}_j^{\mathbf{\Phi}},\quad \boldsymbol{\gamma}_j^{\mathbf{\Phi}},\boldsymbol{\delta}_j^{\mathbf{\Phi}}\in\mathcal{N}(\mathbf{A}_j),\ 1\le j\le d, \quad (5.2.25a)$$

such that

$$\begin{aligned}
&\int_\omega \mathbf{v}^t\left(\frac{\partial(\mathbf{u}_h+\mathbf{E}^{\mathbf{\Phi}})}{\partial t} + \sum_{j=1}^{d}\mathbf{A}_j\frac{\partial\mathbf{u}_h}{\partial x_j} - \mathbf{g}\right)d\mathbf{x}\\
&= \sum_{j=1}^{d}\int_{\gamma_j}\mathbf{v}^t\nu_j\mathbf{A}_j^{\bar\mu_j}\left(\mathbf{u}_h+\mathbf{E}^{\perp}+\mathbf{E}^{\mathbf{\Phi}}-\mathbf{u}_h^-\mathbf{E}^-\right)d\mathbf{s},\\
&\forall\ \mathbf{v}(\mathbf{x}) = \sum_{i=1}^{d}\left(L_{p+1}\left(\frac{x_i}{h}\right)\mathbf{a}_i - L_p\left(\frac{x_i}{h}\right)\mathbf{b}_i\right),\ \mathbf{a}_i,\mathbf{b}_i\in\mathcal{N}(\mathbf{A}_i^t). \quad (5.2.25b)
\end{aligned}$$

By Lemma 1.2.20, $(\mathbf{I}-\mathbf{A}_i\mathbf{A}_i^{\ddagger})^t = (\mathbf{I}-(\mathbf{A}_i^t)^{\ddagger}\mathbf{A}_i^t)$ projects any vector in $\mathbb{R}^m$ into $\mathcal{N}(\mathbf{A}_i^t)$ and the columns of $(\mathbf{I}-\mathbf{A}_i\mathbf{A}_i^{\ddagger})^t$ span $\mathcal{N}(\mathbf{A}_i^t)$. Hence the columns of $L_{p+1}(\xi_i)(\mathbf{I}-\mathsf{P}_i)$ and $L_p(\xi_i)(\mathbf{I}-\mathsf{P}_i),\ 1\le i\le d$, span the space of test functions.

Replacing $\mathbf{v}$ in (5.2.25b) by $L_m(\xi_i)(\mathbf{I}-\mathbf{A}_i\mathbf{A}_i^{\ddagger})^t,\ m=p,p+1,\ 1\le i\le d$, yields

$$\begin{aligned}
&\int_\omega L_m\left(\frac{x_i}{h}\right)(\mathbf{I}-\mathbf{A}_i\mathbf{A}_i^{\ddagger})\left(\frac{\partial(\mathbf{u}_h+\mathbf{E}^{\mathbf{\Phi}})}{\partial t} + \sum_{j=1}^{d}\mathbf{A}_j\frac{\partial\mathbf{u}_h}{\partial x_j} - \mathbf{g}\right)d\mathbf{x}\\
&= \sum_{j=1}^{d}\int_{\gamma_j} L_m\left(\frac{x_i}{h}\right)(\mathbf{I}-\mathbf{A}_i\mathbf{A}_i^{\ddagger})\nu_j\mathbf{A}_j^{\bar\mu_j}\left(\mathbf{u}_h+\mathbf{E}^{\perp}+\mathbf{E}^{\mathbf{\Phi}}-\mathbf{u}_h^-\mathbf{E}^-\right)d\mathbf{s},\\
&\forall\ m=p,p+1,\ 1\le i\le d. \quad (5.2.26)
\end{aligned}$$

By Lemma 1.2.20i and (1.2.25b), $(\mathbf{I}-\mathbf{A}_i\mathbf{A}_i^{\ddagger})\mathbf{A}_i^{\bar\mu_i}=\mathbf{0}$, thus (5.2.26) can be written as

$$\int_\omega L_m\left(\frac{x_i}{h}\right)(\mathbf{I}-\mathbf{A}_i\mathbf{A}_i^{\ddagger})\frac{\partial\mathbf{E}^{\mathbf{\Phi}}}{\partial t}\,d\mathbf{x} - \sum_{j\in D(i)}\int_{\gamma_j} L_m\left(\frac{x_i}{h}\right)(\mathbf{I}-\mathbf{A}_i\mathbf{A}_i^{\ddagger})\nu_j\mathbf{A}_j^{\bar\mu_j}\mathbf{E}^{\mathbf{\Phi}}\,d\mathbf{s} = \mathfrak{r}_{m,i}^{\mathbf{\Phi}}, \quad (5.2.27a)$$

where $\mathfrak{r}^{\Phi}_{m,i}$ is the projection of the residual given by

$$\mathfrak{r}^{\Phi}_{m,i} = (\mathbf{I} - \mathbf{A}_i\mathbf{A}_i^{\ddagger}) \int_{\omega} L_m\left(\frac{x_i}{h}\right)\left(\mathbf{g} - \frac{\partial \mathbf{u}_h}{\partial t} - \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{u}_h}{\partial x_j}\right) d\mathbf{x}$$
$$+ (\mathbf{I} - \mathbf{A}_i\mathbf{A}_i^{\ddagger}) \sum_{j=1}^{d} \int_{\gamma_j} L_m\left(\frac{x_i}{h}\right) \nu_j \mathbf{A}_j^{\bar{\mu}_j} \left(\mathbf{u}_h + \mathbf{E}^{\perp} - \mathbf{u}_h^{-} - \mathbf{E}^{-}\right)\big) \, d\mathbf{s},$$
$$m = p, p+1,\ 1 \le i \le d. \tag{5.2.27b}$$

For $m = p+1$, we use the orthogonality properties (2.2.6) to reduce (5.2.27a) to

$$\int_{\omega} L_{p+1}^2\left(\frac{x_i}{h}\right) \dot{\boldsymbol{\gamma}}_i^{\Phi} \, d\mathbf{x} - \sum_{j \in D(i)} \int_{\gamma_j} L_{p+1}^2\left(\frac{x_i}{h}\right) \nu_j (\mathbf{I} - \mathbf{A}_i\mathbf{A}_i^{\ddagger}) \mathbf{A}_j^{\bar{\mu}_j} \boldsymbol{\gamma}_i^{\Phi} \, d\mathbf{s} = \mathfrak{r}^{\Phi}_{p+1,i}, \tag{5.2.28}$$

which by (2.2.6) is equal to

$$\dot{\boldsymbol{\gamma}}_i^{\Phi} = \frac{1}{h}(\mathbf{I} - \mathbf{A}_i\mathbf{A}_i^{\ddagger}) \sum_{j \in D(i)} (\mathbf{A}_j^{-} - \mathbf{A}_j^{+}) \boldsymbol{\gamma}_i^{\Phi} + \frac{2p+3}{h^d} \mathfrak{r}^{\Phi}_{p+1,i}. \tag{5.2.29a}$$

For $m = p$, we get similarly

$$\dot{\boldsymbol{\delta}}_i^{\Phi} = \frac{1}{h}(\mathbf{I} - \mathbf{A}_i\mathbf{A}_i^{\ddagger}) \sum_{j \in D(i)} (\mathbf{A}_j^{-} - \mathbf{A}_j^{+}) \boldsymbol{\delta}_i^{\Phi} + \frac{2p+1}{h^d} \mathfrak{r}^{\Phi}_{p,i}, \tag{5.2.29b}$$

subject to the initial conditions

$$\boldsymbol{\gamma}_i^{\Phi}(0) = h^{p+1} \mathbf{c}_i^{\Phi}(0),\ \boldsymbol{\delta}_i^{\Phi}(0) = \mathbf{0},\ 1 \le i \le d. \tag{5.2.29c}$$

Note that (5.2.29) and (5.2.27b) ensures that $\boldsymbol{\gamma}_i^{\Phi}, \boldsymbol{\delta}_i^{\Phi} \in \mathcal{N}(\mathbf{A}_i),\ 1 \le i \le d$.

Theorem 5.2.3. *Under the assumptions of Theorem 5.1.3, assume further that $\mathbf{u}_h$ is computed by approximating the initial conditions by $\pi\mathbf{u}_0$ and let*

$$\mathbf{E}^{\Phi}(t, h\boldsymbol{\xi}) = \sum_{j=1}^{d} (L_{p+1}(\xi_j)\boldsymbol{\gamma}_j^{\Phi}(t) - L_p(\xi_j)\boldsymbol{\delta}_j^{\Phi}(t)), \tag{5.2.30}$$

where $\boldsymbol{\gamma}_i^{\Phi}, \boldsymbol{\delta}_i^{\Phi},\ 1 \le i \le d$, are solutions of (5.2.29) and (5.2.27b).
Then, at $t = \mathcal{O}(1)$ and for $p \ge 1$,

$$\mathbf{e}^{\Phi}(t, \mathbf{x}) = \mathbf{E}^{\Phi}(t, \mathbf{x}) + \mathcal{O}(h^{p+2}), \quad \mathbf{x} \in \omega. \tag{5.2.31}$$

Proof. Left-multiplying (5.2.25a) by $\mathbf{R}$ and substituting $\tilde{\mathbf{E}}^{\Phi} = \mathbf{R}\mathbf{E}^{\perp}$, $\tilde{\boldsymbol{\gamma}}_i^{\Phi} = \mathbf{R}\boldsymbol{\gamma}_i^{\Phi}$ and $\tilde{\boldsymbol{\delta}}_i^{\Phi} = \mathbf{R}\boldsymbol{\delta}_i^{\Phi}$ yields

$$\tilde{\mathbf{E}}^{\Phi}(t, h\boldsymbol{\xi}) = \sum_{j=1}^{d} L_{p+1}(\xi_j)\tilde{\boldsymbol{\gamma}}_j^{\Phi} - L_p(\xi_j)\tilde{\boldsymbol{\delta}}_j^{\Phi}. \tag{5.2.32a}$$

Since $\boldsymbol{\gamma}_i^{\boldsymbol{\Phi}}, \boldsymbol{\delta}_i^{\boldsymbol{\Phi}} \in \mathcal{N}(\mathbf{A}_i)$, Lemma 1.2.16i yields

$$\tilde{\boldsymbol{\gamma}}_i^{\boldsymbol{\Phi}}, \tilde{\boldsymbol{\delta}}_i^{\boldsymbol{\Phi}} \in \mathcal{N}(\mathbf{B}_i), \quad 1 \leq i \leq d. \tag{5.2.32b}$$

By Lemma 1.2.16iv, $\mathbf{w} = \mathbf{R}^{-t}\mathbf{v} \in \mathcal{N}(\mathbf{B}_i)$ for all $\mathbf{v} \in \mathcal{N}(\mathbf{A}_i^t)$.
Thus, substituting $\mathbf{v} = \mathbf{R}^t\mathbf{w}$, $\mathbf{u}_h = \mathbf{R}^{-1}\mathbf{U}_h$, $\mathbf{E}^{\boldsymbol{\Phi}} = \mathbf{R}^{-1}\tilde{\mathbf{E}}^{\boldsymbol{\Phi}}$, $\mathbf{E}^{\perp} = \mathbf{R}^{-1}\tilde{\mathbf{E}}^{\perp}$ and $\mathbf{E}^{-} = \mathbf{R}^{-1}\tilde{\mathbf{E}}^{-}$ into (5.2.25b), we obtain

$$\begin{aligned}
&\int_\omega \mathbf{w}^t \left(\frac{\partial(\mathbf{U}_h + \tilde{\mathbf{E}}^{\boldsymbol{\Phi}})}{\partial t} + \sum_{j=1}^{d} \mathbf{B}_j \frac{\partial \mathbf{U}_h}{\partial x_j} - \mathbf{R}\mathbf{g} \right) d\mathbf{x} \\
&= \sum_{j=1}^{d} \int_{\gamma_j} \mathbf{w}^t \nu_j \mathbf{B}_j^{\bar{\mu}_j} \left(\mathbf{U}_h + \tilde{\mathbf{E}}^{\perp} + \tilde{\mathbf{E}}^{\boldsymbol{\Phi}} - \mathbf{U}_h^{-} - \tilde{\mathbf{E}}^{-}) \right) d\mathbf{s}, \\
&\forall\ \mathbf{w}(\mathbf{x}) = \sum_{i=1}^{d} \left(L_{p+1}\left(\frac{x_i}{h}\right) \mathbf{a}_i - L_p\left(\frac{x_i}{h}\right) \mathbf{b}_i \right),\ \mathbf{a}_i, \mathbf{b}_i \in \mathcal{N}(\mathbf{B}_i).
\end{aligned} \tag{5.2.32c}$$

Then, system (5.2.32) satisfies Theorem 4.3.4, which yields, for $p \geq 1$ and $t \in \mathcal{O}(1)$,

$$\tilde{\mathbf{e}}^{\boldsymbol{\Phi}}(t, \mathbf{x}) = \tilde{\mathbf{E}}^{\boldsymbol{\Phi}}(t, \mathbf{x}) + \mathcal{O}(h^{p+2}), \quad \mathbf{x} \in \omega. \tag{5.2.33}$$

Left-multiplying (5.2.33) with $\mathbf{R}^{-1}$ and substituting $\mathbf{e}^{\boldsymbol{\Phi}} = \mathbf{R}^{-1}\tilde{\mathbf{e}}^{\boldsymbol{\Phi}}$ and $\mathbf{E}^{\boldsymbol{\Phi}} = \mathbf{R}^{-1}\tilde{\mathbf{E}}^{\boldsymbol{\Phi}}$, yields (5.2.31). □

5.3 Computational Examples

We present two examples of symmetrizable systems to validate the theoretical results of Chapter 5.

Example 5.3.1. *Let* $\mathbf{u}$ *be defined on* $\mathbf{x} \in \Omega = (0,1)^2$ *and* $0 \leq t \leq 1$ *by*

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}_1 \frac{\partial \mathbf{u}}{\partial x_1} + \mathbf{A}_2 \frac{\partial \mathbf{u}}{\partial x_2} = \mathbf{g}, \quad \mathbf{A}_1 = \begin{pmatrix} 3 & 0 \\ 1 & 0 \end{pmatrix},\ \mathbf{A}_2 = \begin{pmatrix} 2 & 0 \\ -2 & 8 \end{pmatrix}, \tag{5.3.1a}$$

with source term $\mathbf{g}$*, initial and boundary conditions such that*

$$\mathbf{u}(t, \mathbf{x}) = (1,1)^t \exp(t + x_1 + x_2), \quad \mathbf{x} \in \Omega,\ 0 \leq t \leq 1. \tag{5.3.1b}$$

We can symmetrize (5.3.1a) by left-multiplying the system by a symmetric positive definite matrix $\mathbf{S}_0 = \frac{1}{8}\begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$, which yields

$$\mathbf{S}_0 \frac{\partial \mathbf{u}}{\partial t} + \mathbf{S}_1 \frac{\partial \mathbf{u}}{\partial x_1} + \mathbf{S}_2 \frac{\partial \mathbf{u}}{\partial x_2} = \mathbf{S}_0 \mathbf{g}, \quad \mathbf{S}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},\ \mathbf{S}_2 = \begin{pmatrix} 1 & -1 \\ -1 & 3 \end{pmatrix}. \tag{5.3.2}$$

Since $\mathbf{A}_1$ has a zero eigenvalue for eigenvector $\boldsymbol{\varepsilon}_2 = (0,1)^t$, $\mathbf{e}^{\mathbf{\Psi}}$ is nonzero with

$$\tilde{\mathbf{e}}^{\mathbf{\Psi}}(t, h\boldsymbol{\xi}) = \left(h^{p+1} L_{p+1}(\xi_1) c(t) - L_p(\xi_1) d(t)\right) \boldsymbol{\varepsilon}_2. \tag{5.3.3}$$

Thus, we predict that the stationary error estimate $\mathbf{E}^{\perp}$ can only accurately estimate the first component of the error e_1.

In order to validate the theoretical results, we solve (5.3.1) on uniform meshes having $N = 5^2, 10^2, 15^2$ elements with $p = 1, 2, 3$ for $0 \leq t \leq 1$. We present the L_2-errors and effectivity indices for each component of the stationary error estimate of the solution at $t = 1$ in Table 5.3.1. We observe that only $E_1^{\perp}$ is an accurate estimate of e_1. We present the L_2-errors and effectivity indices for the transient error estimate of the solution at $t = 1$ in Table 5.3.2. We observe that the effectivity indices converge to unity under mesh refinement.

Example 5.3.2. *Let us consider Maxwell's equations*

$$\varepsilon_0 \frac{\partial \boldsymbol{\mathcal{E}}}{\partial t} = \nabla \times \boldsymbol{\mathcal{H}}, \quad \nabla \cdot \boldsymbol{\mathcal{E}} = 0, \tag{5.3.4a}$$

$$\mu_0 \frac{\partial \boldsymbol{\mathcal{H}}}{\partial t} = \nabla \times \boldsymbol{\mathcal{E}}, \quad \nabla \cdot \boldsymbol{\mathcal{H}} = 0, \tag{5.3.4b}$$

where $\boldsymbol{\mathcal{E}}(t, \mathbf{x}) = (\mathcal{E}_x, \mathcal{E}_y, \mathcal{E}_z)^t$ *and* $\boldsymbol{\mathcal{H}}(t, \mathbf{x}) = (\mathcal{H}_x, \mathcal{H}_y, \mathcal{H}_z)^t$ *denote the* electric and magnetic field *and* $\mu_0 = 4\pi \cdot 10^{-7}\, NA^{-2}$, $\varepsilon_0 = c^{-2}\mu_0^{-1}$ *denote the* magnetic and electric permittivity *in free space, respectively, with* $c = 299,792,458\, ms^{-2}$ *being the* speed of light.

For a transverse electric wave traveling in the x_1x_2*-plane,* $\mathcal{E}_z = \mathcal{H}_x = \mathcal{H}_y = 0$. *Thus, (5.3.4) yields the symmetrizable hyperbolic system*

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}_1 \frac{\partial \mathbf{u}}{\partial x_1} + \mathbf{A}_2 \frac{\partial \mathbf{u}}{\partial x_2} = \mathbf{0}, \tag{5.3.5a}$$

where

$$\mathbf{u} = \begin{pmatrix} \mathcal{E}_x \\ \mathcal{E}_y \\ \mathcal{H}_z \end{pmatrix}, \quad \mathbf{A}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \varepsilon_0^{-1} \\ 0 & \mu_0^{-1} & 0 \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} 0 & 0 & -\varepsilon_0^{-1} \\ 0 & 0 & 0 \\ -\mu_0^{-1} & 0 & 0 \end{pmatrix}. \tag{5.3.5b}$$

We will select initial and boundary conditions such that the true solution is a typical FM radio wave of wavelength $\lambda = 3m$ *and amplitude* $A = 1.5V/m$ *traveling in direction* $[x, y] = [1, 1]$. *i.e.*

$$\mathbf{u}(t, \mathbf{x}) = \left[1, -1, \sqrt{2\mu_0\varepsilon_0^{-1}}\right]^t A \sin\left(\frac{2\pi}{\lambda}(ct + \frac{x+y}{\sqrt{2}})\right), \ \forall\ t \geq 0, \ \mathbf{x} \in \Omega. \tag{5.3.5c}$$

We solve (5.3.5) for $0 < t \leq 2 \cdot 10^{-8}$, a time interval, in which the wave travels $10^{-8} \cdot c \approx 6\, m$.

Both matrices $\mathbf{A}_1, \mathbf{A}_2$ are singular and admit the eigenvalues $\{-1, 0, 1\}$. Moreover, the eigenvectors $(1,0,0)^t$ and $(0,1,0)^t$ are associated with the zero eigenvalue for $\mathbf{A}_1$ and $\mathbf{A}_2$,

p	N	$\|\mathbf{e}\|^*$	*order*	$\|\mathbf{e}-\mathbf{E}^\perp\|^*$	*order*	θ^*
1	5^2	2.9339*e*-2	–	3.1898*e*-3	–	0.9308
		2.5247*e*-2		9.0253*e*-3		0.8395
	10^2	7.4038*e*-3	1.9865	4.0646*e*-4	2.9723	0.9658
		6.3200*e*-3	1.9981	2.1982*e*-3	2.0377	0.8566
	15^2	3.3009*e*-3	1.9923	1.2111*e*-4	2.9861	0.9773
		2.8183*e*-3	1.9917	9.6980*e*-4	2.0182	0.8648
2	5^2	4.7562*e*-4	–	5.0290*e*-5	–	0.9587
		4.1291*e*-4		1.5255*e*-4		0.8423
	10^2	5.9669*e*-5	2.9947	3.2093*e*-6	3.9700	0.9801
		5.1565*e*-5	3.0014	1.8501*e*-5	3.0436	0.8544
	15^2	1.7704*e*-5	2.9967	6.3810*e*-7	3.9838	0.9869
		1.5279*e*-5	2.9999	5.4450*e*-6	3.0166	0.8585
3	5^2	5.9131*e*-6	–	8.8854*e*-7	–	0.9512
		5.1399*e*-6		2.0124*e*-6		0.8295
	10^2	3.6824*e*-7	4.0052	2.8857*e*-8	4.9445	0.9775
		3.1954*e*-7	4.0077	1.1811*e*-7	4.0908	0.8475
	15^2	7.2710*e*-8	4.0010	3.8566*e*-9	4.9636	0.9855
		6.3063*e*-8	4.0022	2.3011*e*-8	4.0339	0.8526

Table 5.3.1: Componentwise $L^2(\Omega)$-Norm of error and static error estimate and global effectivity index for Example 5.3.1 at $t = 1$ for $p = 1, 2, 3$ and $n = 5, 10, 15$.

p	N	$\|\mathbf{e}\|$	*order*	$\|\mathbf{e}-\mathbf{E}^\perp-\mathbf{E}^\Pi\|$	*order*	θ
1	5^2	3.8706*e*-2	–	3.7645*e*-3	–	0.9529
	10^2	9.7344*e*-3	1.991	4.8038*e*-4	2.97	0.976
	15^2	4.3404*e*-3	1.992	1.4230*e*-4	3.001	0.9844
2	5^2	6.2985*e*-4	–	6.6223*e*-5	–	0.9746
	10^2	7.8863*e*-5	2.998	4.1988*e*-6	3.979	0.9878
	15^2	2.3385*e*-5	2.998	8.3283*e*-7	3.99	0.9921
3	5^2	7.8348*e*-6	–	1.1977*e*-6	–	0.9672
	10^2	4.8755*e*-7	4.006	3.8324*e*-8	4.966	0.9857
	15^2	9.6248*e*-8	4.001	5.0915*e*-9	4.978	0.9911

Table 5.3.2: $L^2(\Omega)$-Norm of error and transient error estimate and global effectivity index for Example 5.3.1 at $t = 1$ for $p = 1, 2, 3$ and $n = 5, 10, 15$.

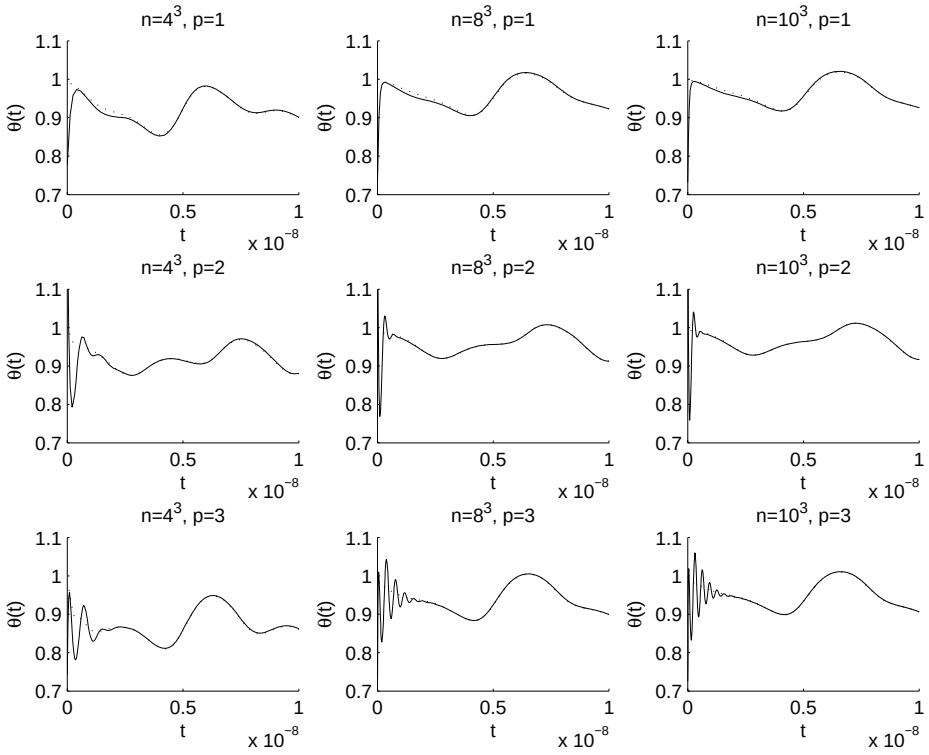

Figure 5.3.1: Global effectivity indices versus time using $\pi\mathbf{u}_0$ (dotted) and $\Pi\mathbf{u}_0$ (solid) for Example 5.3.2.

respectively. Applying our theory, the stationary error estimate $\mathbf{E}^\perp$ can only accurately approximate the component of the error lying in $(\mathcal{N}(\mathbf{A}_1) \bigoplus \mathcal{N}(A_2))^\perp = \text{span}\{(0,0,1)^t\}$, *i.e.,* only $E_3^\perp$ is an accurate estimate of e_3.

We further note, that (5.3.5) does not satisfy the conditions of Lemma 3.1.1, since both matrices are singular and $\mathbf{P}_{1,2}^t\mathbf{A}_2\mathbf{P}_{1,2} = \mathbf{P}_{2,2}^t\mathbf{A}_1\mathbf{P}_{2,2} = 0$.

To validate our theory, we solve (5.3.5) on uniform meshes having $N = 10^2, 20^2, 30^2$ elements for $p = 1, 2, 3$ using $\Pi\mathbf{u}_0$. We present the componentwise L^2-errors and effectivity indices corresponding to the stationary error estimate $\mathbf{E}^\perp$ at $t = 2 \cdot 10^{-8}$ in Table 5.3.3. In Table 5.3.4, we present the L^2-errors and effectivity indices for the transient error estimate $\mathbf{E}^\perp + \mathbf{E}^{\Cross}$ at $t = 2 \cdot 10^{-8}$. We observe that the effectivity indices for the transient error estimate and for the third component of the static estimate converge to unity under mesh refinement. Furthermore, we plot the effectivity indices for the transient error estimate versus time in Figure 5.3.1 to show that $\mathbf{E}^\perp + \mathbf{E}^{\Cross}$ is asymptotically accurate, which is in full agreement with Theorem 5.2.2. We further note that the effectivity indices stay close to unity at all times when using $\pi\mathbf{u}_0$. For $\Pi\mathbf{u}_0$, the effectivity indices oscillates about unity near $t = 0$ before approaching unity.

p	N	$\|\mathbf{e}\|^*$	*order*	$\|\mathbf{e}-\mathbf{E}^\perp\|^*$	*order*	θ^*
1	4^2	3.1090e-2	–	1.8772e-2	–	0.7828
		3.1090e-2		1.8772e-2		0.7828
		8.7984e-5		2.6422e-5		0.9584
	8^2	7.9312e-3	1.9709	4.9144e-3	1.9335	0.7807
		7.9312e-3	1.9709	4.9144e-3	1.9335	0.7807
		2.1622e-5	2.0248	3.3561e-6	2.9769	0.9890
	10^2	5.1169e-3	1.9640	3.2097e-3	1.9091	0.7759
		5.1169e-3	1.9640	3.2097e-3	1.9091	0.7759
		1.3806e-5	2.0103	1.7204e-6	2.9946	0.9929
2	4^2	1.1176e-3	–	7.0864e-4	–	0.7739
		1.1176e-3		7.0864e-4		0.7739
		3.1994e-6		1.0324e-6		0.9617
	8^2	1.4231e-4	2.9732	8.9626e-5	2.9831	0.7761
		1.4231e-4	2.9732	8.9626e-5	2.9831	0.7761
		3.9577e-7	3.0151	6.4991e-8	3.9896	0.9908
	10^2	7.3477e-5	2.9625	4.6663e-5	2.9251	0.7718
		7.3477e-5	2.9625	4.6663e-5	2.9251	0.7718
		2.0247e-7	3.0036	2.6651e-8	3.9949	0.9943
3	4^2	4.5022e-5	–	3.0142e-5	–	0.7462
		4.5022e-5		3.0142e-5		0.7462
		1.2583e-7		3.7623e-8		0.9622
	8^2	2.8075e-6	4.0033	1.8538e-6	4.0232	0.7517
		2.8075e-6	4.0033	1.8538e-6	4.0232	0.7517
		7.6762e-9	4.0349	1.1943e-9	4.9774	0.9898
	10^2	1.1543e-6	3.9832	7.6474e-7	3.9682	0.7494
		1.1543e-6	3.9832	7.6474e-7	3.9682	0.7494
		3.1348e-9	4.0133	3.9216e-10	4.9907	0.9934

Table 5.3.3: Componentwise L^2 errors $\|\mathbf{e}\|^*$, $\|\mathbf{e}-\mathbf{E}^\perp\|^*$ and their order of convergence. Global effectivity indices θ^* for each component for Example 5.3.2 at $t = 10^{-8}$ using $\Pi\mathbf{u}_0$.

p	N	$\|\mathbf{e}\|$	*order*	$\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Phi}\|$	*order*	θ
1	4^2	4.3969*e*-2	–	1.2540*e*-2	–	0.901
	8^2	1.1216*e*-2	1.971	2.3479*e*-3	2.417	0.923
	10^2	7.2364*e*-3	1.964	1.3870*e*-3	2.359	0.9261
2	4^2	1.5805*e*-3	–	6.0655*e*-4	–	0.8812
	8^2	2.0126*e*-4	2.973	5.1210*e*-5	3.566	0.9127
	10^2	1.0391*e*-4	2.962	2.3919*e*-5	3.412	0.917
3	4^2	6.3671*e*-5	–	2.6401*e*-5	–	0.8621
	8^2	3.9704*e*-6	4.003	1.0874*e*-6	4.602	0.8995
	10^2	1.6324*e*-6	3.983	3.9561*e*-7	4.531	0.9064

Table 5.3.4: L^2 errors $\|\mathbf{e}\|$, $\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Phi}\|$, their order of convergence and global effectivity indices for Example 5.3.2 at $t = 2 \cdot 10^{-8}$ using $\Pi\mathbf{u}_0$.

Chapter 6

The DG Method with Lax-Friedrichs Flux

In this chapter, we will consider the effect of substituting the Steger-Warming numerical flux by the *Lax-Friedrichs* flux for solving linear symmetric hyperbolic systems. The Lax-Friedrichs flux is defined in [31] for nonlinear systems as replacing $\sum_{i=1}^{d} \nu_i \mathbf{f}_i(\mathbf{u})$ in the boundary integral term of the weak formulation (2.1.4) by

$$\mathbf{h}(\mathbf{u}_h^+, \mathbf{u}_h^-, \boldsymbol{\nu}) = \sum_{i=1}^{d} \frac{\nu_i}{2} \left(\mathbf{f}_i(\mathbf{u}_h^+) + \mathbf{f}_i(\mathbf{u}_h^-) \right) - \frac{C_i}{2} \left(\mathbf{u}_h^+ - \mathbf{u}_h^- \right), \\ 0 \leq t \leq T,\ \mathbf{x} \in \partial\omega_h,\ \omega_h \in \mathcal{T}_h, \tag{6.0.1}$$

where $\boldsymbol{\nu}$ denotes the outward unit normal, and C_i denotes the largest absolute eigenvalue of the Jacobian $\frac{\partial}{\partial \mathbf{u}} \nu_i \mathbf{f}_i(\mathbf{u}_h)$ in a neighborhood of each edge of ω_h.

This flux splitting is of interest in particular for non-linear systems, since the Steger-Warming flux splitting requires the flux matrices to be diagonalized, while the Lax-Friedrichs flux only requires the modulus of the largest eigenvalue, often called wave propagation speed, to be known, thus minimizing computational cost.

For linear systems, the Lax-Friedrichs flux yields

$$\begin{aligned} \mathbf{h}(\mathbf{u}_h^+, \mathbf{u}_h^-, \boldsymbol{\nu}) &= \sum_{i=1}^{d} \frac{\nu_i}{2} \left(\mathbf{A}_i \mathbf{u}_h^+ + \mathbf{A}_i \mathbf{u}_h^- \right) - \frac{C_i}{2} \left(\mathbf{u}_h^+ - \mathbf{u}_h^- \right) \\ &= \sum_{i=1}^{d} \nu_i \breve{\mathbf{A}}_i^{\mu_i} \mathbf{u}_h^+ + \nu_i \breve{\mathbf{A}}_i^{\bar{\mu}_i} \mathbf{u}_h^-,\ 0 \leq t \leq T,\ \mathbf{x} \in \partial\omega. \end{aligned} \tag{6.0.2}$$

where $\breve{\mathbf{A}}_i^{+}$, $\breve{\mathbf{A}}_i^{-}$, and C_i, $1 \leq i \leq d$, are defined in Definition 1.2.21.

Thus, let $\mathbf{u} \in [C^2([0,T], C^{p+2}(\bar{\Omega}))]^m$ be the true solution of the linear symmetric hyperbolic system

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{i=1}^{d} \mathbf{A}_i \frac{\partial \mathbf{u}}{\partial x_i} = \mathbf{g}(t,\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad 0 < t < T, \tag{6.0.3a}$$

with *source term* $\mathbf{g} : (0,T) \times \Omega \to \mathbb{R}^m$ and subject to the initial and boundary conditions

$$\mathbf{u}(0,\mathbf{x}) = \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{6.0.3b}$$

$$\left(\sum_{i=1}^{d} \nu_i \breve{\mathbf{A}}_i^{\bar{\mu}_i} \right) \mathbf{u}(t,\mathbf{x}) = \left(\sum_{i=1}^{d} \nu_i \breve{\mathbf{A}}_i^{\bar{\mu}_i} \right) \mathbf{u}_B(t,\mathbf{x}), \quad \mathbf{x} \in \partial\Omega,\ 0 < t < T. \tag{6.0.3c}$$

First, we define a proper DG formulation with the Lax-Friedrichs numerical flux. Then we perform a local error analysis to show that, if the order of approximation p is odd or all matrices $\mathbf{A}_1, \ldots, \mathbf{A}_d$ are invertible, the leading term of the discretization error can be expressed as a linear combination of Legendre polynomials of degree p and $p+1$. We note that the superconvergence results of Theorem 4.3.1 for the Steger-Warming numerical flux do *not* extend to the Lax-Friedrichs numerical flux. However, we are able to extend the *a posteriori* error estimation procedure to obtain efficient and asymptotically exact estimates of the discretization error, under the same assumptions as mentioned above. We conclude the chapter by presenting computational results systems with singular and invertible matrices.

6.1 DG formulation

The DG method for the Lax-Friedrichs flux consists of finding $\mathbf{u}_h \in \mathcal{P}_p$ that satisfies

$$\int_\omega \mathbf{v}^t \left(\frac{\partial \mathbf{u}_h}{\partial t} - \mathbf{g} \right) d\mathbf{x} = \sum_{j=1}^{d} \left(\int_\omega \frac{\partial \mathbf{v}^t}{\partial x_j} \mathbf{A}_j \mathbf{u}_h \, d\mathbf{x} - \int_{\partial\omega} \mathbf{v}^t \nu_j (\breve{\mathbf{A}}_j^{\mu_j} \mathbf{u}_h^+ + \breve{\mathbf{A}}_j^{\bar{\mu}_j} \mathbf{u}_h^-) \, d\mathbf{s} \right), \tag{6.1.1a}$$

$$\forall\ \mathbf{v} \in \mathcal{P}_p,\ 0 < t < T,$$

subject to the initial and boundary conditions

$$\mathbf{u}_h(0,\mathbf{x}) = \breve{\pi}\mathbf{u}_0(\mathbf{x}) \text{ or } \mathbf{u}_h(0,\mathbf{x}) = \Pi\mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \omega, \tag{6.1.1b}$$

$$(\nu_i \breve{\mathbf{A}}_i^{\bar{\mu}_i}) \mathbf{u}_h^-(t,\mathbf{x}) = (\nu_i \breve{\mathbf{A}}_i^{\bar{\mu}_i}) \breve{\pi}_i^s \mathbf{u}(t,\mathbf{x}),\ \mathbf{x} \in \gamma_i^s,\ s = +,-,\ 1 \le i \le d,\ 0 < t < T, \tag{6.1.1c}$$

where the initial and boundary approximation operators $\breve{\pi}$, $\breve{\pi}_i^+$ and $\breve{\pi}_i^-$, $1 \le i \le d$, are defined below.

Definition 6.1.1. *If $\mathcal{L}_{p+1}\mathbf{v}$ is the $L^2(\omega)$-projection of $\mathbf{v}$ onto $\mathcal{P}_p$, we define a* corrected $L^2(\omega)$-projection *onto $\mathcal{P}_p$ as*

$$\breve{\pi}\mathbf{v}(\mathbf{x}) = \mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) - \sum_{i=1}^{d} \left(L_{p+1}\left(\frac{x_i}{h}\right) \bar{\mathbf{c}}_i - L_p\left(\frac{x_i}{h}\right) \mathbf{A}_i^{(mod(p,2))} \bar{\mathbf{c}}_i \right), \tag{6.1.2a}$$

where the coefficients $\bar{\mathbf{c}}_i$ *are defined in (2.2.15d) as*

$$\bar{\mathbf{c}}_i = \frac{\int_\Delta \mathbf{v}(\boldsymbol{\xi}) L_{p+1}(\xi_i)\, d\boldsymbol{\xi}}{\int_\Delta L_{p+1}^2(\xi_i)\, d\boldsymbol{\xi}}, \quad 1 \le i \le d. \tag{6.1.2b}$$

Similarly, if $\mathcal{L}_{p+1}^{i,s}\mathbf{v}$ *is the* $L^2(\gamma_i^s)$*-projection of* $\mathbf{v}$ *onto* $\mathcal{P}_p$*, we define a* corrected $L^2(\gamma_i^s)$-projection *as*

$$\begin{aligned}\breve{\pi}_i^s\mathbf{v}(\mathbf{x}) = \mathcal{L}_{p+1}^{i,s}\mathbf{v}(\mathbf{x}) - \sum_{j\in D(i)} \left(L_{p+1}\left(\frac{x_j}{h}\right)\bar{\mathbf{c}}_{ij}^s - L_p\left(\frac{x_j}{h}\right)(\mathbf{A}_j^{(mod(p,2))}\bar{\mathbf{c}}_{ij}^s + \mathbf{d}_j)\right) \qquad (6.1.3a)\\ - \left(L_{p+1}\left(1-\frac{x_i}{h}\right)\mathbf{c}_i - L_p\left(1-\frac{x_i}{h}\right)(\mathbf{A}_i^{(mod(p,2))}\mathbf{c}_i - \mathbf{d}_i)\right),\ s=+,-,\ 1\le i\le d,\end{aligned}$$

where the coefficients $\bar{\mathbf{c}}_{ij}^s$ *are defined in (2.2.16c) as*

$$\bar{\mathbf{c}}_{ij}^s = \frac{\int_{\Gamma_i^s} \mathbf{v}(h\boldsymbol{\xi}) L_{p+1}(\xi_j)\, d\boldsymbol{\sigma}}{\int_{\Gamma_i^s} L_{p+1}^2(\xi_j)\, d\boldsymbol{\sigma}}, \quad s=+,-,\ j\in D(i),\ 1\le i\le d, \tag{6.1.3b}$$

the coefficients $\mathbf{c}_i$ *are defined in (2.2.10c) as*

$$\mathbf{c}_i = \frac{1}{a_{p+1}}\frac{1}{(p+1)!}\frac{\partial^{p+1}\mathbf{v}(\mathbf{0})}{\partial x_i^{p+1}},\ 1\le i\le d, \tag{6.1.3c}$$

and the coefficients $\mathbf{d}_i$ *are chosen such that*

$$\mathbf{d}_i = \mathcal{O}(h^{p+1}),\ \mathbf{d}_i \in \mathcal{N}(\mathbf{A}_i),\ 1\le i\le d. \tag{6.1.3d}$$

Note. *To obtain the boundary interpolation operator* $\breve{\pi}_i^s\mathbf{u}$ *on* γ_i^s*,* $s=+,-$*,* $1\le i\le d$*, we need the coefficients* $\mathbf{c}_{ij}^s$*,* $j\in D(i)$*, which can be obtained by evaluating* $\mathbf{u}$ *on* γ_i^s*, and the coefficients* $\mathbf{c}_i$*,* $\mathbf{d}_i$*, which* cannot *be obtained from* $\mathbf{u}$ *on* γ_i^s*. We instead calculate these coefficients by approximating* $\mathbf{u}$ *by* $\mathbf{u}_h + \mathbf{E}^\perp + \mathbf{E}^\Phi$*, where* $\mathbf{E}^\perp$*,* $\mathbf{E}^\Phi$ *are estimates of the discretization error defined in §6.4. An explicit formula will be given in (6.4.48).*

We can state the following approximation properties of $\breve{\pi}$ and $\breve{\pi}_i^s$.

Lemma 6.1.2. *Let* $\mathbf{v}\in[C^{p+2}(\bar{\omega})]^m$ *and let* $\breve{\pi}$*,* $\breve{\pi}_i^+$*,* $\breve{\pi}_i^-$*,* $\mathbf{c}_i$ *and* $\mathbf{d}_i$*,* $1\le i\le d$*, be defined in Definition 6.1.1. Then* $\breve{\pi}\mathbf{v}$ *and* $\breve{\pi}_i^s\mathbf{v}$ *satisfy the* a priori *bound*

$$\left\| \mathbf{v}(\mathbf{x}) - \breve{\pi}\mathbf{v}(\mathbf{x}) - h^{p+1}\sum_{i=1}^d \left(L_{p+1}(\xi_i)\mathbf{c}_i - L_p(\xi_i)\mathbf{A}_i^{(mod(p,2))}\mathbf{c}_i\right)\left(\frac{x_j}{h}\right)\right\|_{\infty,\omega} \le Ch^{p+2}, \tag{6.1.4}$$

and

$$\left\| \mathbf{v}(\mathbf{x}) - \breve{\pi}_i^s\mathbf{v}(\mathbf{x}) - h^{p+1}\left(\hat{\mathbf{r}}_i^-\left(\frac{x_i}{h}\right) + \sum_{j\in D(i)}\hat{\mathbf{r}}_j\left(\frac{x_j}{h}\right)\right)\right\|_{\infty,\gamma_i^s} \le Ch^{p+2}, \tag{6.1.5a}$$

where

$$\hat{\mathbf{r}}_i(\xi_i) = L_{p+1}(\xi_i)\mathbf{c}_i - L_p(\xi_i)(\mathbf{A}_i^{(mod(p,2))}\mathbf{c}_i + \mathbf{d}_i), \tag{6.1.5b}$$

$$\hat{\mathbf{r}}_i^-(\xi_i) = L_{p+1}(1-\xi_i)\mathbf{c}_i - L_p(1-\xi_i)(\mathbf{A}_i^{(mod(p,2))}\mathbf{c}_i - \mathbf{d}_i),\ 1 \le i \le d. \tag{6.1.5c}$$

Proof. By definition of $\breve{\pi}\mathbf{v}(\mathbf{x})$ and $\hat{\mathbf{r}}_i$, respectively, in (6.1.2a) and (6.1.5b), we have

$$\begin{aligned}
&\left\| \mathbf{v}(\mathbf{x}) - \breve{\pi}\mathbf{v}(\mathbf{x}) - h^{p+1} \sum_{i=1}^{d} \left(L_{p+1}(\xi_i)\mathbf{c}_i - L_p(\xi_i)\mathbf{A}_i^{(\mathrm{mod}(p,2))}\mathbf{c}_i \right) \right\|_{\infty,\omega} \\
&= \left\| \mathbf{v}(\mathbf{x}) - \mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) + \sum_{i=1}^{d} \left(L_{p+1}\left(\frac{x_i}{h}\right) - L_p\left(\frac{x_i}{h}\right)\mathbf{A}_i^{(\mathrm{mod}(p,2))} \right) \left(\bar{\mathbf{c}}_i - h^{p+1}\mathbf{c}_i \right) \right\|_{\infty,\omega} \\
&\le \left\| \mathbf{v}(\mathbf{x}) - \mathcal{L}_{p+1}\mathbf{v}(\mathbf{x}) \right\|_{\infty,\omega} + C' \sum_{i=1}^{d} \left\| \bar{\mathbf{c}}_i - h^{p+1}\mathbf{c}_i \right\|,
\end{aligned} \tag{6.1.6}$$

which, combined with Lemma 2.2.5, yields (6.1.4).

Similarly, by definition of $\breve{\pi}_i^s\mathbf{v}(\mathbf{x})$ and $\hat{\mathbf{r}}_i$, $\hat{\mathbf{r}}_i^-$ in (6.1.3a), we have

$$\begin{aligned}
&\left\| \mathbf{v}(\mathbf{x}) - \breve{\pi}_i^s\mathbf{v}(\mathbf{x}) - h^{p+1} \left(\hat{\mathbf{r}}_i^- \left(\frac{x_i}{h}\right) + \sum_{j \in D(i)} \hat{\mathbf{r}}_j \left(\frac{x_j}{h}\right) \right) \right\|_{\infty,\gamma_i^s} \\
&= \left\| \mathbf{v}(\mathbf{x}) - \mathcal{L}_{p+1}^{i,s}\mathbf{v}(\mathbf{x}) + \sum_{j \in D(i)} \left(L_{p+1}\left(\frac{x_j}{h}\right) - L_p\left(\frac{x_j}{h}\right)\mathbf{A}_j^{(\mathrm{mod}(p,2))} \right) \left(\bar{\mathbf{c}}_{ij}^s - h^{p+1}\mathbf{c}_j \right) \right\|_{\infty,\gamma_i^s} \\
&\le \left\| \mathbf{v}(\mathbf{x}) - \mathcal{L}_{p+1}^{i,s}\mathbf{v}(\mathbf{x}) \right\|_{\infty,\gamma_i^s} + C' \sum_{j \in D(i)} \left\| \bar{\mathbf{c}}_{ij}^s - h^{p+1}\mathbf{c}_j \right\|, \quad 1 \le i \le d,\ s = +,-,
\end{aligned} \tag{6.1.7}$$

which, combined with Lemma 2.2.5, yields (6.1.5). □

6.2 Preliminary Results

We recall the polynomial spaces $\bar{\mathcal{P}}_p$ and $\bar{\mathcal{P}}_p{}^{\perp}$, defined in (4.1.1) as

$$\bar{\mathcal{P}}_p = \left\{ \mathbf{v}(\boldsymbol{\xi}) \in \mathcal{P}_p :\ \sum_{i=1}^{d} \mathbf{A}_i \frac{\partial \mathbf{v}}{\partial \xi_i} = \mathbf{0} \text{ on } \Delta,\ \mathbf{A}_i\mathbf{v} = \mathbf{0} \text{ on } \Gamma_i,\ 1 \le i \le d \right\},$$

and

$$\bar{\mathcal{P}}_p{}^{\perp} = \left\{ \mathbf{w}(\boldsymbol{\xi}) \in [L^2(\Delta)]^m :\ \int_{\Delta} \mathbf{v}^t\mathbf{w}\, d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \bar{\mathcal{P}}_p \right\}.$$

Further we need Lemma 4.1.1, which states that for all integrable functions $\mathbf{f} : (0,1) \to \mathbb{R}^m$, we have

$$\int_\Delta \mathbf{v}^t \mathbf{A}_i \mathbf{f}(\xi_i)\, d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \bar{\mathcal{P}}_p,\ 1 \le i \le d, \tag{6.2.2}$$

and Lemma 4.1.2, which states that for $\mathbf{d} \in \bigoplus_{k=1}^d \mathcal{R}(\mathbf{A}_k)$, we have

$$\int_\Delta \mathbf{v}^t \mathbf{d} L_p(\xi_i)\, d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \bar{\mathcal{P}}_p,\ 1 \le i \le d. \tag{6.2.3}$$

Let us state and show a useful orthogonality property in the following lemma.

Lemma 6.2.1. *Let* $\mathbf{c} \in \mathbb{R}^m,\ \mathbf{d} \in \bigoplus_{k=1}^d \mathcal{R}(\mathbf{A}_k)$. *Then*

$$\int_\Delta \mathbf{v}^t \left(L_{p+1}(\xi_i)\mathbf{c} - L_p(\xi_i)(\mathbf{A}_i^{(mod(p,2))}\mathbf{c} + \mathbf{d}) \right) d\boldsymbol{\xi} = 0, \quad \forall\ \mathbf{v} \in \bar{\mathcal{P}}_p,\ 1 \le i \le d. \tag{6.2.4}$$

Proof. Since $\mathbf{A}_i \mathbf{A}_i^\dagger$ is the identity on $\mathcal{R}(\mathbf{A}_i)$ by Lemma 1.2.13 and $\mathcal{R}(\mathbf{A}_i) = \mathcal{R}(\mathbf{A}_i^{(\mathrm{mod}(p,2))})$ by property (1.2.52c), we obtain

$$\mathbf{A}_i^{(\mathrm{mod}(p,2))} = \mathbf{A}_i \mathbf{A}_i^\dagger \mathbf{A}_i^{(\mathrm{mod}(p,2))}, \quad 1 \le i \le d, \tag{6.2.5}$$

which, when combined with (6.2.2), yields

$$\int_\Delta \mathbf{v}^t \mathbf{A}_i^{(\mathrm{mod}(p,2))} \mathbf{c} L_p(\xi_i)\, d\boldsymbol{\xi} = \int_\Delta \mathbf{v}^t \mathbf{A}_i \left(\mathbf{A}_i^\dagger \mathbf{A}_i^{(\mathrm{mod}(p,2))} \mathbf{c} L_p(\xi_i) \right) d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \bar{\mathcal{P}}_p,\ 1 \le i \le d. \tag{6.2.6}$$

Equation (6.2.3) and the orthogonality properties (2.2.6) combined yield

$$\int_\Delta \mathbf{v}^t \left(L_{p+1}(\xi_i)\mathbf{c} - L_p(\xi_i)\mathbf{d} \right) d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \bar{\mathcal{P}}_p,\ 1 \le i \le d. \tag{6.2.7}$$

Adding (6.2.6) and (6.2.7) yields (6.2.4). □

Next, we show another lemma that is needed for the proof of Theorem 6.3.1.

Lemma 6.2.2. *If* $\mathbf{q} \in \mathcal{P}_p$ *satisfies*

$$\sum_{i=1}^d \int_\Delta \left(\frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \mathbf{q}\, d\boldsymbol{\xi} - \int_{\Gamma_i} \mathbf{v}^t \nu_i \breve{\mathbf{A}}_i^{\mu_i} \mathbf{q}\, d\boldsymbol{\sigma} \right) = 0,\ \forall\ \mathbf{v} \in \mathcal{P}_p, \tag{6.2.8}$$

then $\mathbf{q} \in \bar{\mathcal{P}}_p$.

Proof. First we integrate (6.2.8) by parts to obtain

$$\sum_{i=1}^{d}\left(-\int_{\Delta}\mathbf{v}^t\mathbf{A}_i\frac{\partial\mathbf{q}}{\partial\xi_i}\,d\boldsymbol{\xi}+\int_{\Gamma_i}\mathbf{v}^t\nu_i\breve{\mathbf{A}}_i^{\bar{\mu}_i}\mathbf{q}\,d\boldsymbol{\sigma}\right)=0,\ \forall\ \mathbf{v}\in\mathcal{P}_p. \tag{6.2.9}$$

Adding (6.2.8) to (6.2.9), testing against $\mathbf{v}=-\mathbf{q}$, and using the symmetry of $\mathbf{A}_i$, $1\le i\le d$, we obtain

$$\sum_{i=1}^{d}\int_{\Gamma_i}\mathbf{q}^t\nu_i(\breve{\mathbf{A}}_i^{\mu_i}-\breve{\mathbf{A}}_i^{\bar{\mu}_i})\mathbf{q}\,d\boldsymbol{\sigma}=\sum_{i=1}^{d}\int_{\Gamma_i}\mathbf{q}^t(\breve{\mathbf{A}}_i^{+}-\breve{\mathbf{A}}_i^{-})\mathbf{q}\,d\boldsymbol{\sigma}=0. \tag{6.2.10}$$

$(\breve{\mathbf{A}}_i^{+}-\breve{\mathbf{A}}_i^{-})$ is symmetric positive semi-definite by (1.2.52g), and therefore admits a Cholesky factorization $(\breve{\mathbf{A}}_i^{+}-\breve{\mathbf{A}}_i^{-})=\mathbf{L}_i^t\mathbf{L}_i$. Hence (6.2.10) can be written as

$$\sum_{i=1}^{d}\int_{\Gamma_i}\|\mathbf{L}_i\mathbf{q}\|^2\,d\boldsymbol{\sigma}=0. \tag{6.2.11}$$

Therefore, $\mathbf{L}_i\mathbf{q}=\mathbf{0}$ on Γ_i, which yields

$$\mathbf{L}_i^t(\mathbf{L}_i\mathbf{q})=(\breve{\mathbf{A}}_i^{+}-\breve{\mathbf{A}}_i^{-})\mathbf{q}=\mathbf{0}\text{ on }\Gamma_i,\ 1\le i\le d, \tag{6.2.12}$$

which combined with property (1.2.52b) leads to

$$\mathbf{A}_i\mathbf{v}=\mathbf{0}\text{ on }\Gamma_i,\ 1\le i\le d, \tag{6.2.13}$$

By (6.2.13), the boundary integral in (6.2.9) vanishes, thus substituting $\mathbf{v}=-\sum_{i=1}^{d}\mathbf{A}_i\frac{\partial\mathbf{q}}{\partial\xi_i}$ in (6.2.9) yields

$$\int_{\Delta}\left\|\sum_{i=1}^{d}\mathbf{A}_i\frac{\partial\mathbf{q}}{\partial\xi_i}\right\|^2\,d\boldsymbol{\xi}=0, \tag{6.2.14}$$

which in turn yields

$$\sum_{i=1}^{d}\mathbf{A}_i\frac{\partial\mathbf{q}}{\partial\xi_i}=\mathbf{0}\text{ on }\Delta. \tag{6.2.15}$$

Combining (6.2.13) and (6.2.15) proves the lemma. □

6.3 Error Analysis

Below we prove a theorem for the spatial discretization error for the Lax-Friedrichs numerical flux.

Theorem 6.3.1. *Let $\mathbf{u} \in [C^2([0,T], C^{p+2}(\bar{\omega}))]^m$ be the solution of (6.0.3) and let $\mathbf{u}_h \in \mathcal{P}_p$ satisfy (6.1.1). Further, let either p be odd or all matrices $\mathbf{A}_i$, $1 \leq i \leq d$, be invertible.*

Then the local finite element error on ω, at $t = \mathcal{O}(1)$ and for $p \geq 1$, can be written as

$$\mathbf{u}(t, h\boldsymbol{\xi}) - \mathbf{u}_h(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} \mathbf{r}_i(t, h\xi_i) + \mathcal{O}(h^{p+2}), \ \boldsymbol{\xi} \in \Delta, \tag{6.3.1a}$$

where

$$\mathbf{r}_i(t, h\xi_i) = L_{p+1}(\xi_i)\mathbf{c}_i(t) - L_p(\xi_i)(\mathbf{A}_i^{(mod(p,2))}\mathbf{c}_i(t) + \mathbf{d}_i(t)), \ 1 \leq i \leq d, \tag{6.3.1b}$$

with

$$\mathbf{c}_i(t) = \frac{1}{a_{p+1}} \frac{1}{(p+1)!} \frac{\partial^{p+1}\mathbf{u}(t, \mathbf{0})}{\partial x_i^{p+1}}, \quad \mathbf{d}_i(t) \in \mathcal{N}(\mathbf{A}_i) \cap \bigoplus_{k=1}^{d} \mathcal{R}(\mathbf{A}_k). \tag{6.3.1c}$$

Proof. The proof follows closely that of Theorem 4.2.1, where we replace $\mathbf{A}_i^+$, $\mathbf{A}_i^-$ and $\operatorname{sgn}(\mathbf{A}_i)$ by $\breve{\mathbf{A}}_i^+$, $\breve{\mathbf{A}}_i^-$ and $\mathbf{A}_i^{(\mathrm{mod}(p,2))}$ and the initial and boundary conditions π and π_i^s by $\breve{\pi}$ and $\breve{\pi}_i^s$, $s = +, -$, $1 \leq i \leq d$.

We derive the *orthogonality condition* for the error $\mathbf{e} = \mathbf{u} - \mathbf{u}_h$ by noting that the exact solution $\mathbf{u}$ satisfies

$$\int_\omega \mathbf{v}^t \left(\frac{\partial \mathbf{u}}{\partial t} - \mathbf{g} \right) d\mathbf{x} = \sum_{i=1}^{d} \left(\int_\omega \frac{\partial \mathbf{v}^t}{\partial x_i} \mathbf{A}_i \mathbf{u} \, d\mathbf{x} - \int_{\gamma_i} \mathbf{v}^t \nu_i \mathbf{A}_i \mathbf{u} \, d\mathbf{s} \right), \ \forall \ \mathbf{v} \in \mathcal{P}_p, \ 0 < t < T, \tag{6.3.2}$$

which, subtracted from (6.1.1a), yields

$$\int_\omega \mathbf{v}^t \frac{\partial \mathbf{e}}{\partial t} \, d\mathbf{x} = \sum_{i=1}^{d} \left(\int_\omega \frac{\partial \mathbf{v}^t}{\partial x_i} \mathbf{A}_i \mathbf{e} \, d\mathbf{x} - \int_{\gamma_i} \mathbf{v}^t \nu_i (\breve{\mathbf{A}}_i^{\mu_i} \mathbf{e} + \breve{\mathbf{A}}_i^{\bar{\mu}_i} \mathbf{e}^-) \, d\mathbf{s} \right), \ \forall \ \mathbf{v} \in \mathcal{P}_p, \ 0 < t < T. \tag{6.3.3}$$

Apply the scalings $\tau = T^{-1}t$ and $\boldsymbol{\xi} = h^{-1}\mathbf{x}$ and write $\hat{\mathbf{e}}(\tau, \boldsymbol{\xi}) = \mathbf{e}(T\tau, h\boldsymbol{\xi})$ to write

$$\frac{h}{T} \int_\Delta \mathbf{v}^t \frac{\partial \hat{\mathbf{e}}}{\partial \tau} \, d\boldsymbol{\xi} = \sum_{i=1}^{d} \left(\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \hat{\mathbf{e}} \, d\boldsymbol{\xi} - \int_{\Gamma_i} \mathbf{v}^t \nu_i (\breve{\mathbf{A}}_i^{\mu_i} \hat{\mathbf{e}} + \breve{\mathbf{A}}_i^{\bar{\mu}_i} \hat{\mathbf{e}}^-) \, d\boldsymbol{\sigma} \right), \ \mathbf{v} \in \mathcal{P}_p, \ 0 < \tau < 1. \tag{6.3.4}$$

Now note that, since $\mathcal{P}_p$ is a subspace of $[L^2(\Delta)]^m$, we can split $\hat{\mathbf{e}}$,

$$\hat{\mathbf{e}} = \bar{\mathbf{e}} + \tilde{\mathbf{e}}, \quad \bar{\mathbf{e}} \in \bar{\mathcal{P}}_p, \quad \tilde{\mathbf{e}} \in \bar{\mathcal{P}}_p{}^{\perp}. \tag{6.3.5}$$

We establish Theorem 6.3.1 by showing that

$$\bar{\mathbf{e}}(\tau, \boldsymbol{\xi}) = \mathcal{O}(h^{p+2}), \quad \boldsymbol{\xi} \in \Delta, \ 0 \leq \tau \leq 1, \tag{6.3.6}$$

and

$$\tilde{\mathbf{e}}(\tau,\boldsymbol{\xi}) = h^{p+1}\sum_{i=1}^{d}\hat{\mathbf{r}}_i(\tau,\xi_i) + \mathcal{O}(h^{p+2}),\ \boldsymbol{\xi}\in\Delta, 0\le\tau\le 1, \tag{6.3.7}$$

where $\hat{\mathbf{r}}_i(\tau,\xi_i) = \mathbf{r}_i(T\tau, h\xi_i)$.

Since $\bar{\mathcal{P}}_p$ is a finite dimensional vector space and $\bar{\mathbf{e}}\in\bar{\mathcal{P}}_p$, we have

$$\frac{\partial\bar{\mathbf{e}}}{\partial\tau}(\tau,\boldsymbol{\xi}) = \lim_{h\to 0}\frac{\bar{\mathbf{e}}(\tau+h,\boldsymbol{\xi})-\bar{\mathbf{e}}(\tau,\boldsymbol{\xi})}{h}\in\bar{\mathcal{P}}_p, \tag{6.3.8a}$$

$$\frac{\partial\tilde{\mathbf{e}}}{\partial\tau}(\tau,\boldsymbol{\xi}) = \lim_{h\to 0}\frac{\tilde{\mathbf{e}}(\tau+h,\boldsymbol{\xi})-\tilde{\mathbf{e}}(\tau,\boldsymbol{\xi})}{h}\in\bar{\mathcal{P}}_p{}^{\perp},\quad \boldsymbol{\xi}\in\Delta,\ 0<\tau<1. \tag{6.3.8b}$$

By the definition of $\bar{\mathcal{P}}_p$ and the symmetry of $\mathbf{A}_i$, $\breve{\mathbf{A}}_i^{+}$, and $\breve{\mathbf{A}}_i^{-}$, $1\le i\le d$, (6.3.4) yields for $\mathbf{v}\in\bar{\mathcal{P}}_p$

$$\begin{aligned}\frac{h}{T}\int_\Delta \mathbf{v}^t\frac{\partial\hat{\mathbf{e}}}{\partial\tau}\,d\boldsymbol{\xi} &= \sum_{i=1}^{d}\left(\int_\Delta\left(\mathbf{A}_i\frac{\partial\mathbf{v}}{\partial\xi_i}\right)^t\hat{\mathbf{e}}\,d\boldsymbol{\xi} - \int_{\Gamma_i}\left(\breve{\mathbf{A}}_i^{\mu_i}\mathbf{v}\right)^t\nu_i\hat{\mathbf{e}} + \left(\breve{\mathbf{A}}_i^{\bar{\mu}_i}\mathbf{v}\right)^t\nu_i\hat{\mathbf{e}}^{-}\,d\boldsymbol{\sigma}\right)\\ &= 0,\quad \forall\ \mathbf{v}\in\bar{\mathcal{P}}_p,\quad 0<\tau<1.\end{aligned} \tag{6.3.9}$$

Thus, $\frac{\partial\hat{\mathbf{e}}}{\partial\tau}\in\bar{\mathcal{P}}_p{}^{\perp}$, which combined with (6.3.5) and (6.3.8b) yields

$$\frac{\partial\bar{\mathbf{e}}}{\partial\tau} = \frac{\partial\hat{\mathbf{e}}}{\partial\tau} - \frac{\partial\tilde{\mathbf{e}}}{\partial\tau}\in\bar{\mathcal{P}}_p{}^{\perp},\quad 0<\tau<1. \tag{6.3.10}$$

Combining (6.3.10) and (6.3.8a) shows that

$$\frac{\partial\bar{\mathbf{e}}}{\partial\tau}(\tau,\boldsymbol{\xi}) = \mathbf{0},\quad \boldsymbol{\xi}\in\Delta,\quad 0<\tau<1. \tag{6.3.11}$$

Now we show that

$$\bar{\mathbf{e}}(0,\boldsymbol{\xi}) = \mathcal{O}(h^{p+2}),\quad \boldsymbol{\xi}\in\Delta. \tag{6.3.12}$$

if the initial conditions are given by either $\mathbf{u}_h|_{t=0} = \Pi\mathbf{u}_0$ or $\mathbf{u}_h|_{t=0} = \pi\mathbf{u}_0$.

If $\mathbf{u}_h|_{t=0} = \Pi\mathbf{u}_0$, then Lemma 2.2.4 yields

$$\hat{\mathbf{e}}(0,\boldsymbol{\xi}) = \mathbf{u}_0(h\boldsymbol{\xi}) - \Pi\mathbf{u}_0(h\boldsymbol{\xi}) = h^{p+1}\sum_{i=1}^{d}L_{p+1}(\xi_i)\mathbf{c}_i + \mathcal{O}(h^{p+2}),\quad \boldsymbol{\xi}\in\Delta. \tag{6.3.13}$$

By the orthogonality properties (2.2.6), we obtain

$$h^{p+1}\sum_{i=1}^{d}L_{p+1}(\xi_i)\mathbf{c}_i\in\bar{\mathcal{P}}_p{}^{\perp},\quad 1\le i\le d, \tag{6.3.14}$$

Splitting the $\mathcal{O}(h^{p+2})$-term into parts in $\bar{\mathcal{P}}_p$ and $\bar{\mathcal{P}}_p{}^{\perp}$, combined with (6.3.14), yields (6.3.12).

If $\mathbf{u}_h|_{t=0} = \pi\mathbf{u}_0$, then Lemma 6.1.2 yields

$$\hat{\mathbf{e}}(0,\boldsymbol{\xi}) = \mathbf{u}_0(h\boldsymbol{\xi}) - \breve{\pi}\mathbf{u}_0(h\boldsymbol{\xi}) = h^{p+1}\sum_{i=1}^{d}\mathbf{r}_i(0,\xi_i) + \mathcal{O}(h^{p+2}), \quad \boldsymbol{\xi}\in\Delta. \tag{6.3.15}$$

By Lemma 6.2.1, we obtain

$$h^{p+1}\sum_{i=1}^{d}\mathbf{r}_i(0,\xi_i) \in \bar{\mathcal{P}}_p{}^{\perp}, \quad 1\le i\le d, \tag{6.3.16}$$

Splitting the $\mathcal{O}(h^{p+2})$-term into parts in $\bar{\mathcal{P}}_p$ and $\bar{\mathcal{P}}_p{}^{\perp}$, combined with (6.3.15), yields (6.3.12).

By the Fundamental Theorem of Calculus, (6.3.11) and (6.3.12) yields (6.3.6).

In the remainder of the proof, we will investigate the asymptotic behavior of $\tilde{\mathbf{e}}$. We write the Maclaurin series of $\hat{\mathbf{e}}$ with respect to the mesh parameter h as

$$\hat{\mathbf{e}}(\tau,\boldsymbol{\xi}) = \sum_{k=0}^{p+1} h^k\mathbf{q}_k(\tau,\boldsymbol{\xi}) + \mathcal{O}(h^{p+2}), \quad \boldsymbol{\xi}\in\Delta,\ 0<\tau<1, \tag{6.3.17}$$

where, since $\mathbf{u}_h$ is a function of $T\tau$, $h\boldsymbol{\xi}$, and h,

$$\mathbf{q}_k(\tau,\boldsymbol{\xi}) = \frac{1}{k!}\frac{d^k(\mathbf{u}(T\tau,h\boldsymbol{\xi}) - \mathbf{u}_h(T\tau,h\boldsymbol{\xi},h)}{dh^k}\bigg|_{h=0}. \tag{6.3.18}$$

We write the Maclaurin series of $\tilde{\mathbf{e}}\in\bar{\mathcal{P}}_p{}^{\perp}$ with respect to the mesh parameter h as

$$\tilde{\mathbf{e}}(\tau,\boldsymbol{\xi}) = \sum_{k=0}^{\infty} h^k\tilde{\mathbf{q}}_k(\tau,\boldsymbol{\xi}), \quad \tilde{\mathbf{q}}_k\in\bar{\mathcal{P}}_p{}^{\perp},\ \boldsymbol{\xi}\in\Delta,\ 0<\tau<1. \tag{6.3.19}$$

By (6.3.5) and (6.3.6), $\hat{\mathbf{e}} = \tilde{\mathbf{e}} + \mathcal{O}(h^{p+2})$, thus subtracting (6.3.17) from (6.3.19) and setting all terms having the same power of h equal yields

$$\mathbf{q}_k = \tilde{\mathbf{q}}_k \in \bar{\mathcal{P}}_p{}^{\perp},\ 0\le k\le p+1. \tag{6.3.20}$$

By Lemma 6.1.2, the boundary conditions satisfy

$$\begin{aligned}\hat{\mathbf{e}}^-(\tau,\boldsymbol{\xi}) &= \mathbf{u}(t,h\boldsymbol{\xi}) - \breve{\pi}_i^s\mathbf{u}(t,h\boldsymbol{\xi}) &\text{(6.3.21a)}\\ &= h^{p+1}\big(\hat{\mathbf{r}}_i^-(\tau,\xi_j) + \sum_{j\in D(i)}\hat{\mathbf{r}}_j(\tau,\xi_j)\big) + \mathcal{O}(h^{p+2}),\ \boldsymbol{\xi}\in\Gamma_i^s,\ s=+,-,\ 1\le i\le d,\end{aligned}$$

where

$$\hat{\mathbf{r}}_i^-(\xi_i) = L_{p+1}(1-\xi_i)\mathbf{c}_i(T\tau) - L_p(1-\xi_i)(\mathbf{A}_i^{(\mathrm{mod}(p,2))}\mathbf{c}_i(T\tau) - \mathbf{d}_i(T\tau)),\ 1 \le i \le d. \quad (6.3.21b)$$

Substituting (6.3.17) and (6.3.21) into the orthogonality condition (6.3.4) yields

$$\sum_{k=0}^{p+1} h^k \left(\frac{h}{T} \int_\Delta \mathbf{v}^t \frac{\partial \mathbf{q_k}}{\partial \tau}\, d\boldsymbol{\xi} - \sum_{i=1}^{d} \left(\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \mathbf{q}_k\, d\boldsymbol{\xi} + \int_{\Gamma_i} \mathbf{v}^t \nu_i \breve{\mathbf{A}}_i^{\mu_i} \mathbf{q}_k\, d\boldsymbol{\sigma} \right) \right)$$
$$= -h^{p+1} \sum_{i=1}^{d} \int_{\Gamma_i} \mathbf{v}^t \nu_i \breve{\mathbf{A}}_i^{\bar{\mu}_i} \left(\hat{\mathbf{r}}_i^- + \sum_{j \in D(i)} \hat{\mathbf{r}}_j \right) d\boldsymbol{\sigma} + \mathcal{O}(h^{p+2}),\ \mathbf{v} \in \mathcal{P}_p. \quad (6.3.22)$$

Now assume $T = \mathcal{O}(1)$ and set to zero all terms in (6.3.22) having the same power of h.

Thus the $\mathcal{O}(1)$ term $\mathbf{q}_0$ satisfies the orthogonality condition

$$\sum_{i=1}^{d} \left(\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \mathbf{q}_0\, d\boldsymbol{\xi} - \int_{\Gamma_i} \mathbf{v}^t \nu_i \breve{\mathbf{A}}_i^{\mu_i} \mathbf{q}_0\, d\boldsymbol{\sigma} \right) = 0,\ \mathbf{v} \in \mathcal{P}_p. \quad (6.3.23)$$

Lemma 6.2.2 yields $\mathbf{q}_0 \in \bar{\mathcal{P}}_p$, which combined with (6.3.20) shows that $\mathbf{q}_0 = \mathbf{0}$ on Δ.

Assume that $\mathbf{q}_j = \mathbf{0}$ for all $0 \le j \le k-1$, where $k \le p$. Thus, the $\mathcal{O}(h^k)$ term is written as

$$\sum_{i=1}^{d} \left(\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \mathbf{q}_k\, d\boldsymbol{\xi} - \int_{\Gamma_i} \mathbf{v}^t \nu_i \breve{\mathbf{A}}_i^{\mu_i} \mathbf{q}_k\, d\boldsymbol{\sigma} \right) = 0,\ \mathbf{v} \in \mathcal{P}_p. \quad (6.3.24)$$

Lemma 6.2.2 yields $\mathbf{q}_k \in \bar{\mathcal{P}}_p$, which combined with (6.3.20) shows that $\mathbf{q}_k = \mathbf{0}$ on Δ for $0 \le k \le p$.

The $\mathcal{O}(h^{p+1})$ term satisfies the orthogonality condition

$$\sum_{i=1}^{d} \left(\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \mathbf{q}_{p+1}\, d\boldsymbol{\xi} - \int_{\Gamma_i} \mathbf{v}^t \nu_i \left(\breve{\mathbf{A}}_i^{\mu_i} \mathbf{q}_{p+1} - \breve{\mathbf{A}}_i^{\bar{\mu}_i} \left(\sum_{j \in D(i)} \hat{\mathbf{r}}_j + \hat{\mathbf{r}}_i^- \right) \right) d\boldsymbol{\sigma} \right) = 0,\ \forall\ \mathbf{v} \in \mathcal{P}_p. \quad (6.3.25)$$

We will first show that $\mathbf{q}_{p+1} = \sum_{i=1}^{d} \hat{\mathbf{r}}_i + \mathbf{p},\ \mathbf{p} \in \mathcal{P}_p$.

Since $\frac{\partial^{p+1}}{\partial x_i^{p+1}} \mathbf{u}_h = \mathbf{0}$ for $1 \le i \le d$, (6.3.18) yields for $k = p+1$

$$\mathbf{q}_{p+1}(\tau, \boldsymbol{\xi}) = \frac{1}{k!} \frac{d^k (\mathbf{u} - \mathbf{u}_h)(T\tau, \boldsymbol{\xi} h, h)}{dh^k} \bigg|_{h=0} = \sum_{|\boldsymbol{\alpha}| \le p+1} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} (\mathbf{u} - \mathbf{u}_h)(T\tau, \mathbf{0}) \boldsymbol{\xi}^{\boldsymbol{\alpha}}$$
$$= \sum_{i=1}^{d} \frac{1}{(p+1)!} \frac{\partial^{p+1} \mathbf{u}(T\tau, \mathbf{0})}{\partial x_i^{p+1}} \xi_i^{p+1} + \mathbf{p}_1(\tau, \boldsymbol{\xi}),\ \mathbf{p}_1 \in \mathcal{P}_p \quad (6.3.26)$$

By the definition of $\mathbf{c}_i$ in (6.3.1c),

$$\begin{aligned} L_{p+1}(\xi_i)\mathbf{c}_i(T\tau) &= \frac{1}{a_{p+1}}\frac{1}{(p+1)!}\frac{\partial^{p+1}\mathbf{u}(T\tau,\mathbf{0})}{\partial x_i^{p+1}}L_{p+1}(\xi_i) \\ &= \frac{1}{(p+1)!}\frac{\partial^{p+1}\mathbf{u}(T\tau,\mathbf{0})}{\partial x_i^{p+1}}\xi_i^{p+1} + \breve{\mathbf{p}}_i(\tau,\boldsymbol{\xi}),\ \breve{\mathbf{p}}_i \in \mathcal{P}_p \end{aligned} \tag{6.3.27}$$

Substituting (6.3.27) into (6.3.1b) yields

$$\begin{aligned} \sum_{i=1}^{d}\hat{\mathbf{r}}_i(\tau,\xi_i) &= \sum_{i=1}^{d}(L_{p+1}(\xi_i)\mathbf{c}_i(T\tau) - L_p(\xi_i)(\mathbf{A}_i^{(\mathrm{mod}(p,2))}\mathbf{c}_i(T\tau) + \mathbf{d}_i(T\tau))) \\ &= \sum_{i=1}^{d}\frac{1}{(p+1)!}\frac{\partial^{p+1}\mathbf{u}(T\tau,\mathbf{0})}{\partial x_i^{p+1}}\xi_i^{p+1} + \mathbf{p}_2(\tau,\boldsymbol{\xi}), \end{aligned} \tag{6.3.28}$$

where

$$\mathbf{p}_2 = \sum_{i=1}^{d}\Big(\breve{\mathbf{p}}_i - L_p(\xi_i)(\mathbf{A}_i^{(\mathrm{mod}(p,2))}\mathbf{c}_i + \mathbf{d}_i)\Big) \in \mathcal{P}_p. \tag{6.3.29}$$

Combining (6.3.28) and (6.3.26) yields

$$\mathbf{q}_{p+1}(\tau,\boldsymbol{\xi}) = \sum_{i=1}^{d}\hat{\mathbf{r}}_i(\tau,\xi_i) + \mathbf{p}(\tau,\boldsymbol{\xi}),\ \mathbf{p} = \mathbf{p}_1 - \mathbf{p}_2 \in \mathcal{P}_p. \tag{6.3.30}$$

Substituting (6.3.30) into (6.3.25) yields

$$\begin{aligned} &\sum_{i=1}^{d}\left(\int_\Delta \frac{\partial \mathbf{v}^t}{\partial\xi_i}\mathbf{A}_i\mathbf{p}\,d\boldsymbol{\xi} - \int_{\Gamma_i}\mathbf{v}^t\nu_i\breve{\mathbf{A}}_i^{\mu_i}\mathbf{p}\,d\boldsymbol{\sigma}\right) \\ &= \sum_{i=1}^{d}\left(-\int_\Delta\frac{\partial\mathbf{v}^t}{\partial\xi_i}\mathbf{A}_i\sum_{j=1}^{d}\hat{\mathbf{r}}_j\,d\boldsymbol{\xi} + \int_{\Gamma_i}\mathbf{v}^t\nu_i\left(\breve{\mathbf{A}}_i^{\mu_i}\sum_{j=1}^{d}\hat{\mathbf{r}}_j + \breve{\mathbf{A}}_i^{\bar{\mu}_i}\left(\sum_{j\in D(i)}\hat{\mathbf{r}}_j + \hat{\mathbf{r}}_i^-\right)\right)d\boldsymbol{\sigma}\right) \\ &= \sum_{i=1}^{d}\big(T_1^i(\mathbf{v}) + \sum_{j\in D(i)}T_2^{i,j}(\mathbf{v}) + T_3^i(\mathbf{v})\big), \quad \forall\ \mathbf{v}\in\mathcal{P}_p, \end{aligned} \tag{6.3.31a}$$

where

$$T_1^i(\mathbf{v}) = -\int_\Delta\frac{\partial\mathbf{v}^t}{\partial\xi_i}\mathbf{A}_i\hat{\mathbf{r}}_i\,d\boldsymbol{\xi}, \tag{6.3.31b}$$

$$T_2^{i,j}(\mathbf{v}) = -\int_\Delta\frac{\partial\mathbf{v}^t}{\partial\xi_i}\mathbf{A}_i\hat{\mathbf{r}}_j\,d\boldsymbol{\xi} + \int_{\Gamma_i}\mathbf{v}^t\nu_i\mathbf{A}_i\hat{\mathbf{r}}_j\,d\boldsymbol{\sigma}, \tag{6.3.31c}$$

$$T_3^i(\mathbf{v}) = \int_{\Gamma_i}\mathbf{v}^t\nu_i(\breve{\mathbf{A}}_i^{\mu_i}\hat{\mathbf{r}}_i + \breve{\mathbf{A}}_i^{\bar{\mu}_i}\hat{\mathbf{r}}_i^-)\,d\boldsymbol{\sigma},\ j\in D(i),\ 1\le i\le d. \tag{6.3.31d}$$

Next, we show that $T_1^i(\mathbf{v}) = T_2^{i,j}(\mathbf{v}) = T_3^i(\mathbf{v}) = 0$ for all $\mathbf{v} \in \mathcal{P}_p,\ j \in D(i),\ 1 \leq i \leq d$.
By the orthogonality properties of Legendre polynomials, we have

$$T_1^i(\mathbf{v}) = -\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \hat{\mathbf{r}}_i \, d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \mathcal{P}_p,\ 1 \leq i \leq d. \tag{6.3.32}$$

Integration (6.3.31c) by parts w.r.t. ξ_i yields

$$T_2^{i,j}(\mathbf{v}) = \int_\Delta \mathbf{v}^t \mathbf{A}_i \frac{\partial \hat{\mathbf{r}}_j}{\partial \xi_i} \, d\boldsymbol{\xi} = 0,\ \forall\ \mathbf{v} \in \mathcal{P}_p,\ 1 \leq i \leq d,\ j \in D(i), \tag{6.3.33}$$

since $\hat{\mathbf{r}}_j(t, \xi_j)$ is independent of ξ_i for $j \in D(i)$.

To show that $T_3^i(\mathbf{v}) = 0$ for all $\mathbf{v} \in \mathcal{P}_p,\ 1 \leq i \leq d$, we first show

$$\breve{\mathbf{A}}_i^{\mu_i} \hat{\mathbf{r}}_i + \breve{\mathbf{A}}_i^{\bar{\mu}_i} \hat{\mathbf{r}}_i^- \big|_{\Gamma_i^s} = \mathbf{0}, \quad 1 \leq i \leq d,\ s = +, -. \tag{6.3.34}$$

If p is odd, then substituting the definitions of $\mathbf{r}_i$ and $\mathbf{r}_i^-$ into (6.3.34) yields

$$\begin{aligned}
\breve{\mathbf{A}}_i^{+} \hat{\mathbf{r}}_i + \breve{\mathbf{A}}_i^{-} \hat{\mathbf{r}}_i^- \big|_{\Gamma_i^+} &= \breve{\mathbf{A}}_i^{+} (\mathbf{c}_i - \mathbf{A}_i^{(1)} \mathbf{c}_i - \mathbf{d}_i) + \breve{\mathbf{A}}_i^{-} (\mathbf{c}_i + \mathbf{A}_i^{(1)} \mathbf{c}_i - \mathbf{d}_i) && \text{(6.3.35a)} \\
&= \left(\breve{\mathbf{A}}_i^{+} - \breve{\mathbf{A}}_i^{+} \mathbf{A}_i^{(1)} + \breve{\mathbf{A}}_i^{-} + \breve{\mathbf{A}}_i^{-} \mathbf{A}_i^{(1)} \right) \mathbf{c}_i - \mathbf{A}_i \mathbf{d}_i = \mathbf{0}, \\
\breve{\mathbf{A}}_i^{-} \hat{\mathbf{r}}_i + \breve{\mathbf{A}}_i^{+} \hat{\mathbf{r}}_i^- \big|_{\Gamma_i^-} &= \breve{\mathbf{A}}_i^{+} (\mathbf{c}_i - \mathbf{A}_i^{(1)} \mathbf{c}_i + \mathbf{d}_i) + \breve{\mathbf{A}}_i^{-} (\mathbf{c}_i + \mathbf{A}_i^{(1)} \mathbf{c}_i + \mathbf{d}_i) && \text{(6.3.35b)} \\
&= \left(\breve{\mathbf{A}}_i^{+} - \breve{\mathbf{A}}_i^{+} \mathbf{A}_i^{(1)} + \breve{\mathbf{A}}_i^{-} + \breve{\mathbf{A}}_i^{-} \mathbf{A}_i^{(1)} \right) \mathbf{c}_i + \mathbf{A}_i \mathbf{d}_i = \mathbf{0},\ 1 \leq i \leq d,
\end{aligned}$$

where we used the fact that $L_p(0) = -1$, $L_{p+1}(0) = L_{p+1}(1) = L_p(1) = 1$, $\breve{\mathbf{A}}_i^{+} - \breve{\mathbf{A}}_i^{+} \mathbf{A}_i^{(1)} + \breve{\mathbf{A}}_i^{-} + \breve{\mathbf{A}}_i^{-} \mathbf{A}_i^{(0)} = \mathbf{0}$ by (1.2.52d), and $\mathbf{d}_i \in \mathcal{N}(\mathbf{A}_i)$.

If p is even and all matrices $\mathbf{A}_i$, $1 \leq i \leq d$, are invertible, *i.e.* $\mathbf{d}_i = \mathbf{0}$, then substituting the definitions of $\mathbf{r}_i$ and $\mathbf{r}_i^-$ into (6.3.34) yields

$$\begin{aligned}
\breve{\mathbf{A}}_i^{+} \hat{\mathbf{r}}_i + \breve{\mathbf{A}}_i^{-} \hat{\mathbf{r}}_i^- \big|_{\Gamma_i^+} &= \breve{\mathbf{A}}_i^{+} (\mathbf{c}_i - \mathbf{A}_i^{(0)} \mathbf{c}_i) - \breve{\mathbf{A}}_i^{-} (\mathbf{c}_i + \mathbf{A}_i^{(0)} \mathbf{c}_i) \\
&= \left(\breve{\mathbf{A}}_i^{+} - \breve{\mathbf{A}}_i^{+} \mathbf{A}_i^{(0)} - \breve{\mathbf{A}}_i^{-} - \breve{\mathbf{A}}_i^{-} \mathbf{A}_i^{(0)} \right) \mathbf{c}_i = \mathbf{0}, && \text{(6.3.36a)} \\
\breve{\mathbf{A}}_i^{-} \hat{\mathbf{r}}_i + \breve{\mathbf{A}}_i^{+} \hat{\mathbf{r}}_i^- \big|_{\Gamma_i^-} &= \breve{\mathbf{A}}_i^{+} (\mathbf{c}_i - \mathbf{A}_i^{(0)} \mathbf{c}_i) - \breve{\mathbf{A}}_i^{-} (\mathbf{c}_i + \mathbf{A}_i^{(0)} \mathbf{c}_i) \\
&= \left(\breve{\mathbf{A}}_i^{+} - \breve{\mathbf{A}}_i^{+} \mathbf{A}_i^{(0)} - \breve{\mathbf{A}}_i^{-} - \breve{\mathbf{A}}_i^{-} \mathbf{A}_i^{(0)} \right) \mathbf{c}_i = \mathbf{0},\ 1 \leq i \leq d, && \text{(6.3.36b)}
\end{aligned}$$

where we used the fact that $L_{p+1}(0) = -1$, $L_p(0) = L_p(1) = L_{p+1}(1) = 1$ and $\breve{\mathbf{A}}_i^{+} - \breve{\mathbf{A}}_i^{+} \mathbf{A}_i^{(0)} - \breve{\mathbf{A}}_i^{-} - \breve{\mathbf{A}}_i^{-} \mathbf{A}_i^{(0)} = \mathbf{0}$ by (1.2.52e).

Thus, (6.3.34) is true, if either p is odd or all matrices $\mathbf{A}_i$, $1 \leq i \leq d$, are invertible.

Substituting (6.3.34) into (6.3.31d), yields

$$T_3^i(\mathbf{v}) = \int_{\Gamma_i} \mathbf{v}^t \nu_i (\breve{\mathbf{A}}_i^{\mu_i} \hat{\mathbf{r}}_i + \breve{\mathbf{A}}_i^{\bar{\mu}_i} \hat{\mathbf{r}}_i^-)\, d\boldsymbol{\sigma} = 0, \quad \forall\ \mathbf{v} \in \mathcal{P}_p,\ 1 \le i \le d. \tag{6.3.37}$$

Substituting (6.3.32), (6.3.33), and (6.3.37) into (6.3.31a) leads to

$$\int_\Delta \frac{\partial \mathbf{v}^t}{\partial \xi_i} \mathbf{A}_i \mathbf{p}\, d\boldsymbol{\xi} - \int_{\Gamma_i} \mathbf{v}^t \nu_i \breve{\mathbf{A}}_i^{\mu_i} \mathbf{p}\, d\boldsymbol{\sigma} = 0,\ \mathbf{v} \in \mathcal{P}_p, \tag{6.3.38}$$

which, combined with Lemma 6.2.2, yields

$$\mathbf{p} \in \bar{\mathcal{P}}_p. \tag{6.3.39}$$

On the other hand, by Lemma 6.2.1, (6.3.20) and (6.3.30) we obtain

$$\mathbf{p} = \mathbf{q}_{p+1} - \sum_{i=1}^{d} \hat{\mathbf{r}}_i \in \bar{\mathcal{P}}_p{}^{\perp}, \tag{6.3.40}$$

Combining (6.3.39) and (6.3.40) yields $\mathbf{p} = \mathbf{0}$ on Δ, which by (6.3.30) leads to

$$\mathbf{q}_{p+1}(\tau, \xi) = \sum_{i=1}^{d} \hat{\mathbf{r}}_i(\tau, \xi_i), \quad \boldsymbol{\xi} \in \Delta. \tag{6.3.41}$$

Substituting $\mathbf{q}_k = \mathbf{0},\ 0 \le k \le p$, and (6.3.41) into (6.3.17) yields (6.3.1a). This completes the proof. □

We note that the results of Corollary 4.2.2 and the Theorem 4.3.1 on superconvergence do *not* extend to the DG Method with Lax-Friedrichs flux.

6.4 *A Posteriori* Error Estimation for Lax-Friedrichs Flux for Symmetric Systems

In this section we present an *a posteriori* error estimation procedure which consists of computing asymptotically exact local and global error estimates of the DG error. In Theorem 6.3.1 we showed that the local discretization error for the DG method on a physical element $\omega = (0, h)^d$ can be written as

$$\mathbf{e}(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} L_{p+1}(\xi_i)\mathbf{c}_i(t) - L_p(\xi_i)(\mathbf{A}_i^{(\mathrm{mod}(p,2))}\mathbf{c}_i(t) + \mathbf{d}_i(t)) + \mathcal{O}(h^{p+2}), \tag{6.4.1}$$

where

$$\mathbf{d}_i(t) \in \mathcal{N}(\mathbf{A}_i) \cap \bigoplus_{k=1}^{d} \mathcal{R}(\mathbf{A}_k), \quad 1 \leq i \leq d. \tag{6.4.2}$$

We apply the pseudoinverse $\mathbf{A}_i^\dagger$ of $\mathbf{A}_i$ to split $\mathbf{c}_i$ into

$$\mathbf{c}_i = \mathbf{c}_i^\perp + \mathbf{c}_i^\Phi, \text{ where } \mathbf{c}_i^\perp = \mathbf{A}_i^\dagger \mathbf{A}_i \mathbf{c}_i, \quad 1 \leq i \leq d. \tag{6.4.3}$$

We note that by Lemma 1.2.13, $\mathbf{A}_i^\dagger \mathbf{A}_i$ is the projection onto $\mathcal{N}(\mathbf{A}_i)^\perp$, thus

$$\mathbf{c}_i^\perp = \mathbf{A}_i^\dagger \mathbf{A}_i \mathbf{c}_i \in \mathcal{N}(\mathbf{A}_i)^\perp, \ \mathbf{c}_i^\Phi \in \mathcal{N}(\mathbf{A}_i), \ 1 \leq i \leq d. \tag{6.4.4}$$

By (1.2.52c), $\mathbf{A}_i^{(\mathrm{mod}(p,2))} \mathbf{c}_i^\Phi = \mathbf{0}$, which yields $\mathbf{A}_i^{(\mathrm{mod}(p,2))} \mathbf{c}_i = \mathbf{A}_i^{(\mathrm{mod}(p,2))} \mathbf{c}_i^\perp \in \mathcal{R}(\mathbf{A}_i) = \mathcal{N}(\mathbf{A}_i)^\perp$.

Hence, the leading term of the spatial discretization error can be split into two parts as

$$\mathbf{e} = \mathbf{e}^\perp + \mathbf{e}^\Phi + \mathcal{O}(h^{p+2}), \tag{6.4.5}$$

where

$$\mathbf{e}^\perp(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} L_{p+1}(\xi_i) \mathbf{c}_i^\perp(t) - L_p(\xi_i) \mathbf{A}_i^{(\mathrm{mod}(p,2))} \mathbf{c}_i^\perp(t), \tag{6.4.6}$$

and

$$\mathbf{e}^\Phi(t, h\boldsymbol{\xi}) = h^{p+1} \sum_{i=1}^{d} L_{p+1}(\xi_i) \mathbf{c}_i^\Phi(t) - L_p(\xi_i) \mathbf{d}_i(t). \tag{6.4.7}$$

We note that for invertible matrices $\mathbf{A}_i, \ 1 \leq i \leq d$, the error component $\mathbf{e}^\Phi(t, \mathbf{x})$ is zero.

Next, we develop an *a posteriori* error estimation procedure for estimating both $\mathbf{e}^\perp$ and $\mathbf{e}^\Phi$.

6.4.1 The Stationary Component of the Error Estimate

The error estimation procedure and analysis in §4.3.3 for the stationary part of the *a posteriori* error analysis holds true for the Lax-Friedrichs flux splitting, if we replace $\mathrm{sgn}(\mathbf{A}_i)$ with $\mathbf{A}_i^{(\mathrm{mod}(p,2))}$, $1 \leq i \leq d$.

Theorem 6.4.1. *Under the assumptions of Theorem 6.3.1, let us consider the error estimate*

$$\mathbf{E}^\perp(t, h\boldsymbol{\xi}) = \sum_{i=1}^{d} \left(L_{p+1}(\xi_i) - L_p(\xi_i) \mathbf{A}_i^{(mod(p,2))} \right) \frac{h^{1-d}}{2} \mathbf{A}_i^\dagger \mathfrak{r}_{p,i}^\perp, \tag{6.4.8}$$

where $\mathfrak{r}_{p,i}^\perp, \ 1 \leq i \leq d$, *are defined in (4.3.24b) as*

$$\mathfrak{r}_{p,i}^\perp = \mathsf{P}_i \int_\omega L_p\left(\frac{x_i}{h}\right) \left(\mathbf{g} - \frac{\partial \mathbf{u}_h}{\partial t} - \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{u}_h}{\partial x_j} \right) d\mathbf{x}, \quad 1 \leq i \leq d, \tag{6.4.9}$$

with $\mathsf{P}_i = \mathbf{A}_i\mathbf{A}_i^\dagger$.

Then, for $p \geq 1$ *and* $t = \mathcal{O}(1)$,

$$\mathbf{e}^\perp(t,\mathbf{x}) = \mathbf{E}^\perp(t,\mathbf{x}) + \mathcal{O}(h^{p+2}), \quad \mathbf{x} \in \omega. \tag{6.4.10}$$

The proof of Theorem 6.4.1 is the same as for Theorem 4.3.2, if we replace $\operatorname{sgn}(\mathbf{A}_i)$ with $\mathbf{A}_i^{(\operatorname{mod}(p,2))}$, $1 \leq i \leq d$, and is therefore omitted.

6.4.2 The Transient Component of the Error Estimate

The *a posteriori* error estimation procedure in §4.3.4 to compute estimates for $\mathbf{e}^\Phi$ holds for the Lax-Friedrichs flux splitting, if we replace π, π_i^s, $\mathbf{A}_i^s$ and $\operatorname{sgn}(\mathbf{A}_i)$, respectively, by $\breve{\pi}$, $\breve{\pi}_i^s$, $\breve{\mathbf{A}}_i^s$ and $\mathbf{A}_i^{(\operatorname{mod}(p,2))}$, $s = +,-$, $1 \leq i \leq d$. We will state it below for clarity.

By Lemma 6.1.2, the approximations $\breve{\pi}\mathbf{u}_0$ on ω and $\breve{\pi}_i\mathbf{u}$ on the boundary $\partial\omega$ satisfy

$$\begin{aligned}
\mathbf{e}(0,\mathbf{x}) &= \mathbf{u}_0(\mathbf{x}) - \breve{\pi}\mathbf{u}_0(\mathbf{x}) \\
&= h^{p+1}\sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j(0) - L_p\left(\frac{x_j}{h}\right)\mathbf{A}_j^{(\operatorname{mod}(p,2))}\mathbf{c}_j(0) + \mathcal{O}(h^{p+2}),\ \mathbf{x} \in \omega, \qquad (6.4.11)\\
\mathbf{e}^-(t,\mathbf{x}) &= \mathbf{u}(t,\mathbf{x}) - \breve{\pi}_i^s\mathbf{u}(t,\mathbf{x}) \\
&= h^{p+1}\sum_{j\in D(i)} L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j(t) - L_p\left(\frac{x_j}{h}\right)\mathbf{A}_j^{(\operatorname{mod}(p,2))}\mathbf{c}_j(t) \\
&\quad - h^{p+1}\left(L_{p+1}\left(1-\frac{x_i}{h}\right)\mathbf{c}_i(t) - L_p\left(1-\frac{x_i}{h}\right)\mathbf{A}_i^{(\operatorname{mod}(p,2))}\mathbf{c}_i(t)\right) + \mathcal{O}(h^{p+2}), \\
&\qquad \mathbf{x} \in \gamma_i,\ s = +,-,\ 1 \leq i \leq d. \qquad (6.4.12)
\end{aligned}$$

We split the error at $t = 0$ into $\mathbf{e} = \mathbf{e}^\perp + \mathbf{e}^\Phi + \mathcal{O}(h^{p+2})$ as in (6.4.5) and define $\mathbf{E}^\Phi(0,\mathbf{x})$ by

$$\mathbf{E}^\Phi(0,\mathbf{x}) = \mathbf{e}^\Phi(0,\mathbf{x}) = h^{p+1}\sum_{i=1}^{d} L_{p+1}\left(\frac{x_i}{h}\right)\mathbf{c}_i^\Phi(0),\ \mathbf{c}_i^\Phi(0) = (\mathbf{I}-\mathsf{P}_i)\mathbf{c}_i(0) \in \mathcal{N}(\mathbf{A}_i), \tag{6.4.13}$$

where $\mathsf{P}_i = \mathbf{A}_i\mathbf{A}_i^\dagger$ denotes the projection from $\mathbb{R}^m$ into $\mathcal{N}(\mathbf{A}_i)^\perp$.

On the boundary, we define $\mathbf{E}^-$ by the leading term of (6.4.12),

$$\begin{aligned}
\mathbf{E}^-(t,\mathbf{x}) &= h^{p+1}\sum_{j\in D(i)} L_{p+1}\left(\frac{x_j}{h}\right)\mathbf{c}_j(t) - L_p\left(\frac{x_j}{h}\right)\mathbf{A}_j^{(\operatorname{mod}(p,2))}\mathbf{c}_j(t) \qquad (6.4.14)\\
&\quad - h^{p+1}\left(L_{p+1}\left(1-\frac{x_i}{h}\right)\mathbf{c}_i(t) - L_p\left(1-\frac{x_i}{h}\right)\mathbf{A}_i^{(\operatorname{mod}(p,2))}\mathbf{c}_i(t)\right),\ \mathbf{x} \in \gamma_i,\ 1 \leq i \leq d,
\end{aligned}$$

where $\left(1-\frac{x_i}{h}\right) = 1$ on γ_i^- and 0 on γ_i^+.

Now let us approximate $\mathbf{e}^{\mathbf{\Phi}}$ by determining

$$\mathbf{E}^{\mathbf{\Phi}}(t,\mathbf{x}) = \sum_{j=1}^{d} L_{p+1}\left(\frac{x_j}{h}\right)\boldsymbol{\gamma}_j^{\mathbf{\Phi}}(t) - L_p\left(\frac{x_j}{h}\right)\boldsymbol{\delta}_j^{\mathbf{\Phi}}(t), \quad \boldsymbol{\gamma}_j^{\mathbf{\Phi}}, \boldsymbol{\delta}_j^{\mathbf{\Phi}} \in \mathcal{N}(\mathbf{A}_j),\ 1 \le j \le d, \tag{6.4.15a}$$

such that

$$\begin{aligned}&\int_\omega \mathbf{v}^t \left(\frac{\partial(\mathbf{u}_h + \mathbf{E}^{\mathbf{\Phi}})}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{u}_h}{\partial x_j} - \mathbf{g}\right) d\mathbf{x} \\ &= \sum_{j=1}^{d} \int_{\gamma_j} \mathbf{v}^t \nu_j \breve{\mathbf{A}}_j^{\bar{\mu}_j} \left(\mathbf{u}_h + \mathbf{E}^{\perp} + \mathbf{E}^{\mathbf{\Phi}} - \mathbf{u}_h^- - \mathbf{E}^-\right) d\mathbf{s},\ \forall\ \mathbf{v} \in \mathcal{E}_p,\end{aligned} \tag{6.4.15b}$$

where $\mathbf{E}^{\perp}$ is the stationary component defined by (6.4.8) and $\mathcal{E}_p$ is defined in (4.3.41c).

By Lemma 1.2.14, $(\mathbf{I} - \mathsf{P}_i)$ projects any vector in $\mathbb{R}^m$ into $\mathcal{N}(\mathbf{A}_i)$ and the columns of $(\mathbf{I} - \mathsf{P}_i)$ span $\mathcal{N}(\mathbf{A}_i)$. Hence the columns of $L_{p+1}(\xi_i)(\mathbf{I} - \mathsf{P}_i)$ and $L_p(\xi_i)(\mathbf{I} - \mathsf{P}_i),\ 1 \le i \le d$, span $\mathcal{E}_p$.

Testing (6.4.15b) against $\mathbf{v} = L_m(\xi_i)(\mathbf{I} - \mathsf{P}_i),\ m = p, p+1,\ 1 \le i \le d$, yields

$$\begin{aligned}&\int_\omega L_m\left(\frac{x_i}{h}\right)(\mathbf{I} - \mathsf{P}_i)\left(\frac{\partial(\mathbf{u}_h + \mathbf{E}^{\mathbf{\Phi}})}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{u}_h}{\partial x_j} - \mathbf{g}\right) d\mathbf{x} \\ &= \sum_{j=1}^{d} \int_{\gamma_j} L_m\left(\frac{x_i}{h}\right)(\mathbf{I} - \mathsf{P}_i)\nu_j \breve{\mathbf{A}}_j^{\bar{\mu}_j} \left(\mathbf{u}_h + \mathbf{E}^{\perp} + \mathbf{E}^{\mathbf{\Phi}} - \mathbf{u}_h^- - \mathbf{E}^-\right) d\mathbf{s}, \\ &\forall\ m = p, p+1,\ 1 \le i \le d.\end{aligned} \tag{6.4.16}$$

Then (6.4.16) can be written as

$$\int_\omega L_m\left(\frac{x_i}{h}\right)(\mathbf{I} - \mathsf{P}_i)\frac{\partial \mathbf{E}^{\mathbf{\Phi}}}{\partial t}\, d\mathbf{x} - \sum_{j=1}^{d} \int_{\gamma_j} L_m\left(\frac{x_i}{h}\right)(\mathbf{I} - \mathsf{P}_i)\nu_j \breve{\mathbf{A}}_j^{\bar{\mu}_j} \mathbf{E}^{\mathbf{\Phi}}\, d\mathbf{s} = \mathfrak{r}_{m,i}^{\mathbf{\Phi}}, \tag{6.4.17a}$$

where $\mathfrak{r}_{m,i}^{\mathbf{\Phi}}$ is the projection of the residual onto $\mathcal{N}(\mathbf{A}_i)$, given by

$$\begin{aligned}\mathfrak{r}_{m,i}^{\mathbf{\Phi}} = (\mathbf{I} - \mathsf{P}_i)&\int_\omega L_m\left(\frac{x_i}{h}\right)\left(\mathbf{g} - \frac{\partial \mathbf{u}_h}{\partial t} - \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{u}_h}{\partial x_j}\right) d\mathbf{x} \\ &+ (\mathbf{I} - \mathsf{P}_i)\sum_{j=1}^{d} \int_{\gamma_j} L_m\left(\frac{x_i}{h}\right)\nu_j \breve{\mathbf{A}}_j^{\bar{\mu}_j} \left(\mathbf{u}_h + \mathbf{E}^{\perp} - \mathbf{u}_h^- - \mathbf{E}^-\right) d\mathbf{s}, \\ &m = p, p+1,\ 1 \le i \le d.\end{aligned} \tag{6.4.17b}$$

For $m = p+1$, we use the orthogonality properties (2.2.6) to reduce (6.4.17a) to

$$\int_\omega L_{p+1}^2\left(\frac{x_i}{h}\right)\dot{\boldsymbol{\gamma}}_i^{\mathbf{\Phi}}\, d\mathbf{x} - \sum_{j=1}^{d} \int_{\gamma_j} L_{p+1}^2\left(\frac{x_i}{h}\right)\nu_j(\mathbf{I} - \mathsf{P}_i)\breve{\mathbf{A}}_j^{\bar{\mu}_j} \boldsymbol{\gamma}_i^{\mathbf{\Phi}}\, d\mathbf{s} = \mathfrak{r}_{p+1,i}^{\mathbf{\Phi}}, \tag{6.4.18}$$

which, by (2.2.6), is equal to

$$\dot{\boldsymbol{\gamma}}_i^{\Phi} = \frac{1}{h}(\mathbf{I} - \mathsf{P}_i)\sum_{j=1}^{d}(\breve{\mathbf{A}}_j^- - \breve{\mathbf{A}}_j^+)\boldsymbol{\gamma}_i^{\Phi} + \frac{2p+3}{h^d}\mathfrak{r}_{p+1,i}^{\Phi}. \tag{6.4.19a}$$

For $m = p$, we get similarly

$$\dot{\boldsymbol{\delta}}_i^{\Phi} = \frac{1}{h}(\mathbf{I} - \mathsf{P}_i)\sum_{j=1}^{d}(\breve{\mathbf{A}}_j^- - \breve{\mathbf{A}}_j^+)\boldsymbol{\delta}_i^{\Phi} + \frac{2p+1}{h^d}\mathfrak{r}_{p,i}^{\Phi}, \tag{6.4.19b}$$

subject to the initial conditions

$$\boldsymbol{\gamma}_i^{\Phi}(0) = h^{p+1}\mathbf{c}_i^{\Phi}(0),\ \boldsymbol{\delta}_i^{\Phi}(0) = \mathbf{0}. \tag{6.4.19c}$$

It remains to state and prove the asymptotic exactness of $\mathbf{E}^{\Phi}$.

In the next Lemma, we will prove an orthogonality on $\mathcal{E}_p$, which simplifies the proof of Theorem 6.4.3.

Lemma 6.4.2. *Let $\mathbf{A}_i \neq \mathbf{0}$ for at least one $1 \leq i \leq d$. If $\mathbf{q} \in \mathcal{E}_p$ satisfies the orthogonality condition*

$$\sum_{i=1}^{d}\left(\int_{\Delta}\mathbf{v}^t\mathbf{A}_i\frac{\partial\mathbf{q}}{\partial\xi_i}\,d\boldsymbol{\xi} - \int_{\Gamma_i}\mathbf{v}^t\nu_i\breve{\mathbf{A}}_i^{\bar{\mu}_i}\mathbf{q}\,d\boldsymbol{\sigma}\right) = 0,\ \forall\ \mathbf{v} \in \mathcal{E}_p, \tag{6.4.20}$$

then $\mathbf{q} = \mathbf{0}$.

Proof. First we integrate equation (6.4.20) by parts to write

$$\sum_{j=1}^{d}\left(-\int_{\Delta}\frac{\partial\mathbf{v}^t}{\partial\xi_j}\mathbf{A}_j\mathbf{q}\,d\boldsymbol{\xi} + \int_{\Gamma_j}\mathbf{v}^t\nu_j\breve{\mathbf{A}}_j^{\mu_j}\mathbf{q}\,d\boldsymbol{\sigma}\right) = 0,\ \forall\ \mathbf{v} \in \mathcal{E}_p, \tag{6.4.21}$$

Adding (6.4.20) and (6.4.21) and setting $\mathbf{v} = \mathbf{q}$, the integral on Δ vanishes because of the symmetry of $\mathbf{A}_j$, $1 \leq j \leq d$, and we get

$$\sum_{j=1}^{d}\int_{\Gamma_j}\mathbf{q}^t(\breve{\mathbf{A}}_j^+ - \breve{\mathbf{A}}_j^-)\mathbf{q}\,d\boldsymbol{\sigma} = 0. \tag{6.4.22}$$

Since $(\breve{\mathbf{A}}_j^+ - \breve{\mathbf{A}}_j^-)$ is positive semi-definite by (1.2.52g), there exists a matrix $\mathbf{L}_j$ such that $\mathbf{L}_j^t\mathbf{L}_j = (\breve{\mathbf{A}}_j^+ - \breve{\mathbf{A}}_j^-)$, and (6.4.22) yields

$$\sum_{j=1}^{d}\int_{\Gamma_j}\|\mathbf{L}_j\mathbf{q}\|^2\,d\boldsymbol{\sigma} = 0, \tag{6.4.23}$$

which yields

$$\mathbf{L}_j\mathbf{q} = \mathbf{0} \text{ on } \Gamma_j,\ 1 \le j \le d. \tag{6.4.24}$$

Since $\mathbf{A}_i \neq 0$, combining (1.2.52b) and the Cholesky factorization for $\breve{\mathbf{A}}_i^+ - \breve{\mathbf{A}}_i^-$, we write

$$\mathbf{L}_i^t\mathbf{L}_i = \breve{\mathbf{A}}_i^+ - \breve{\mathbf{A}}_i^- = C_i\mathbf{I},\ C_i \neq 0. \tag{6.4.25}$$

Pre-multiplying (6.4.24) by $\mathbf{L}_i^t$ yields

$$\mathbf{L}_i^t\mathbf{L}_i\mathbf{q} = C_i\mathbf{q} = \mathbf{0} \text{ on } \Gamma_i. \tag{6.4.26}$$

Therefore, by (4.3.41c), every $\mathbf{q} \in \mathcal{E}_p$ can be written as

$$\mathbf{q}(\boldsymbol{\xi}) = \sum_{j=1}^{d} (L_{p+1}(\xi_j)\mathbf{a}_j - L_p(\xi_j)\mathbf{b}_j) = \mathbf{0} \text{ on } \Gamma_i. \tag{6.4.27}$$

We note that $L_{p+1}(\xi_j)\mathbf{a}_j$, $L_p(\xi_j)\mathbf{b}_j$, $j \in D(i)$, and the constant vector $L_{p+1}(\xi_i)\mathbf{a}_i + L_p(\xi_i)\mathbf{b}_i$ are pairwise orthogonal, and thus linearly independent, on Γ_i^+ and Γ_i^-, therefore

$$\mathbf{a}_j = \mathbf{b}_j = \mathbf{0},\ j \in D(i), \tag{6.4.28a}$$
$$L_{p+1}(0)\mathbf{a}_i - L_p(0)\mathbf{b}_i = (-1)^{p+1}(\mathbf{a}_i + \mathbf{b}_i) = \mathbf{0}, \tag{6.4.28b}$$
$$L_{p+1}(1)\mathbf{a}_i - L_p(1)\mathbf{b}_i = \mathbf{a}_i - \mathbf{b}_i = \mathbf{0}, \tag{6.4.28c}$$

which yields $\mathbf{q} = \mathbf{0}$ on Δ. □

Theorem 6.4.3. *Under the assumptions of Theorem 6.3.1, assume further that $\mathbf{u}_h$ is computed by approximating the initial conditions by $\breve{\pi}\mathbf{u}_0$ and let*

$$\mathbf{E}^{\Phi}(t, h\boldsymbol{\xi}) = \sum_{j=1}^{d} (L_{p+1}(\xi_j)\boldsymbol{\gamma}_j^{\Phi}(t) - L_p(\xi_j)\boldsymbol{\delta}_j^{\Phi}(t)), \tag{6.4.29}$$

where $\boldsymbol{\gamma}_i^{\Phi}, \boldsymbol{\delta}_i^{\Phi},\ 1 \le i \le d$, are solutions of (6.4.19) and (6.4.17b).
Then, at $t = \mathcal{O}(1)$ and for $p \ge 1$,

$$\mathbf{e}^{\Phi}(t, \mathbf{x}) = \mathbf{E}^{\Phi}(t, \mathbf{x}) + \mathcal{O}(h^{p+2}), \quad \mathbf{x} \in \omega. \tag{6.4.30}$$

Proof. Since the true solution $\mathbf{u}$ is continuous and $\mathbf{u} = \mathbf{u}^-$ on $\partial\omega$, $\mathbf{u}$ satisfies

$$\int_\omega \mathbf{v}^t \left(\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{u}}{\partial x_j} - \mathbf{g} \right) d\mathbf{x} = \sum_{j=1}^{d} \int_{\gamma_j} \mathbf{v}^t \nu_j \breve{\mathbf{A}}_j^{\bar{\mu}_j} \left(\mathbf{u} - \mathbf{u}^- \right) d\mathbf{s},\ \forall\ \mathbf{v} \in \mathcal{E}_p. \tag{6.4.31}$$

Subtracting (6.4.15b) from (6.4.31) gives

$$\int_\omega \mathbf{v}^t \left(\frac{\partial(\mathbf{e} - \mathbf{E}^{\mathbf{\Psi}})}{\partial t} + \sum_{j=1}^{d} \mathbf{A}_j \frac{\partial \mathbf{e}}{\partial x_j} \right) d\mathbf{x}$$
$$= \sum_{j=1}^{d} \int_{\gamma_j} \mathbf{v}^t \nu_j \breve{\mathbf{A}}_j^{\bar{\mu}_j} \left(\mathbf{e} - \mathbf{E}^{\perp} - \mathbf{E}^{\mathbf{\Psi}} - \mathbf{e}^{-} + \mathbf{E}^{-} \right) \right) d\mathbf{s}, \ \forall\ \mathbf{v} \in \mathcal{E}_p. \tag{6.4.32}$$

Since $\mathbf{v} \in \mathcal{E}_p$, we can write

$$\mathbf{v}(\mathbf{x}) = \sum_{i=1}^{d} L_{p+1}\left(\frac{x_i}{h}\right) \mathbf{a}_i - L_p\left(\frac{x_i}{h}\right) \mathbf{b}_i, \quad \mathbf{a}_i, \mathbf{b}_i \in \mathcal{N}(\mathbf{A}_i), \tag{6.4.33}$$

while $\mathbf{E}^{\perp}$ is defined in (6.4.8) as

$$\mathbf{E}^{\perp}(t, h\boldsymbol{\xi}) = \sum_{j=1}^{d} L_{p+1}(\xi_j) \boldsymbol{\gamma}_j^{\perp}(t) - L_p(\xi_j) \mathbf{A}_j^{(\mathrm{mod}(p,2))} \boldsymbol{\gamma}_j^{\perp}(t), \quad \boldsymbol{\gamma}_j^{\perp} \in \mathcal{N}(\mathbf{A}_j)^{\perp}. \tag{6.4.34}$$

By (1.2.52c) and (1.2.12), $\mathbf{A}_j^{(\mathrm{mod}(p,2))} \dot{\boldsymbol{\gamma}}_j^{\perp} \in \mathcal{R}(\mathbf{A}_j) = \mathcal{N}(\mathbf{A}_j)^{\perp}$, which, combined with (4.3.64), yields

$$\langle \mathbf{a}_i, \dot{\boldsymbol{\gamma}}_i^{\perp}(t) \rangle = 0, \ \langle \mathbf{b}_i, \mathbf{A}_j^{(\mathrm{mod}(p,2))} \dot{\boldsymbol{\gamma}}_j^{\perp}(t) \rangle = 0, \quad 1 \le i \le d. \tag{6.4.35}$$

By substituting $\mathbf{v}$ and $\mathbf{E}^{\perp}$, as defined in (6.4.33) and (6.4.34), into $\int_\omega \mathbf{v}^t \frac{\partial \mathbf{E}^{\perp}}{\partial t}\, d\mathbf{x}$ and applying the orthogonality property (2.2.6), we obtain

$$\int_\omega \mathbf{v}^t \frac{\partial \mathbf{E}^{\perp}}{\partial t}\, d\mathbf{x} = \sum_{i=1}^{d} \sum_{j=1}^{d} \int_\omega \left(L_{p+1}\left(\frac{x_j}{h}\right) \mathbf{a}_i - L_p\left(\frac{x_j}{h}\right) \mathbf{b}_i \right)^t$$
$$\left(L_{p+1}\left(\frac{x_i}{h}\right) \dot{\boldsymbol{\gamma}}_i^{\perp}(t) - L_p\left(\frac{x_i}{h}\right) \mathbf{A}_i^{(\mathrm{mod}(p,2))} \dot{\boldsymbol{\gamma}}_i^{\perp}(t) \right) d\mathbf{x}$$
$$= \sum_{i=1}^{d} \int_\omega L_{p+1}^2\left(\frac{x_i}{h}\right) \langle \mathbf{a}_i, \dot{\boldsymbol{\gamma}}_i^{\perp}(t) \rangle + L_p^2\left(\frac{x_i}{h}\right) \langle \mathbf{b}_i, \mathbf{A}_j^{(\mathrm{mod}(p,2))} \boldsymbol{\gamma}_j^{\perp}(t) \rangle\, d\mathbf{x}$$
$$= 0, \quad \forall\ \mathbf{v} \in \mathcal{E}_p. \tag{6.4.36}$$

Furthermore, by substituting $\mathbf{v}$, $\mathbf{E}^{\perp}$ and $\mathbf{E}^{\mathbf{\Psi}}$, as defined in (6.4.33), (6.4.34) and (6.4.15a), into $\int_\omega \mathbf{v}^t \mathbf{A}_i \frac{\partial(\mathbf{E}^{\perp} + \mathbf{E}^{\mathbf{\Psi}})}{\partial x_i}\, d\mathbf{x}$ and applying the orthogonality property (2.2.6), we obtain

$$\int_\omega \mathbf{v}^t \mathbf{A}_i \frac{\partial(\mathbf{E}^{\perp} + \mathbf{E}^{\mathbf{\Psi}})}{\partial x_i}\, d\mathbf{x} = \frac{1}{h} \sum_{j=1}^{d} \int_\omega \left(L_{p+1}\left(\frac{x_j}{h}\right) \mathbf{a}_i - L_p\left(\frac{x_j}{h}\right) \mathbf{b}_i \right)^t \mathbf{A}_i$$
$$\left(L'_{p+1}\left(\frac{x_i}{h}\right) (\boldsymbol{\gamma}_i^{\perp} + \boldsymbol{\gamma}_i^{\mathbf{\Psi}}) - L'_p\left(\frac{x_i}{h}\right) (\mathbf{A}_i^{(\mathrm{mod}(p,2))} \boldsymbol{\gamma}_i^{\perp} + \boldsymbol{\delta}_i^{\mathbf{\Psi}}) \right) d\mathbf{x}$$
$$= \frac{1}{h} \int_\omega L_p\left(\frac{x_i}{h}\right) L'_{p+1}\left(\frac{x_i}{h}\right) (\mathbf{A}_i \mathbf{b}_i)^t (\boldsymbol{\gamma}_i^{\perp} + \boldsymbol{\gamma}_i^{\mathbf{\Psi}})\, d\mathbf{x}$$
$$= 0, \quad \forall\ \mathbf{v} \in \mathcal{E}_p,\ 1 \le i \le d, \tag{6.4.37}$$

where we used the fact that $\mathbf{b}_i \in \mathcal{N}(\mathbf{A}_i)$.

Subtracting (6.4.36) and (6.4.37) from (6.4.32) yields for $\boldsymbol{\epsilon} = \mathbf{e} - \mathbf{E}^{\perp} - \mathbf{E}^{\Phi}$ and $\boldsymbol{\epsilon}^- = \mathbf{e}^- - \mathbf{E}^-$

$$\int_\omega \mathbf{v}^t \left(\frac{\partial \boldsymbol{\epsilon}}{\partial t} + \sum_{j=1}^d \mathbf{A}_j \frac{\partial \boldsymbol{\epsilon}}{\partial x_j} \right) d\mathbf{x} = \sum_{j=1}^d \int_{\gamma_j} \mathbf{v}^t \nu_j \breve{\mathbf{A}}_j^{\bar{\mu}_j} (\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^-)\, d\mathbf{s},\ \forall\ \mathbf{v} \in \mathcal{E}_p. \tag{6.4.38}$$

By (6.4.5) we can write

$$\boldsymbol{\epsilon} = (\mathbf{e}^{\perp} - \mathbf{E}^{\perp}) + (\mathbf{e}^{\Phi} - \mathbf{E}^{\Phi}) + \mathcal{O}(h^{p+2}), \tag{6.4.39}$$

thus, since $\mathbf{e}^{\perp} - \mathbf{E}^{\perp} = \mathcal{O}(h^{p+2})$ by Theorem 6.4.1, we obtain

$$\boldsymbol{\epsilon} = (\mathbf{e}^{\Phi} - \mathbf{E}^{\Phi}) + \mathcal{O}(h^{p+2}). \tag{6.4.40}$$

We will now show that $\boldsymbol{\epsilon} = \mathcal{O}(h^{p+2})$.

Applying the linear transformations $t = T\tau,\ T > 0$, and $\mathbf{x} = h\boldsymbol{\xi}$, (6.4.38) becomes

$$\int_\Delta \mathbf{v}^t \left(\frac{h}{T} \frac{\partial \hat{\boldsymbol{\epsilon}}}{\partial \tau} + \sum_{j=1}^d \mathbf{A}_j \frac{\partial \hat{\boldsymbol{\epsilon}}}{\partial \xi_j} \right) d\boldsymbol{\xi} = \sum_{j=1}^d \int_{\Gamma_j} \mathbf{v}^t \nu_j \breve{\mathbf{A}}_j^{\bar{\mu}_j} (\hat{\boldsymbol{\epsilon}} - \hat{\boldsymbol{\epsilon}}^-)\, d\boldsymbol{\sigma},\ \forall\ \mathbf{v} \in \mathcal{E}_p, \tag{6.4.41}$$

where $\hat{\boldsymbol{\epsilon}}(\tau, \boldsymbol{\xi}) = \boldsymbol{\epsilon}(T\tau, h\boldsymbol{\xi})$.

The Maclaurin series of $\mathbf{e}^{\Phi} - \mathbf{E}^{\Phi} \in \mathcal{E}_p$ with respect to h is

$$(\mathbf{e}^{\Phi} - \mathbf{E}^{\Phi})(t, h\boldsymbol{\xi}) = \sum_{k=0}^{\infty} h^k \mathbf{q}_k(t, \boldsymbol{\xi}),\ \mathbf{q}_k \in \mathcal{E}_p,\ k \geq 0, \tag{6.4.42}$$

which together with (6.4.40) yields

$$\hat{\boldsymbol{\epsilon}}(t, \boldsymbol{\xi}) = \sum_{k=0}^{p+1} h^k \mathbf{q}_k(t, \boldsymbol{\xi}) + \mathcal{O}(h^{p+2}). \tag{6.4.43}$$

By the definition of $\mathbf{e}^-$ and $\mathbf{E}^-$ in (6.4.12) and (6.4.14),

$$\hat{\boldsymbol{\epsilon}}^-(\tau, \boldsymbol{\xi}) = (\mathbf{e}^- - \mathbf{E}^-)(T\tau, h\boldsymbol{\xi}) = \mathcal{O}(h^{p+2}),\ \boldsymbol{\xi} \in \Delta. \tag{6.4.44}$$

Substituting (6.4.43) and (6.4.44) into (6.4.38) yields

$$\sum_{k=0}^{p+1} h^k \left(\int_\Delta \mathbf{v}^t \left(\frac{h}{T} \frac{\partial \mathbf{q}_k}{\partial \tau} + \sum_{j=1}^d \mathbf{A}_j \frac{\partial \mathbf{q}_k}{\partial \xi_j} \right) d\boldsymbol{\xi} - \sum_{j=1}^d \int_{\Gamma_j} \mathbf{v}^t \nu_j \breve{\mathbf{A}}_j^{\bar{\mu}_j} \mathbf{q}_k\, d\boldsymbol{\sigma} \right) = \mathcal{O}(h^{p+2}),\ \forall\ \mathbf{v} \in \mathcal{E}_p, \tag{6.4.45}$$

which infers that all terms having the same power in h are zero.

Thus, the $\mathcal{O}(1)$ term leads to the orthogonality condition for $\mathbf{q}_0$,

$$\sum_{j=1}^{d}\left(\int_{\Delta}\mathbf{v}^t\mathbf{A}_j\frac{\partial\mathbf{q}_0}{\partial\xi_j}\,d\boldsymbol{\xi}-\int_{\Gamma_j}\mathbf{v}^t\nu_j\breve{\mathbf{A}}_j^{\bar{\mu}_j}\mathbf{q}_0\,d\boldsymbol{\sigma}\right)=0,\ \forall\ \mathbf{v}\in\mathcal{E}_p. \tag{6.4.46}$$

Since $\mathbf{q}_0\in\mathcal{E}_p$ by (6.4.42) and satisfies (6.4.46), Lemma 6.4.2 infers $\mathbf{q}_0=\mathbf{0}$.

Using induction, we assume that $\mathbf{q}_l=\mathbf{0},\ 0\le l\le k-1,\ k\le p+1$, and apply the $\mathcal{O}(h^k)$ term to obtain the orthogonality condition

$$\sum_{j=1}^{d}\left(\int_{\Delta}\mathbf{v}^t\mathbf{A}_j\frac{\partial\mathbf{q}_k}{\partial\xi_j}\,d\boldsymbol{\xi}-\int_{\Gamma_j}\mathbf{v}^t\nu_j\breve{\mathbf{A}}_j^{\bar{\mu}_j}\mathbf{q}_k\,d\boldsymbol{\sigma}\right)=0,\ \forall\ \mathbf{v}\in\mathcal{E}_p, \tag{6.4.47}$$

which, by Lemma 6.4.2 and (6.4.42), infers $\mathbf{q}_k=\mathbf{0},\ k\le p+1$.

Substituting $\mathbf{q}_k=\mathbf{0},\ 0\le k\le p+1$ into (6.4.43) yields $\hat{\boldsymbol{\epsilon}}=\mathcal{O}(h^{p+2})$, which, when substituted into (6.4.40), yields (6.4.30). This completes the proof. □

Note. *Now, we state an explicit formula to obtain the coefficients $\mathbf{c}_i$, $\mathbf{d}_i$ needed to evaluate the boundary approximation operator $\pi_i^s\mathbf{u}$ on γ_i^s, $s=+,-$, $1\le i\le d$. We set*

$$\mathbf{c}_i(t+\Delta t)=h^{-p-1}(\boldsymbol{\gamma}_i^{\perp}(t)+\boldsymbol{\gamma}_i^{\Phi}(t)),\quad \mathbf{d}_i(t+\Delta t)=h^{-p-1}\boldsymbol{\delta}_i^{\Phi}(t),\ 1\le i\le d, \tag{6.4.48}$$

where $\boldsymbol{\gamma}_i^{\perp}(t)$, $\boldsymbol{\gamma}_i^{\Phi}(t)$ and $\boldsymbol{\delta}_i^{\Phi}(t)$ are the coefficients of $\mathbf{E}^{\perp}(t)$ and $\mathbf{E}^{\Phi}(t)$, obtained in the previous timestep.

6.5 Computational Examples

Example 6.5.1. *Let us consider the two-dimensional wave equation, described in Example 3.3.4, which can be written as the first-order linear hyperbolic system*

$$\mathbf{u}_{,t}+\mathbf{A}_1\mathbf{u}_{,x}+\mathbf{A}_2\mathbf{u}_{,y}=0,\ (x,y)\in(0,1)^2,\ 0<t\le 1, \tag{6.5.1a}$$

where

$$\mathbf{u}=\begin{pmatrix}v_{,t}+v_{,x}\\ v_{,y}\end{pmatrix},\ \mathbf{A}_1=\begin{pmatrix}-1&0\\0&1\end{pmatrix},\ \mathbf{A}_2=\begin{pmatrix}0&-1\\-1&0\end{pmatrix}, \tag{6.5.1b}$$

and select initial and boundary conditions such that the true solution is

$$\mathbf{u}=\begin{pmatrix}\sin(\sqrt{2}t+x+y)-\cos(-\sqrt{2}t+x+y)\\(\sqrt{2}-1)\sin(\sqrt{2}t+x+y)+(1+\sqrt{2})\cos(-\sqrt{2}t+x+y))\end{pmatrix}. \tag{6.5.1c}$$

We solve (6.5.1) on uniform meshes having $N=5^2,10^2,15^2$ elements for $p=1,2,3$ using $\Pi\mathbf{u}_0$ and present the L^2 errors and effectivity indices corresponding to the stationary error estimates $\mathbf{E}^{\perp}$ at $t=1$ in Table 3.3.4. We observe that the effectivity indices converge to unity under mesh refinement, which is in full agreement with Theorem 6.4.1.

p	N	$\|\|\mathbf{e}\|\|$	*order*	$\|\|\mathbf{e}-\mathbf{E}^{\perp}\|\|$	*order*	θ
1	5^2	8.3265*e*-3	–	6.5837*e*-4	–	0.9769
	10^2	2.0569*e*-3	2.017	8.4485*e*-5	2.962	0.9889
	15^2	9.1068*e*-4	2.009	2.5224*e*-5	2.981	0.9927
2	5^2	8.4018*e*-5	–	1.5344*e*-5	–	1.025
	10^2	1.0593*e*-5	2.988	9.9241*e*-7	3.951	1.016
	15^2	3.1500*e*-6	2.991	1.9830*e*-7	3.972	1.011
3	5^2	1.6232*e*-6	–	1.4919*e*-7	–	0.9889
	10^2	1.0077*e*-7	4.01	4.7772*e*-9	4.965	0.9954
	15^2	1.9869*e*-8	4.005	6.3386*e*-10	4.981	0.9971

Table 6.5.1: L^2-errors $\|\mathbf{e}\|_{2,\Omega}$, $\|\mathbf{e}-\mathbf{E}^{\perp}\|_{2,\Omega}$ and their order of convergence. Global effectivity indices corresponding to static estimates for Example 6.5.1 at $t=1$ using $\Pi\mathbf{u}_0$ on Ω and $\mathbf{u}_h^- = \breve{\pi}\mathbf{u}_B$ on $\partial\Omega$.

Example 6.5.2. *Let* $\mathbf{u}$ *be defined on* $\mathbf{x}\in\Omega=(0,1)$ *and* $0\le t\le 1$ *by*

$$\frac{\partial \mathbf{u}}{\partial t}+\mathbf{A}_1\frac{\partial \mathbf{u}}{\partial x_1}+\mathbf{A}_2\frac{\partial \mathbf{u}}{\partial x_2}=\mathbf{g},\quad \mathbf{A}_1=\begin{pmatrix}0&0\\0&2\end{pmatrix},\ \mathbf{A}_2=\begin{pmatrix}1&-2\\-2&1\end{pmatrix}, \tag{6.5.2a}$$

with source term, initial and boundary conditions such that

$$\mathbf{u}(t,\mathbf{x})=\exp(t+x_1+x_2),\quad \mathbf{x}\in\Omega,\ 0\le t\le 1. \tag{6.5.2b}$$

Basic linear algebra yields that the eigenvector $(1,0)^t$ $\mathbf{A}_1$ corresponds to a zero eigenvalue. Thus, we only have theoretical results for odd orders of approximations p. Applying our theory, for odd p, the stationary error estimate $\mathbf{E}^b ot$ can accurately predict the second component of the error e_2, while we expect the transient error estimate $\mathbf{E}^{\perp}+\mathbf{E}^{\Pi}$ to estimate the full error $\mathbf{e}$.

To validate our theory, we solve (6.5.2) on uniform meshes having $N=5^2,10^2,15^2$ elements for $p=1,2,3$ using $\Pi\mathbf{u}_0$. We present the L^2-errors, componentwise L^2-errors and effectivity indices corresponding to the stationary error estimate $\mathbf{E}^{\perp}$ at $t=1$ in Tables 6.5.2 and 6.5.3. In Tables 6.5.4 and 6.5.5, we present the L^2-errors, componentwise L^2-errors and effectivity indices for the transient error estimate $\mathbf{E}^{\perp}+\mathbf{E}^{\Pi}$ at $t=1$. We observe that the effectivity indices for the transient error estimate and for the first component of the static estimate converge to unity under mesh refinement.

p	N	$\|\|\mathbf{e}\|\|$	*order*	$\|\|\mathbf{e}-\mathbf{E}^{\perp}\|\|$	*order*	θ
1	5^2	3.1963*e*-2	–	1.3148*e*-2	–	0.9048
	10^2	7.9760*e*-3	2.003	3.2487*e*-3	2.017	0.9093
	15^2	3.5418*e*-3	2.002	1.4407*e*-3	2.005	0.9107
2	5^2	1.0554*e*-3	–	4.9652*e*-4	–	1.084
	10^2	1.4049*e*-4	2.909	5.6190*e*-5	3.143	1.036
	15^2	4.2974*e*-5	2.921	1.6107*e*-5	3.082	1.01
3	5^2	6.7800*e*-6	–	3.2644*e*-6	–	0.8587
	10^2	4.1568*e*-7	4.028	1.8887*e*-7	4.111	0.8816
	15^2	8.1772*e*-8	4.01	3.6750*e*-8	4.037	0.887

Table 6.5.2: L^2-errors $\|\mathbf{e}\|_{2,\Omega}$, $\|\mathbf{e}-\mathbf{E}^{\perp}\|_{2,\Omega}$ and their order of convergence. Global effectivity indices corresponding to static estimates for Example 6.5.2 at $t = 1$ using $\Pi\mathbf{u}_0$ on Ω and $\mathbf{u}_h^- = \breve{\pi}\mathbf{u}_B$ on $\partial\Omega$.

p	N	$\|\|\mathbf{e}\|\|^*$	*order*	$\|\|\mathbf{e}-\mathbf{E}^{\perp}\|\|^*$	*order*	θ^*
1	5^2	1.9382*e*-2	–	1.3045*e*-2	–	0.7416
		2.5415*e*-2		1.6371*e*-3		0.9873
	10^2	4.8187*e*-3	2.0080	3.2419*e*-3	2.0086	0.7395
		6.3558*e*-3	1.9996	2.1124*e*-4	2.9542	0.9938
	15^2	2.1369*e*-3	2.0054	1.4394*e*-3	2.0025	0.7385
		2.8245*e*-3	2.0003	6.3025*e*-5	2.9829	0.9959
2	5^2	7.6744*e*-4	–	4.5458*e*-4	–	1.0112
		7.2455*e*-4		1.9973*e*-4		1.1610
	10^2	1.0183*e*-4	2.9139	5.4060*e*-5	3.0719	0.9692
		9.6796*e*-5	2.9041	1.5324*e*-5	3.7042	1.1055
	15^2	3.1102*e*-5	2.9251	1.5786*e*-5	3.0360	0.9465
		2.9656*e*-5	2.9175	3.2003*e*-6	3.8626	1.0764
3	5^2	4.3614*e*-6	–	3.1644*e*-6	–	0.6718
		5.1910*e*-6		8.0204*e*-7		0.9693
	10^2	2.6411*e*-7	4.0456	1.8721*e*-7	4.0792	0.6964
		3.2100*e*-7	4.0154	2.4992*e*-8	5.0041	0.9875
	15^2	5.1823*e*-8	4.0164	3.6598*e*-8	4.0256	0.7016
		6.3254*e*-8	4.0060	3.3366*e*-9	4.9662	0.9922

Table 6.5.3: Componentwise L^2-errors $\|\mathbf{e}\|^*$, $\|\mathbf{e}-\mathbf{E}^{\perp}\|^*$ and their order of convergence. Global effectivity indices corresponding to static estimates for Example 6.5.2 at $t = 1$ using $\Pi\mathbf{u}_0$ on Ω and $\mathbf{u}_h^- = \breve{\pi}\mathbf{u}_B$ on $\partial\Omega$.

p	N	$\|\|\mathbf{e}\|\|$	*order*	$\|\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Phi}\|\|$	*order*	θ
1	5^2	3.1963*e*-2	–	2.6009*e*-3	–	0.9932
	10^2	7.9760*e*-3	2.003	3.2740*e*-4	2.99	0.9986
	15^2	3.5418*e*-3	2.002	9.7195*e*-5	2.995	0.9998
2	5^2	1.0554*e*-3	–	4.7369*e*-4	–	1.091
	10^2	1.4049*e*-4	2.909	5.3183*e*-5	3.155	1.043
	15^2	4.2974*e*-5	2.921	1.5195*e*-5	3.09	1.018
3	5^2	6.7800*e*-6	–	1.8204*e*-6	–	0.9523
	10^2	4.1568*e*-7	4.028	8.0747*e*-8	4.495	0.9779
	15^2	8.1772*e*-8	4.01	1.4371*e*-8	4.257	0.9831

Table 6.5.4: L^2-errors $\|\mathbf{e}\|_{2,\Omega}$, $\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Phi}\|_{2,\Omega}$ and their order of convergence. Global effectivity indices corresponding to transient estimates for Example 6.5.2 at $t=1$ using $\Pi\mathbf{u}_0$ on Ω and $\mathbf{u}_h^- = \breve{\pi}\mathbf{u}_B$ on $\partial\Omega$.

p	N	$\|\|\mathbf{e}\|\|^*$	*order*	$\|\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Phi}\|\|^*$	*order*	θ^*
1	5^2	1.9382*e*-2	–	2.0210*e*-3	–	1.0034
		2.5415*e*-2		1.6371*e*-3		0.9873
	10^2	4.8187*e*-3	2.0080	2.5014*e*-4	3.0143	1.0069
		6.3558*e*-3	1.9996	2.1124*e*-4	2.9542	0.9938
	15^2	2.1369*e*-3	2.0054	7.3991*e*-5	3.0041	1.0065
		2.8245*e*-3	2.0003	6.3025*e*-5	2.9829	0.9959
2	5^2	7.6744*e*-4	–	4.2952*e*-4	–	1.0255
		7.2455*e*-4		1.9973*e*-4		1.1610
	10^2	1.0183*e*-4	2.9139	5.0927*e*-5	3.0762	0.9835
		9.6796*e*-5	2.9041	1.5324*e*-5	3.7042	1.1055
	15^2	3.1102*e*-5	2.9251	1.4854*e*-5	3.0387	0.9609
		2.9656*e*-5	2.9175	3.2003*e*-6	3.8626	1.0764
3	5^2	4.3614*e*-6	–	1.6341*e*-6	–	0.9278
		5.1910*e*-6		8.0204*e*-7		0.9693
	10^2	2.6411*e*-7	4.0456	7.6782*e*-8	4.4116	0.9635
		3.2100*e*-7	4.0154	2.4992*e*-8	5.0041	0.9875
	15^2	5.1823*e*-8	4.0164	1.3978*e*-8	4.2013	0.9694
		6.3254*e*-8	4.0060	3.3366*e*-9	4.9662	0.9922

Table 6.5.5: Componentwise L^2-errors $\|\mathbf{e}\|^*$, $\|\mathbf{e}-\mathbf{E}^{\perp}-\mathbf{E}^{\Phi}\|^*$ and their order of convergence. Global effectivity indices corresponding to transient estimates for Example 6.5.2 at $t=1$ using $\Pi\mathbf{u}_0$ on Ω and $\mathbf{u}_h^- = \breve{\pi}\mathbf{u}_B$ on $\partial\Omega$.

Chapter 7

A DG Adaptive Mesh Refinement Algorithm

In this chapter we are going to develop adaptive DG schemes that use the estimates of the discretization error, which we described in previous chapters, to guide the adaptive algorithm. We will first describe both an h- and a p-adaptive DG scheme. Then we will apply these schemes to a model problem which showing that our *a posteriori* error estimates successfully predict the local discretization error, even for irregular meshes. We finish the chapter with some applications of our error estimations to nonlinear problems.

Until now we required a uniform mesh size h and a uniform order p for all elements. However, since the smoothness of a solution, and therefore the approximation error, can vary largely within a given domain, we can reduce the computational cost by allowing different mesh size and/or order for each element. We can then lower the computational cost by using large elements and/or low order in smooth regions, and decrease the error by using small elements and/or high order in less smooth regions. This technique is called $h-$, $p-$, or $hp-$refinement, depending on which parameters, mesh size h or order p we allow to vary between elements. We will only consider the case $d = 2$.

We will aim to obtain a solution that satisfies

$$|\omega|^{-1/2}\|\mathbf{e}(t,\cdot)\|_{2,\omega} \lesssim tol \quad \forall\ \omega \in \mathcal{T}(t), 0 \le t \le T, \tag{7.0.1}$$

where $\mathcal{T}(t)$ denotes the partition used at time t and tol denotes a given tolerance.

To achieve this, we define, for fixed mesh size h and polynomial order p, a set $\Sigma^0 = (\mathcal{T}^0, \mathcal{V}^0, \mathcal{W}^0)$, where $\mathcal{T}^0 = \mathcal{T}_h$ denotes the initial partition and $\mathcal{V}^0, \mathcal{W}^0$ denote the piecewise polynomial finite element spaces defined as

$$\mathcal{V}^0 = \{\mathbf{v}(\mathbf{x}) :\ \mathbf{v}|_\omega \in \mathcal{P}_p,\ \omega \in \mathcal{T}^n\}, \tag{7.0.2}$$

$$\mathcal{W}^0 = \left\{ \mathbf{v}(\mathbf{x}(\boldsymbol{\xi})) = \sum_{i=1}^{2} (L_{p+1}(\xi_j)\mathbf{a}_j - L_p(\xi_j)\mathbf{b}_j) : \ \mathbf{a}_i, \mathbf{b}_i \in \mathbb{R}^m, \ \boldsymbol{\xi} \in \Delta \right\}. \tag{7.0.3}$$

We define initial approximations $\mathbf{U}(0,\cdot) \in \mathcal{V}^0$ and $\mathbf{E}(0,\cdot) \in \mathcal{W}^0$ by

$$\mathbf{U}(0,\mathbf{x}) = \pi \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \omega, \ \omega \in \mathcal{T}_h, \tag{7.0.4}$$

$$\mathbf{E}(0,\mathbf{x}) = \mathcal{L}^{\omega}_{p+1}\mathbf{u}_0(\mathbf{x}) - \pi^{\omega}_p \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \omega, \ \omega \in \mathcal{T}_h, \tag{7.0.5}$$

where $\pi^{\omega}_p \mathbf{u}_0$, $\mathcal{L}^{\omega}_{p+1}\mathbf{u}_0$ are interpolations defined in §2.2.

We refine and coarsen the partition at each time step t_n,

$$t_0 = 0 < t_1 < \cdots < t_N = T, \tag{7.0.6}$$

where we refine after each fifth time step of the Runge-Kutta time-stepping scheme.

We refine all elements $\omega \in \mathcal{T}^n$, $0 \le n \le N$, for which

$$|\omega|^{-1/2} \|\mathbf{E}(t^n,\cdot)\|_{2,\omega} \le tol. \tag{7.0.7}$$

In order to efficiently obtain a solution, we will choose $\theta \in \mathbb{R}$, for which we coarsen all elements $\omega \in \mathcal{T}^n$ that satisfy

$$|\omega|^{-1/2} \|\mathbf{E}(t^n,\cdot)\|_{2,\omega} \ge \theta tol. \tag{7.0.8}$$

We obtain a new set $\Sigma^{n+1} = (\mathcal{T}^{n+1}, \mathcal{V}^{n+1}, \mathcal{W}^{n+1})$, and replace $\mathbf{U}(t^n,\cdot) \in \mathcal{V}^n$ and $\mathbf{E}(t^n,\cdot) \in \mathcal{W}^n$ by functions $\mathbf{U}_{new}(t^n,\cdot) \in \mathcal{V}^{n+1}$ and $\mathbf{E}_{new}(t^n,\cdot) \in \mathcal{W}^{n+1}$.

In §7.1, we will describe an h-adaptive mesh refinement process. First we will describe how to obtain a new set Σ^{n+1} and functions $\mathbf{U}_{new}(t^n,\cdot)$, and $\mathbf{E}_{new}(t^n,\cdot)$ from Σ^n, $\mathbf{U}(t^n,\cdot)$, and $\mathbf{E}(t^n,\cdot)$. Then we will describe the criterium we use to coarsen an element ω. Finally, we describe how to integrate $\mathbf{U}(t,\cdot)$, $\mathbf{E}(t,\cdot)$ in time from $t = t^{n-1}$ to t^n.

In §7.2, we will describe the main features for a p-adaptive mesh refinement process.

Below we provide an algorithm description of the method.

Input: Set Σ^0, functions $\mathbf{U}(0,\cdot)$, $\mathbf{E}(0,\cdot)$, final time T, tolerance *tol*
foreach $0 \leq n \leq N$ **do**
  foreach *element* $\omega \in \mathcal{T}^n$ **do**
    if $|\omega|^{-1/2}\|\mathbf{E}(t^n,\cdot)\|_{2,\omega} > tol$ **then**
      Mark ω for refinement
    end
    if *Criterium for coarsening* **then**
      Mark ω for coarsening
    end
  end
  $(\Sigma^{n+1}, \mathbf{U}(t^n,\cdot), \mathbf{E}(t^n,\cdot))$ = Refine-and-Coarsen $(\Sigma^n, \mathbf{U}(t^n,\cdot), \mathbf{E}(t^n,\cdot)$, tol)
  $(\mathbf{U}(t^{n+1},\cdot),\mathbf{E}(t^{n+1},\cdot))$ = Time-integrate $(\mathbf{U}(t^n,\cdot), \mathbf{E}(t^n,\cdot), \Sigma^{n+1})$
end
Output: Set Σ^N, functions $\mathbf{U}(T,\cdot)$, $\mathbf{E}(T,\cdot)$

Algorithm 1: Basic time-stepping refinement algorithm

7.1 An h-Adaptive Mesh Refinement Algorithm

For simplicity, we use only two methods to h-refine and coarsen any given mesh $\mathcal{T}^n$: To refine an element ω, we will split it into four elements $\omega_i,\ 1 \leq i \leq 4$ as shown in Figure 7.1.1. To coarsen four elements $\omega_i,\ 1 \leq i \leq 4$ of the same size h that form a square of size $2h$, we will merge them into one element ω, as shown in Figure 7.1.2.

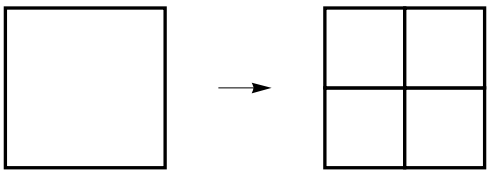

Figure 7.1.1: Refining of one element into four elements

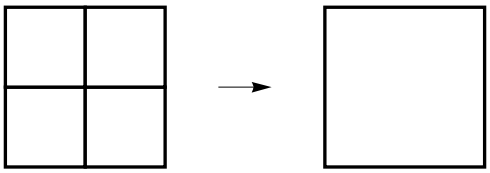

Figure 7.1.2: Coarsening of four elements into one element

This refinement process can create irregular nodes. However, we can create a partition allowing at most one irregular node on each face by enforcing the following two definitions:

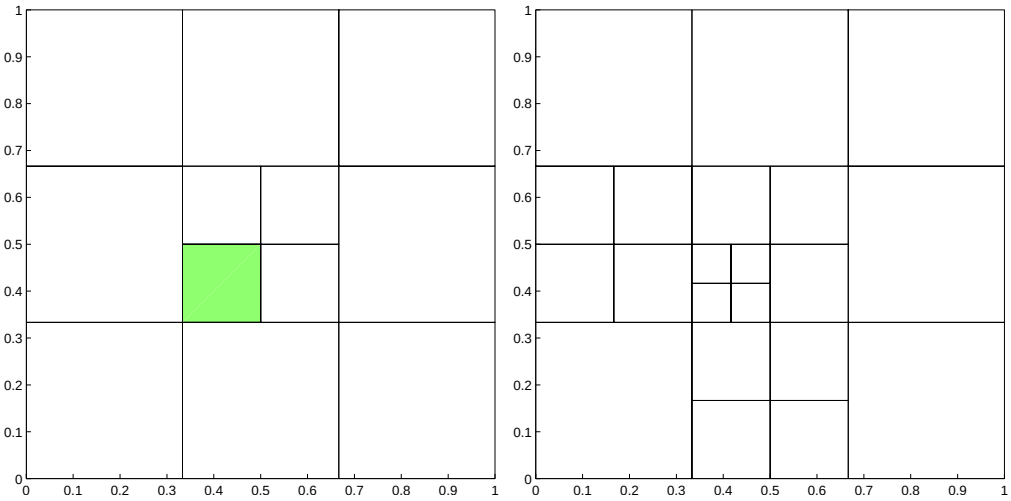

Figure 7.1.3: Refining an element (green) that is not refinable by first refining its neighbors

An element ω of width h is a candidate for h-refinement, if and only if all neighboring elements are of width h or $h/2$; an element ω of width h is a candidate for coarsening, if the element is part of a square of size $2h$ containing 4 elements ω_i, $1 \leq i \leq 4$, and all neighboring elements of ω_i, $1 \leq i \leq 4$ are of width h or $2h$. Thus, an element of size h can only have neighbors of size $h/2$, h or $2h$.

If we want to refine an element ω that is not a candidate for h-refinement, (because at least one neighboring element ω^- is of size $2h$), we first refine the neighboring elements of size $2h$, which is illustrated in Figure 7.1.3. If we want to coarsen a set of elements $\{\omega_i\}_{i=1}^4$ of size h that is not a candidate for coarsening (because at least one element ω^- neighboring the set is of size $h/2$), we first have to coarsen the neighboring elements until $\{\omega_i\}_{i=1}^4$ becomes a candidate for coarsening.

An example for a refined mesh can be seen in Figure 7.1.4.

We define the finite element spaces $\mathcal{V}^n$, $\mathcal{W}^n$ by

$$\mathcal{V}^n = \{\mathbf{v}(t,\mathbf{x}) : \ \mathbf{v}|_\omega \in \mathcal{P}_p, \ \omega \in \mathcal{T}^n, \ t^{n-1} \leq t \leq t^n\}, \tag{7.1.1}$$

$$\mathcal{W}^n = \Bigg\{ \mathbf{v}(t,\mathbf{x}(\boldsymbol{\xi})) = \sum_{i=1}^{2} \left(L_{p+1}(\xi_j)\mathbf{a}_j(t) - L_p(\xi_j)\mathbf{b}_j(t)\right) : \\ \mathbf{a}_i(t), \mathbf{b}_i(t) \in \mathbb{R}^m, \ \boldsymbol{\xi} \in \Delta, \ \omega \in \mathcal{T}^n, \ t^{n-1} \leq t \leq t^n \Bigg\}, \quad 1 \leq n \leq N. \tag{7.1.2}$$

Now we show how to obtain a solution $\mathbf{U}_{new}(t^n,\cdot)$ and an error estimate $\mathbf{E}_{new}(t^n,\cdot)$ on Σ^{n+1}.

If we refined ω into $\{\omega_i\}_{i=1}^4$, then we obtain $\mathbf{U}_{new}(t^n,\cdot) \in \mathcal{V}^{n+1}$ and $\mathbf{E}_{new}(t^n,\cdot) \in \mathcal{W}^{n+1}$ from

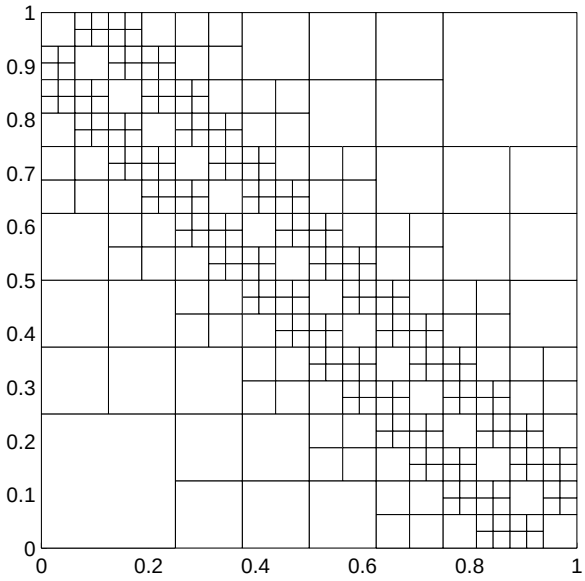

Figure 7.1.4: Example of an adaptive mesh obtained from a 4×4 initial mesh

$U(t^n, \cdot) \in \mathcal{V}^n$ and $E(t^n, \cdot) \in \mathcal{W}^n$ by

$$\mathbf{U}_{new}(t^n, \mathbf{x}) = \pi_p^{\omega_i}(\mathbf{U} + \mathbf{E})(t^n, \mathbf{x}), \quad \mathbf{x} \in \omega_i,\ 1 \le i \le 4, \tag{7.1.3}$$

$$\mathbf{E}_{new}(t^n, \mathbf{x}) = (\mathbf{U} + \mathbf{E} - \mathbf{U}_{new})(t^n, \mathbf{x}), \quad \mathbf{x} \in \omega_i,\ 1 \le i \le 4. \tag{7.1.4}$$

Similarly, If we coarsened $\{\omega_i\}_{i=1}^4$ into ω, then we obtain $\mathbf{U}_{new}(t^n, \cdot) \in \mathcal{V}^{n+1}$ and $\mathbf{E}_{new}(t^n, \cdot) \in \mathcal{W}^{n+1}$ from $U(t^n, \cdot) \in \mathcal{V}^n$ and $E(t^n, \cdot) \in \mathcal{W}^n$ by

$$\mathbf{U}_{new}(t^n, \mathbf{x}) = \pi_p^{\omega}(\mathbf{U} + \mathbf{E})(t^n, \mathbf{x}), \quad \mathbf{x} \in \omega, \tag{7.1.5}$$

$$\mathbf{E}_{new}(t^n, \mathbf{x}) = (\mathbf{U} + \mathbf{E} - \mathbf{U}_{new})(t^n, \mathbf{x}), \quad \mathbf{x} \in \omega. \tag{7.1.6}$$

Next, we give a criterium for mesh coarsening. To obtain an efficient algorithm, we try to coarsen such that

$$|\omega|^{-1/2}\|\mathbf{E}_{new}\|_{2,\omega} \lesssim \frac{1}{2} tol. \tag{7.1.7}$$

However, since we do not know $\mathbf{E}_{new}$ before we coarsen the mesh, we need to develop a criteria that uses $\mathbf{E}$ instead of $\mathbf{E}_{new}$. Since $|\omega_i|^{-1/2}\|\mathbf{E}\|_{2,\omega_i} \approx h^{-1}Ch^{p+1} = Ch^p$, where h denotes the size of ω_i, $1 \le i \le 4$, we expect $|\omega|^{-1/2}\|\mathbf{E}_{new}\|_{2,\omega} \approx (2h)^{-1}C(2h)^{p+1} = 2^pCh^p$. Thus, we coarsen $\omega \in \mathcal{T}^n$, if

$$|\omega|^{-1/2}\|\mathbf{E}(t^n, \cdot)\|_{2,\omega} < 2^{-p-1} tol, \tag{7.1.8}$$

which yields $|\omega|^{-1/2}\|\mathbf{E}_{new}\|_{2,\omega} \lesssim \frac{1}{2}tol$. If, after coarsening, $|\omega|^{-1/2}\|\mathbf{E}_{new}\|_{2,\omega} > tol$, then we reject the coarsening step and return to the original partition $\{\omega_i\}_{i=1}^4$.

Finally, we will describe how to integrate $\mathbf{U}(t,\cdot)$, $\mathbf{E}(t,\cdot)$ in time from $t = t^{n-1}$ to t^n.

We say that γ_i^s contains a irregular node, if the neighboring element is of either double or half the size of ω.

On (t^{n-1}, t^n), where the partition $\mathcal{T}^n$ is fixed, we integrate $\mathbf{U} \in \mathcal{V}^n$ in time such that

$$\int_\omega \mathbf{v}^t \left(\frac{\partial \mathbf{U}}{\partial t} - \mathbf{g}\right) d\mathbf{x} = \sum_{j=1}^2 \left(\int_\omega \frac{\partial \mathbf{v}^t}{\partial x_j}\mathbf{A}_j\mathbf{U}\, d\mathbf{x} - \int_{\gamma_i} \mathbf{v}^t\nu_j(\mathbf{A}_j^{\mu_j}\mathbf{U}^+ + \mathbf{A}_j^{\bar\mu_j}h(\mathbf{U}^-,\mathbf{E}^-))\, d\mathbf{s}\right),$$
$$\forall\ \mathbf{v} \in \mathcal{V}^n,\ \omega \in \mathcal{T}^n,\ t^{n-1} < t < t^n, \tag{7.1.9a}$$

subject to the boundary conditions

$$(\nu_i\mathbf{A}_i^{\bar\mu_i})\mathbf{U}^-(t,\mathbf{x}) = (\nu_i\mathbf{A}_i^{\bar\mu_i})\pi_p^{\gamma_i^s}\mathbf{u}_B(t,\mathbf{x}), \quad \mathbf{x} \in \gamma_i^s \cap \partial\Omega,\ s = +,-,\ i = 1,2,\ \omega \in \mathcal{T}^n, \tag{7.1.9b}$$

where

$$h(\mathbf{U}^-,\mathbf{E}^-)|_{\gamma_i^s} = \begin{cases} \pi_p^{\gamma_i^s}(\mathbf{U}^- + \mathbf{E}^-) & \text{if } \gamma_i^s \text{ contains a irregular node,} \\ \mathbf{U}^- & \text{else.} \end{cases} \tag{7.1.9c}$$

To integrate $\mathbf{E} = \mathbf{E}^\perp + \mathbf{E}^\Phi \in \mathcal{W}^n$ in time, we define

$$\mathbf{E}^\perp(t,\mathbf{x}(\boldsymbol{\xi})) = \sum_{j=1}^2 (L_{p+1}(\xi_j) - L_p(\xi_j)\mathrm{sgn}(\mathbf{A}_j))\,\frac{h^{-1}}{2}\mathbf{A}_j^\dagger\mathfrak{r}_{p,j}^\perp,$$
$$\boldsymbol{\xi} \in \Delta,\ \omega \in \mathcal{T}^n,\ t^{n-1} \le t \le t^n, \tag{7.1.10a}$$

where $\mathfrak{r}_{p,j}^\perp,\ j = 1,2$, are defined in (4.3.24b).

Then we define

$$\mathbf{E}^\Phi(t,\mathbf{x}(\boldsymbol{\xi})) = \sum_{j=1}^2 L_{p+1}(\xi_j)\boldsymbol{\gamma}_j^\Phi(t) - L_p(\xi_j)\boldsymbol{\delta}_j^\Phi(t),$$
$$\boldsymbol{\gamma}_j^\Phi, \boldsymbol{\delta}_j^\Phi \in \mathcal{N}(\mathbf{A}_j),\ j = 1,2,\ \boldsymbol{\xi} \in \Delta,\ \omega \in \mathcal{T}^n,\ t^{n-1} \le t \le t^n, \tag{7.1.10b}$$

such that

$$\int_\omega \mathbf{v}^t \left(\frac{\partial(\mathbf{U} + \mathbf{E}^\Phi)}{\partial t} + \sum_{j=1}^2 \mathbf{A}_j\frac{\partial \mathbf{U}}{\partial x_j} - \mathbf{g}\right) d\mathbf{x}$$
$$= \sum_{j=1}^2 \int_{\gamma_j} \mathbf{v}^t\nu_j\mathbf{A}_j^{\bar\mu_j}\left(\mathbf{U} + \mathbf{E}^\perp + \mathbf{E}^\Phi - k(\mathbf{U}^-,\mathbf{E}^-)\right) d\mathbf{s},\ \forall\ \mathbf{v} \in \mathcal{E}_p,\ \omega \in \mathcal{T}^n, \tag{7.1.10c}$$

subject to the boundary conditions

$$(\nu_i \mathbf{A}_i^{\bar{\mu}_i})\mathbf{E}^-(t,\mathbf{x}) = (\nu_i \mathbf{A}_i^{\bar{\mu}_i})(\mathcal{L}_{p+1}\mathbf{u} - \pi_p^{\gamma_i^s}\mathbf{u})(t,\mathbf{x}), \quad \mathbf{x} \in \gamma_i^s \cap \partial\Omega,\ s = +,-,\ i = 1,2,\ \omega \in \mathcal{T}^n, \tag{7.1.10d}$$

where

$$k(\mathbf{U}^-, \mathbf{E}^-)|_{\gamma_i^s} = \begin{cases} (\mathbf{U}^- + \mathbf{E}^-) - \pi_p^{\gamma_i^s}(\mathbf{U}^- + \mathbf{E}^-) & \text{if } \gamma_i^s \text{ contains a irregular node,} \\ \mathbf{E}^- & \text{else.} \end{cases} \tag{7.1.10e}$$

Note that the definition of $\mathbf{U}$ and $\mathbf{E}$ is equivalent to the DG formulation for $\mathbf{u}_h$ and $\mathbf{E}^\perp + \mathbf{E}^{\maltese}$, if $\mathcal{T}^n = \mathcal{T}_h$, $1 \le n \le N$.

We introduced functions k, h to obtain boundary conditions consistent with the boundary conditions on regular elements.

7.2 An p-Adaptive Enrichment Algorithm

For p-refinement, we keep the initial mesh $\mathcal{T}^0 = \mathcal{T}_h$ fixed during all refinement steps, *i.e.* $\mathcal{T}^n = \mathcal{T}_h,\ 1 \le n \le N.$ but vary the polynomial order on each element.

For $p_\omega^n \ge 1,\ \omega \in \mathcal{T}_h$, we define the finite element spaces

$$\mathcal{V}^n = \{\mathbf{v}(t,\mathbf{x}) :\ \mathbf{v}|_\omega \in \mathcal{P}_{p_\omega^n},\ \omega \in \mathcal{T}_h,\ t^{n-1} \le t \le t^n\}. \tag{7.2.1}$$

$$\mathcal{W}^n = \Bigg\{ \mathbf{v}(t,\mathbf{x}(\boldsymbol{\xi})) = \sum_{i=1}^{2} \left(L_{p_\omega^n+1}(\xi_j)\mathbf{a}_j(t) - L_{p_\omega^n}(\xi_j)\mathbf{b}_j(t)\right) : \mathbf{a}_i(t), \mathbf{b}_i(t) \in \mathbb{R}^m,\ \boldsymbol{\xi} \in \Delta,\ \omega \in \mathcal{T}_h,\ t^{n-1} \le t \le t^n \Bigg\},\ 1 \le n \le N. \tag{7.2.2}$$

To refine ω, we increase the order p_ω^n by 1, to coarsen ω, we decrease the order p_ω^n by 1, up to a minimum order of 1. In order not to create too abrupt changes of order, an element ω of order p_ω^n is a candidate for enrichment, if and only if all neighboring elements are of order p_ω^n or $p_\omega^n + 1$; an element ω of width h is a candidate for coarsening, if all neighboring elements are of order p_ω^n or $p_\omega^n - 1$. Thus, an element of order p_ω^n can only have neighbors of order $p_\omega^n - 1$, p_ω^n or $p_\omega^n + 1$.

If we want to refine an element ω that is not a candidate for enrichment (because at least one neighboring element ω^- is of order $p_\omega^n - 1$), we first refine the neighboring elements of order $p_\omega^n - 1$. Similar, if we want to coarsen an element ω that is not a candidate for coarsening, we first coarsen the neighboring elements of order $p_\omega^n + 1$.

Now we show how to obtain a solution $\mathbf{U}_{new}(t^n,\cdot)$ and an error estimate $\mathbf{E}_{new}(t^n,\cdot)$ on Σ^{n+1}.

If we refined $\omega\in\mathcal{T}_h$, s.t. $p_\omega^{n+1}=p_\omega^n+1$, then we obtain $\mathbf{U}_{new}(t^n,\cdot)\in\mathcal{V}^{n+1}$ and $\mathbf{E}_{new}(t^n,\cdot)\in\mathcal{W}^{n+1}$ from $U(t^n,\cdot)\in\mathcal{V}^n$ and $E(t^n,\cdot)\in\mathcal{W}^n$ by

$$\mathbf{U}_{new}(t^n,\mathbf{x})=(\mathbf{U}+\mathbf{E})(t^n,\mathbf{x}),\quad \mathbf{x}\in\omega_i, \tag{7.2.3}$$
$$\mathbf{E}_{new}(t^n,\mathbf{x})=0. \tag{7.2.4}$$

If we coarsened $\omega\in\mathcal{T}_h$, s.t. $p_\omega^{n+1}=p_\omega^n-1$, then we obtain $\mathbf{U}_{new}(t^n,\cdot)\in\mathcal{V}^{n+1}$ and $\mathbf{E}_{new}(t^n,\cdot)\in\mathcal{W}^{n+1}$ from $U(t^n,\cdot)\in\mathcal{V}^n$ and $E(t^n,\cdot)\in\mathcal{W}^n$ by

$$\mathbf{U}_{new}(t^n,\mathbf{x})=\pi_{p-1}^n(\mathbf{U}+\mathbf{E})(t^n,\mathbf{x}),\quad \mathbf{x}\in\omega_i, \tag{7.2.5}$$
$$\mathbf{E}_{new}(t^n,\mathbf{x})=\mathcal{L}_p^\omega(\mathbf{U}+\mathbf{E})(t^n,\mathbf{x})-\mathbf{U}_{new}(t^n,\mathbf{x}). \tag{7.2.6}$$

Since it is simple to calculate $\|\mathbf{E}_{new}\|_{2,\omega}$, we use

$$|\omega|^{-1/2}\|\mathbf{E}_{new}(t^n,\cdot)\|_{2,\omega}<\frac{1}{2}tol, \tag{7.2.7}$$

as condition for mesh coarsening.

Finally, we will describe how to integrate $\mathbf{U}(t,\cdot)$, $\mathbf{E}(t,\cdot)$ in time from $t=t^{n-1}$ to t^n.

On (t^{n-1},t^n), where the partition $\mathcal{T}^n$ is fixed, we integrate $\mathbf{U}\in\mathcal{V}^n$ in time such that

$$\int_\omega \mathbf{v}^t\left(\frac{\partial\mathbf{U}}{\partial t}-\mathbf{g}\right)d\mathbf{x}=\sum_{j=1}^2\left(\int_\omega\frac{\partial\mathbf{v}^t}{\partial x_j}\mathbf{A}_j\mathbf{U}\,d\mathbf{x}-\int_{\gamma_i}\mathbf{v}^t\nu_j(\mathbf{A}_j^{\mu_j}\mathbf{U}^+ + \mathbf{A}_j^{\bar\mu_j}h(\mathbf{U}^-,\mathbf{E}^-))\,d\mathbf{s}\right),$$
$$\forall\ \mathbf{v}\in\mathcal{V}^n,\ \omega\in\mathcal{T}^n,\ t^{n-1}<t<t^n, \tag{7.2.8a}$$

subject to the boundary conditions

$$(\nu_i\mathbf{A}_i^{\bar\mu_i})\mathbf{U}^-(t,\mathbf{x})=(\nu_i\mathbf{A}_i^{\bar\mu_i})\pi_p^{\gamma_i^s}\mathbf{u}(t,\mathbf{x}),\quad \mathbf{x}\in\gamma_i^s\cap\partial\Omega,\ s=+,-,\ i=1,2,\ \omega\in\mathcal{T}^n, \tag{7.2.8b}$$

with

$$h(\mathbf{U}^-,\mathbf{E}^-)|_{\gamma_i^s}=\begin{cases}\pi_{p_\omega^n-1}^{\gamma_i^s}(\mathbf{U}^-+\mathbf{E}^-) & \text{if } p_{\omega^-}^n<p_\omega^n,\\ \mathbf{U}^-+\mathbf{E}^- & \text{if } p_{\omega^-}^n>p_\omega^n,\\ \mathbf{U}^- & \text{else,}\end{cases} \tag{7.2.8c}$$

where ω^- denotes the neighboring element on each face γ_i^s.

To integrate $\mathbf{E}=\mathbf{E}^\perp+\mathbf{E}^{\maltese}\in\mathcal{W}^n$ in time, we define

$$\mathbf{E}^\perp(t,\mathbf{x}(\boldsymbol{\xi}))=\sum_{j=1}^2\left(L_{p_\omega^n+1}(\xi_j)-L_{p_\omega^n}(\xi_j)\mathrm{sgn}(\mathbf{A}_j)\right)\frac{h^{-1}}{2}\mathbf{A}_j^\dagger\mathfrak{r}_{p_\omega^n,j}^\perp,$$
$$\boldsymbol{\xi}\in\Delta,\ \omega\in\mathcal{T}^n,\ t^{n-1}\le t\le t^n, \tag{7.2.9a}$$

where $\mathfrak{r}^{\perp}_{p^n_\omega,j},\ j=1,2$, are defined in (4.3.24b).

Then we define

$$\mathbf{E}^{\mathbf{\Phi}}(t,\mathbf{x}(\boldsymbol{\xi})) = \sum_{j=1}^{2} L_{p^n_\omega+1}(\xi_j)\boldsymbol{\gamma}^{\mathbf{\Phi}}_j(t) - L_{p^n_\omega}(\xi_j)\boldsymbol{\delta}^{\mathbf{\Phi}}_j(t),$$
$$\boldsymbol{\gamma}^{\mathbf{\Phi}}_j, \boldsymbol{\delta}^{\mathbf{\Phi}}_j \in \mathcal{N}(\mathbf{A}_j),\ j=1,2,\ \boldsymbol{\xi}\in\Delta,\ \omega\in\mathcal{T}^n,\ t^{n-1}\le t\le t^n, \tag{7.2.9b}$$

such that

$$\int_\omega \mathbf{v}^t\left(\frac{\partial(\mathbf{U}+\mathbf{E}^{\mathbf{\Phi}})}{\partial t} + \sum_{j=1}^{2}\mathbf{A}_j\frac{\partial\mathbf{U}}{\partial x_j} - \mathbf{g}\right)d\mathbf{x}$$
$$= \sum_{j=1}^{2}\int_{\gamma_j}\mathbf{v}^t\nu_j\mathbf{A}_j^{\bar{\mu}_j}\left(\mathbf{U}+\mathbf{E}^{\perp}+\mathbf{E}^{\mathbf{\Phi}}-k(\mathbf{U}^-,\mathbf{E}^-)\right)d\mathbf{s},\ \forall\ \mathbf{v}\in\mathcal{E}_p,\ \omega\in\mathcal{T}^n, \tag{7.2.9c}$$

subject to the boundary conditions

$$(\nu_i\mathbf{A}_i^{\bar{\mu}_i})\mathbf{E}^-(t,\mathbf{x}) = (\nu_i\mathbf{A}_i^{\bar{\mu}_i})(\mathcal{L}_{p^n_\omega+1}\mathbf{u} - \pi^{\gamma_i^s}_{p^n_\omega}\mathbf{u})(t,\mathbf{x}), \tag{7.2.9d}$$
$$\mathbf{x}\in\gamma_i^s\cap\partial\Omega,\ s=+,-,\ i=1,2,\ \omega\in\mathcal{T}^n,$$

where

$$k(\mathbf{U}^-,\mathbf{E}^-)|_{\gamma_i^s} = \begin{cases} \mathcal{L}^{\gamma_i^s}_{p^n_\omega}(\mathbf{U}^-+\mathbf{E}^-) - \pi^{\gamma_i^s}_{p^n_\omega-1}(\mathbf{U}^-+\mathbf{E}^-) & \text{if } p^n_{\omega^-} < p^n_\omega, \\ \mathbf{0} & \text{if } p^n_{\omega^-} > p^n_\omega, \\ \mathbf{E}^- & \text{else.}\end{cases} \tag{7.2.9e}$$

Note that the definition of $\mathbf{U}$ and $\mathbf{E}$ is equivalent to the DG formulation for $\mathbf{U}=\mathbf{u}_h$ and $\mathbf{E}=\mathbf{E}^{\perp}+\mathbf{E}^{\mathbf{\Phi}}$, if $\mathcal{T}^n=\mathcal{T}_h$, $1\le n\le N$.

7.3 Computational Examples

Example 7.3.1. *Let us consider the equation*

$$\mathbf{u}_{,t} + \begin{pmatrix}0&1\\1&0\end{pmatrix}\mathbf{u}_{,x} + \begin{pmatrix}1&0\\0&-1\end{pmatrix}\mathbf{u}_{,y} = \mathbf{g}, (x,y)\in(0,1)^2,\ 0\le t\le 2. \tag{7.3.1a}$$

The source term $\mathbf{g}$ *and initial and boundary conditions are such that the true solution is*

$$\mathbf{u}(t,x,y) = (1\ 1)^t\tanh(10(x+y-t)). \tag{7.3.1b}$$

The solution $\mathbf{u}$ is steep close to the line $x+y=t$ and smooth elsewhere, see Figure 7.3.1. Hence the error close to the line $x+y=t$ should be larger and thus the mesh will be refined

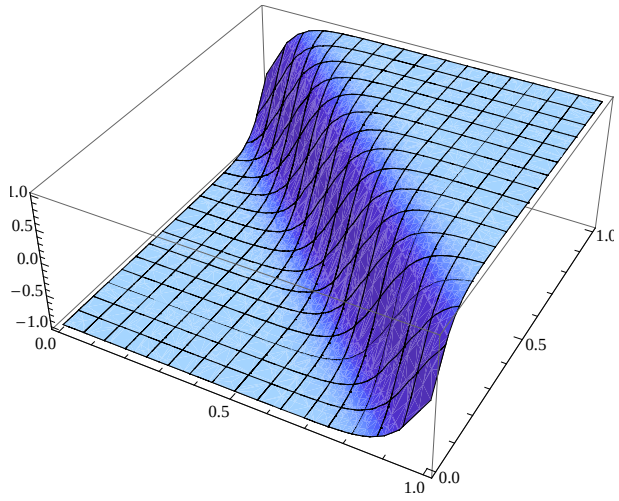

Figure 7.3.1: Example 7.3.1: $\tanh(10(x+y-t))$ at $t=1$

near $x+y=t$. Using the static error estimate $\mathbf{E}^{\perp}$ as criterium for refinement (since $\mathbf{E}^{\Phi}=0$), we can drive both h- and p-refinement.

We first find an h-refined solution of (7.3.1) for an initial 4×4, tolerance $tol=0.01$ and order of approximation $p=1$. We plot the refined mesh at timesteps $t=0,\ 0.8542,\ 1.5293$ in Figure 7.3.2. We observe that the algorithm refines the mesh close to the line $x+y=t$ and coarsens the mesh away from $x+y=t$. We plot the effectivity index θ in Figure 7.3.3 and the L_2-error $\|\mathbf{e}\|_{2,\Omega}$ (solid) and estimate $\|\mathbf{E}^{\perp}\|_{2,\Omega}$ (dotted) in Figure 7.3.4, where $\circ$'s denote refinement and coarsening steps. We observe that the error decreases when we refine the mesh, and the error estimate approaches unity soon after each refinement step.

Similarly, we find a p-refined solution of (7.3.1) for a 20×20 mesh, tolerance $tol=0.01$ and initial order of approximation $p=1$. We plot the refined mesh at timesteps $t=0.31843,\ 0.90719$, and 1.5236 in Figure 7.3.5. Again, we observe that the algorithm refines the mesh close to the line $x+y=t$ and coarsens the mesh away from $x+y=t$.

7.4 A Nonlinear Problem

We finish this chapter with an outlook into the nonlinear case. Therefore assume that $\mathbf{u}$ satisfies the nonlinear system

$$\frac{\partial \mathbf{u}}{\partial t}+\sum_{i=1}^{d}\mathbf{A}_i(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x_i}=\mathbf{g}(t,\mathbf{x}),\quad \mathbf{x}\in\Omega,\quad 0<t<T, \tag{7.4.1a}$$

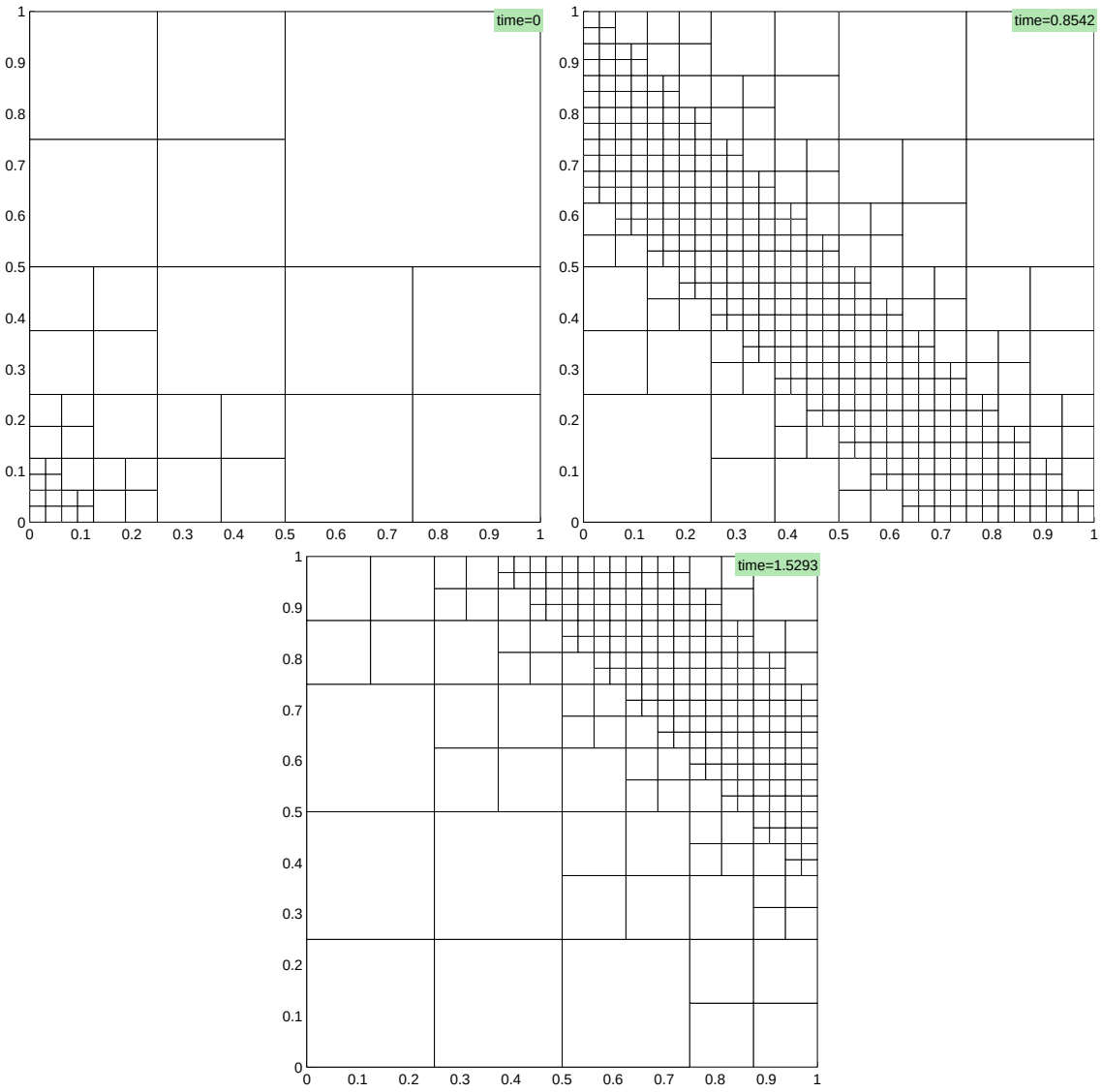

Figure 7.3.2: h-refined mesh for Example 7.3.1 with $p = 1$ and $tol = 10^{-2}$ at $t = 0,\ 0.8542,\ 1.5293$ for an initial 4×4 mesh.

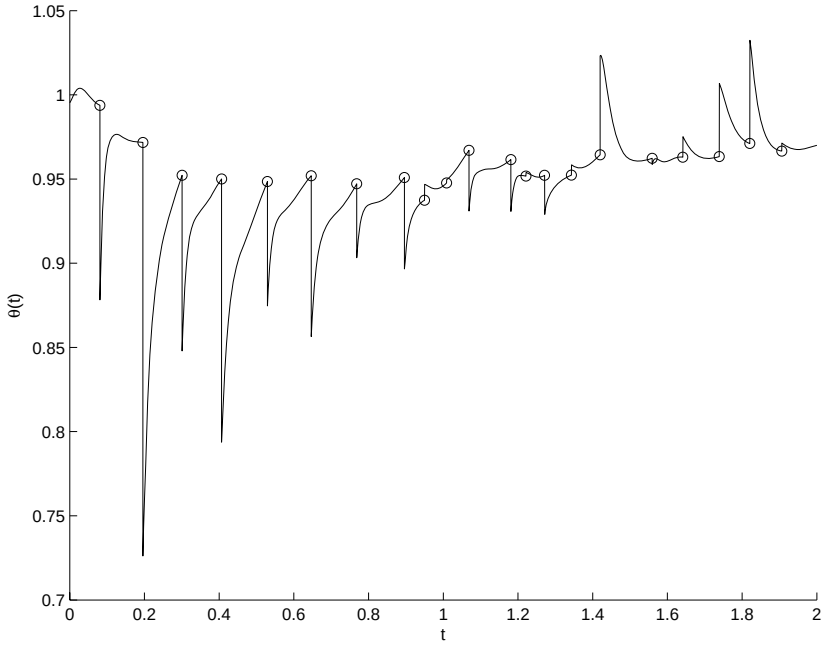

Figure 7.3.3: Effectivity index θ over time for h-refined mesh in Example 7.3.1 with $p = 1$, $tol = 10^{-2}$ and an initial 4×4 mesh. $\circ$ denote refinement steps.

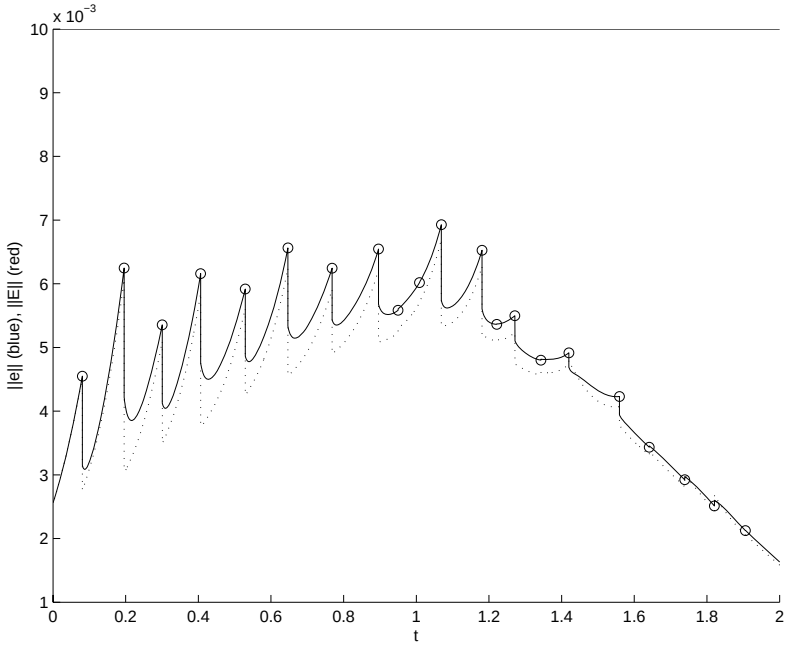

Figure 7.3.4: L_2-error $\|\mathbf{e}\|_{2,\Omega}$ (solid) and estimate $\|\mathbf{E}^\perp\|_{2,\Omega}$ (dotted) over time for h-refined mesh in Example 7.3.1 with $p = 1$, $tol = 10^{-2}$ and an initial 4×4 mesh. $\circ$ denote refinement steps.

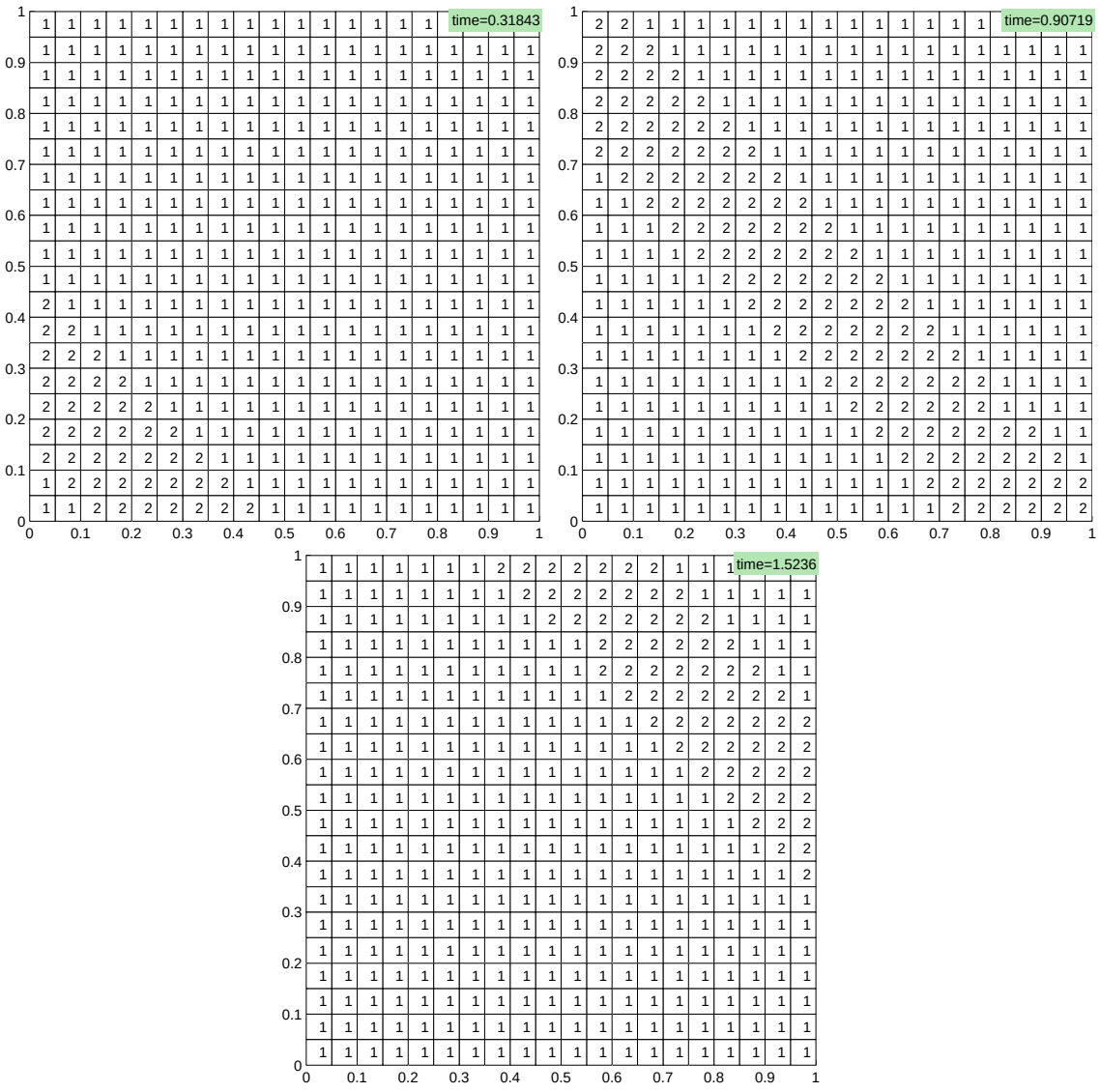

Figure 7.3.5: p-enriched mesh for Example 7.3.1 with $h = 1/20$ and $tol = 10^{-2}$ and initial order $p = 1$ for $t = 0.31843,\ 0.90719,\ 1.5236$

with *source term* $\mathbf{g} : (0,T) \times \Omega \to \mathbb{R}^m$ and subject to the initial and boundary conditions

$$\mathbf{u}(0,\mathbf{x}) = \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{7.4.1b}$$

$$\left(\sum_{i=1}^{d} \nu_i \mathbf{A}_i^{\bar{\mu}_i}(\mathbf{u}_B)\right) \mathbf{u} = \left(\sum_{i=1}^{d} \nu_i \mathbf{A}_i^{\bar{\mu}_i}(\mathbf{u}_B)\right) \mathbf{u}_\mathbf{B}, \quad \mathbf{x} \in \partial\Omega,\ 0 < t < T, \tag{7.4.1c}$$

and let the DG method of the Steger-Warming numerical flux consist of finding $\mathbf{u}_h \in \mathcal{V}_p^h$ that satisfies

$$\begin{aligned}\int_\omega \mathbf{v}^t \left(\frac{\partial \mathbf{u}_h}{\partial t} - \mathbf{g}\right) d\mathbf{x} = \sum_{j=1}^{d} \Bigg(&\int_\omega \frac{\partial \mathbf{v}^t}{\partial x_j} \mathbf{A}_j(\mathbf{u}_h)\mathbf{u}_h \, d\mathbf{x} \\ &- \int_{\partial\omega} \mathbf{v}^t \nu_j (\mathbf{A}_j^{\mu_j}(\mathbf{u}_h)\mathbf{u}_h + \mathbf{A}_j^{\bar{\mu}_j}(\mathbf{u}_h^-)\mathbf{u}_h^-)\, d\mathbf{s}\Bigg), \quad \forall\ \mathbf{v} \in \mathcal{V}_p^h,\ \omega \in \mathcal{T}_h,\ 0 < t < T,\end{aligned} \tag{7.4.2a}$$

subject to the initial and boundary conditions

$$\mathbf{u}_h(0,\mathbf{x}) = \pi\mathbf{u}_0(\mathbf{x}) \text{ or } \mathbf{u}_h(0,\mathbf{x}) = \Pi\mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \omega, \tag{7.4.2b}$$

$$(\nu_i \mathbf{A}_i^{\bar{\mu}_i}(\mathbf{u}))\mathbf{u}_h^- = (\nu_i \mathbf{A}_i^{\bar{\mu}_i}(\mathbf{u}))\pi_i^s \mathbf{u}, \quad \mathbf{x} \in \gamma_i^s \cap \partial\Omega,\ s = +,-,\ 1 \le i \le d,\ \omega \in \mathcal{T}_h,\ 0 < t < T. \tag{7.4.2c}$$

Below we present a few simple examples for nonlinear problems, where we will use $\Pi\mathbf{u}_0$ as approximations for the initial conditions.

Example 7.4.1. *Let us consider the non-linear scalar equation*

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x_1} = \mathbf{g}, \tag{7.4.3a}$$

on $\Omega = (0,1)$ *and* $\frac{3}{2} \le t \le \frac{5}{2}$*, with source term, initial and boundary conditions such that*

$$\mathbf{u}(t,x) = \sin(t+x). \tag{7.4.3b}$$

We note that the real solution allows positive and negative values. If Theorem 4.2.1 can be generalized to the nonlinear case, we would expect that the discretization error $\mathbf{e} = \mathbf{u} - \mathbf{u}_h$ satisfies

$$\begin{aligned}\mathbf{e}(t, h\boldsymbol{\xi}) &= h^{p+1}\left(L_{p+1}(\xi)\mathbf{c}(t) - L_p(\xi)(\operatorname{sgn}(\mathbf{u}_h)\mathbf{c}(t) + \mathbf{d}_j)\right) + \mathcal{O}(h^{p+2}) \\ &= \begin{cases} h^{p+1}R_{p+1}^+(\xi)\mathbf{c}(t) + \mathcal{O}(h^{p+2}), & \text{if } \mathbf{u}_h > 0, \forall\ \mathbf{x} \in \omega, \\ h^{p+1}R_{p+1}^-(\xi)\mathbf{c}(t) + \mathcal{O}(h^{p+2}), & \text{if } \mathbf{u}_h < 0, \forall\ \mathbf{x} \in \omega, \\ \mathcal{O}(h^{p+1}) & \text{else,}\end{cases} \quad \forall\ \omega \in \mathcal{T}_h.\end{aligned} \tag{7.4.4}$$

In Figure 7.4.1 we plot the error $\mathbf{e}$ on $\Omega = (0,1)$ at time $t = 2.5$ for $p = 1,2,3$ and $N = 10, 20, 30$. We can see that the error transitions smoothly from left to right Radau

polynomials at $x = \pi - 2.5$, where $\mathbf{u}$ changes sign. In Figure 7.4.2 we plot the effectivity index over time for the static error estimate $\mathbf{E}^\perp$. We observe that the error estimate behaves erratically in some time intervals. In Figure 7.4.3 we plot the error estimate $\mathbf{E}^\perp$ on $\Omega = (0, 1)$ at times where the effectivity index is large. We chose $t = 2.47$ for $N = 20$, $t = 2.495$ for $N = 30$ and $t = 2.48$ for $N = 40$. We observe that the error estimate behaves erratically only near $x = \pi - 2.5$.

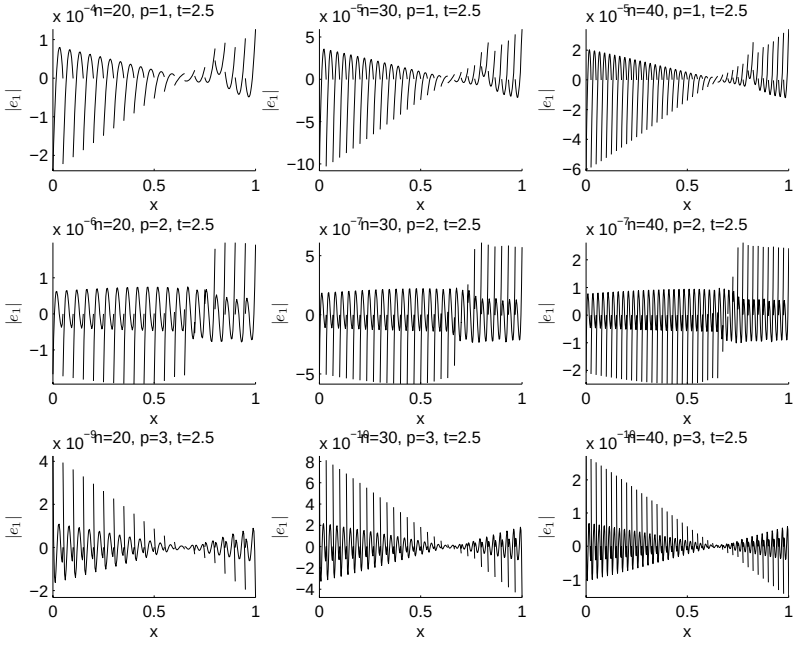

Figure 7.4.1: Error $\mathbf{e}(t, x)$ over $x \in (0, 1)$ for Example 7.4.1 at $t = 2.5$

Example 7.4.2. *Let us consider the non-linear equation*

$$\frac{\partial \mathbf{u}}{\partial t} + \begin{pmatrix} u_1 & 0 \\ 0 & 1 \end{pmatrix} \frac{\partial \mathbf{u}}{\partial x_1} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \frac{\partial \mathbf{u}}{\partial x_2} = \mathbf{g}, \tag{7.4.5a}$$

on $\Omega = (0, 1)^2$ *and* $1 \leq t \leq 2$*, with source term, initial and boundary conditions such that*

$$\mathbf{u}(t, x, y) = (1\ 1)^t \sin(t + x + y). \tag{7.4.5b}$$

We note that $\begin{pmatrix} u_1 & 0 \\ 0 & 1 \end{pmatrix}$ contains zero eigenvalues in element near the line $x + y = \pi - t$. In Figure 7.4.4 we plot the local effectivity index θ_ω on $\Omega = (0, 1)^2$ at time $t = 2$ for $p = 1, 2, 3$ and $N = 10, 15, 20$. We observe that the estimates behave well except for the region near the line $x + y = \pi - 2$.

The results obtained from Examples 7.4.1 and 7.4.2 suggest that a modification of the error estimation technique is necessary near regions where the Jacobians of the flux $\mathbf{A}_i$, $1 \leq i \leq d$ contain 0 eigenvalues.

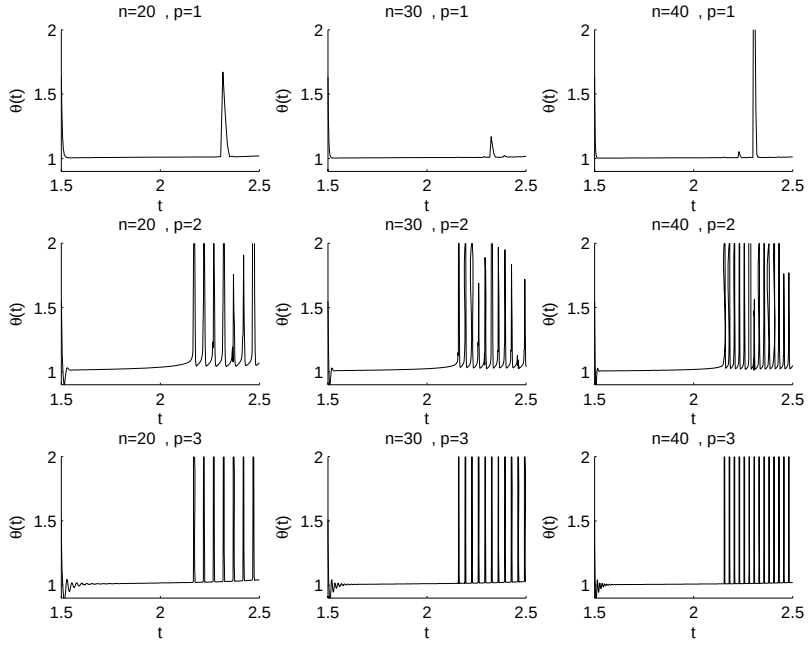

Figure 7.4.2: Global effectivity index for static error estimate θ on $t \in (1.5, 2.5)$ for Example 7.4.1

.

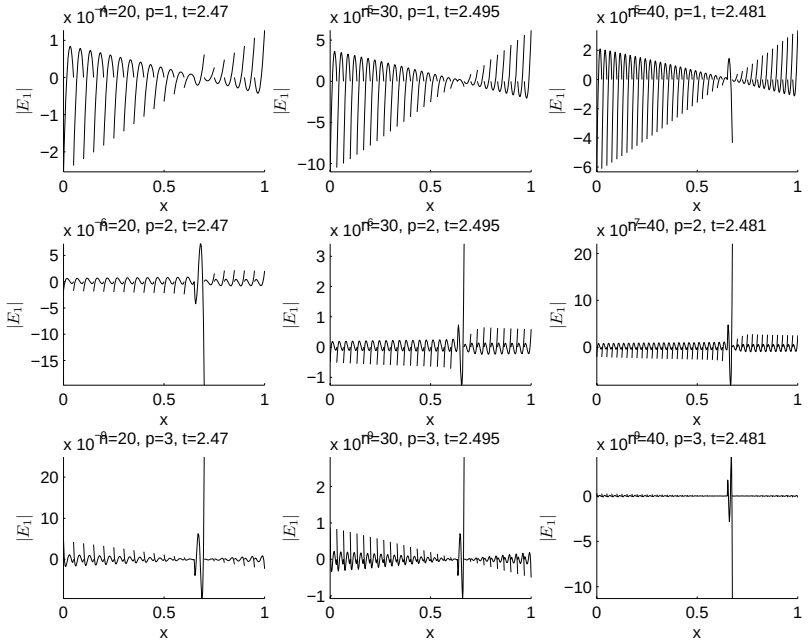

Figure 7.4.3: Static error estimate $\mathbf{E}(t, x)$ over $x \in (0, 1)$ at $t = 2.47$ for $N = 20$, $t = 2.495$ for $N = 30$ and $t = 2.48$ for $N = 40$ for Example 7.4.1

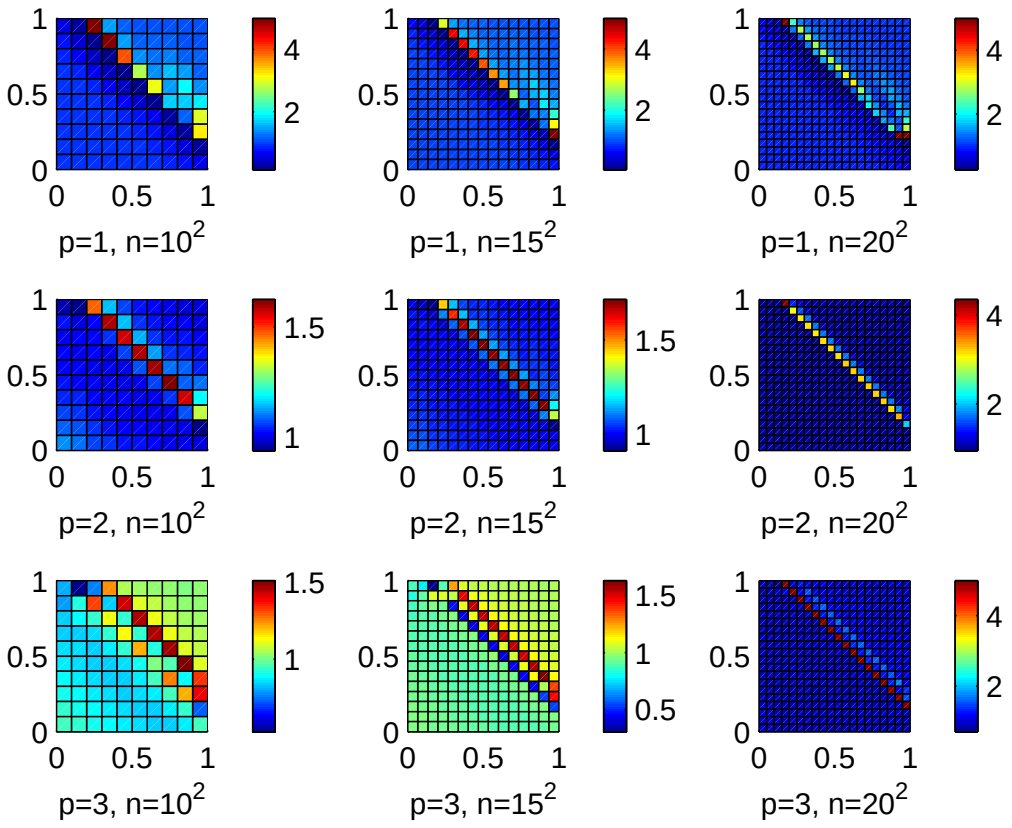

Figure 7.4.4: Local effectivity indices θ_ω on $\Omega = (0,1)^2$ for Example 7.4.2 at $t = 2$.

We will finally turn our attention to a example with discontinuous solution.

Example 7.4.3. *Let us consider the non-linear scalar equation*

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x_1} = \mathbf{0} \tag{7.4.6}$$

on $\Omega = (0,1)$ *and* $0 \leq t \leq \frac{1}{3}$*, with initial conditions* $\mathbf{u}_0(x) = 1$*,* $0 \leq x \leq 1$*, and boundary conditions* $\mathbf{u}_B(t,0) = 2$*,* $0 \leq t \leq \frac{1}{3}$*.*

Since $2 > 1$, we obtain by the Rankine-Hugoniot condition

$$\mathbf{u}(t,x) = \begin{cases} 2 & \text{if } x \leq \frac{3}{2}t, \\ 1 & \text{else.} \end{cases} \tag{7.4.7}$$

In Figure 7.4.5 we plot the error $u - u_h$ on $x \in (0,1)$ for $n = 10$ and $p = 1, 2, 3$. We observe that the error is large near the discontinuity. This suggests that a modification of the DG method is necessary near discontinuities.

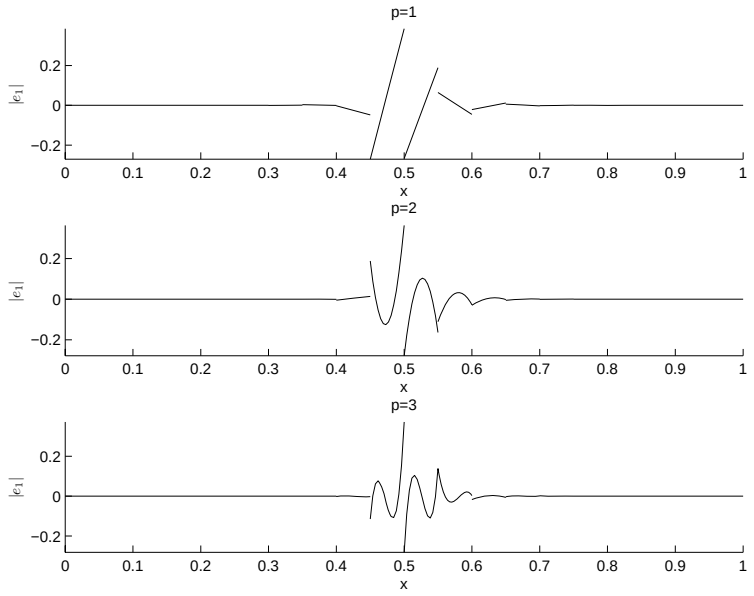

Figure 7.4.5: Error $(u - u_h)(t,x)$ on $x \in (0,1)$ for Example 7.4.3 at $t = \frac{1}{3}$ for $n = 20$ and $p = 1, 2, 3$

Chapter 8

Conclusions

8.1 Contributions

In this dissertation we developed a new, modified discontinuous Galerkin scheme for the space discretization of linear multi-dimensional hyperbolic systems of conservation laws. We chose an enriched polynomial space $\mathcal{P}_p$, $\mathbb{P}_p \subset \mathcal{P}_p \subset \mathbb{P}_{p+1}$, as a basis for the function space $\mathcal{V}_p^h$ on each element, and used corrected L_2-projections to approximate the initial- and boundary conditions. We performed a local error analysis which showed that the leading term of the discretization error lies in a polynomial subspace spanned by a linear combination of Legendre polynomials of order p and $p+1$. For special hyperbolic systems, where the coefficient matrices are nonsingular we showed that the leading term of the error is spanned by $(p+1)^{th}$-degree Radau polynomials. We also established new pointwise and averaged $\mathcal{O}(h^{p+2})$ superconvergence results.

We then turned our attention to the construction of a new implicit residual-based *a posteriori* error estimation procedure. We split the error into two parts and estimated each part separately by solving a small system of equations based on the local residual of the PDE. Thus, we were able to compute an efficient estimate of the discretization error locally on each element. For systems with invertible matrices, the error could be estimated by a static problem, while, for general systems, part of the error had to be computed by solving a transient system of equations. Local error analysis suggests that, for smooth solutions, both error estimates are asymptotically correct, that is they converge to the real error under mesh refinement. We first showed these results for linear symmetric systems that satisfy certain assumptions, then for general linear symmetric systems. Next, we generalized these results to linear symmetrizable systems by considering an equivalent symmetric formulation, which required us to make small modifications in the error estimation procedure. Numerical results confirmed the results of our analysis, for both the symmetric and the symmetrizable case in one, two and three space dimensions. Examples included the linearized Euler's equations,

Maxwell's equations and the acoustic wave equation, as well as several other systems.

We further investigated the behavior of the discretization error when other numerical fluxes such as Lax-Friedrichs are used. We observed that, while no superconvergence results could be obtained, an error estimation procedure can be developed for most cases. We further developed simple h- and p-refinement techniques to show that the error estimates can be successfully used to guide the refinement and coarsening process. Finally, we presented numerical results where we applied our DG formulation to some nonlinear problems.

8.2 Future Work

We note that, up to this point, we are not able to prove the asymptotic exactness of our global *a posteriori* error estimates. However, the computational results in this dissertation suggest that global *a posteriori* error estimates are asymptotically exact for smooth solutions. Thus, a focal point of research in the near future will be to establish a global error analysis. Further work has to be done on the extension of these results to more general linear and nonlinear hyperbolic systems. Since our theory does not hold near singularities, we plan to devise a strategy to detect discontinuities, which enables us to solve linear and nonlinear problems with discontinuous solutions and thus test the scope of applicability of our results. We also plan to investigate the extension of the work of Adjerid and Baccouch [2, 3] on triangular meshes to hyperbolic systems. Another point of interest is the development of truncation error estimates, which requires the solution to an adjoint problem to measure the pollution error as well as the local error, and to compare them to our discretization error estimates when applied to adaptivity algorithms.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1965.

[2] S. Adjerid and M. Baccouch. The discontinuous Galerkin method for two-dimensional hyperbolic problems Part I: Superconvergence error analysis. *Journal of Scientific Computing*, 33:75–113, 2007.

[3] S. Adjerid and M. Baccouch. The discontinuous Galerkin method for two-dimensional hyperbolic problems Part II: *A Posteriori* error estimation. *Journal of Scientifc Computing*, 2008. to appear.

[4] S. Adjerid, K. D. Devine, J. E. Flaherty, and L. Krivodonova. A posteriori error estimation for discontinuous Galerkin solutions of hyperbolic problems. *Computer Methods in Applied Mechanics and Engineering*, 191:1097–1112, 2002.

[5] S. Adjerid, J. E. Flaherty, and I. Babuška. A posteriori error estimation for the finite element method-of-lines solution of parabolic problems. *Mathematic Models and Methods in Applied Sciences*, 9(2):261–286, 1999.

[6] S. Adjerid and A. Klauser. Superconvergence of discontinuous finite element solutions for transient convection-diffusion problems. *Journal of Scientific computing*, 22:5–24, 2005.

[7] S. Adjerid and T. C. Massey. A posteriori discontinuous finite element error estimation for two-dimensional hyperbolic problems. *Computer Methods in Applied Mechanics and Engineering*, 191:5877–5897, 2002.

[8] S. Adjerid and T. C. Massey. Superconvergence of discontinuous finite element solutions for nonlinear hyperbolic problems. *Computer Methods in Applied Mechanics and Engineering*, 195:3331–3346, 2006.

[9] M. Ainsworth and J. T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. John Wiley, New York, 2000.

[10] I. Babuška and W. Rheinboldt. A-posteriori error estimates for the finite element method. *International Journal for Numerical Methods in Engineering*, 12:1597–1615, 1978.

[11] I. Babuška and W. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM Journal Numerical Analysis*, 18:736–754, 1978.

[12] I. Babuška and W. Rheinboldt. A posteriori error analysis of finite element solutions for onedimensional problems. *SIAM Journal Numerical Analysis*, 18:565–589, 1981.

[13] I. Babuška, O. C. Zienkiewicz, J. Gago, and E.R. de A. Oliveira. *Accuracy Estimates and Adaptive Refinements in Finite Element Computations*. John Wiley & Sons Ltd., 1986.

[14] R.E. Bank. Analysis of a local a posteriori error estimate for elliptic equations. In *Accuracy Estimates and Adaptive Refinements in Finite Element Computations [13]*, chapter 7, pages 119–128.

[15] R.E. Bank and A. Weiser. Some a posteriori error estimators for elliptic partial differential equations. *Mathematics of Computation*, 44(170):283–301, 1985.

[16] S. Benzoni-Gavage and D. Serre. *Multidimensional Hyperbolic Partial Differential Equations*. Oxford University Press, 2007.

[17] K. Bottcher and R. Rannacher. Adaptive error control in solving ordinary differential equations by the discontinuous Galerkin method. Tech. report, University of Heidelberg, 1996.

[18] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral Methods: Fundamentals in Single Domains*. Springer Verlag, New York, 2006.

[19] P. Castillo. A superconvergence result for discontinuous Galerkin methods applied to elliptic problems. *Computer Methods in Applied Mechanics and Engineering*, 192:4675–4685, 2003.

[20] F. Celiker and B. Cockburn. Superconvergence of the numerical traces for discontinuous Galerkin and hybridized methods for convection-diffusion problems in one space dimension. *Math. Comp.*, 76:67–96, 2007.

[21] G. Chavent and B. Cockburn. The local projection P0-P1-discontinuous Galerkin method for scalar conservation laws. *Modél. Math. Anal. Numér*, 23:565–592, 1989.

[22] G. Chavent and G. Salzano. A finite-element method for the 1-d water flooding problem with gravity. *J. Comput. Phys*, 45:307–344, 1982.

[23] B. Cockburn. A simple introduction to error estimation for nonlinear hyperbolic conservation laws. In *Proceedings of the 1998 EPSRC Summer School in Numerical Analysis, SSCM, volume 26 of the Graduate Student's Guide for Numerical Analysis, pages 1-46*, Berlin, 1999. Springer.

[24] B. Cockburn and P. A. Gremaud. Error estimates for finite element methods for nonlinear conservation laws. *SIAM Journal on Numerical Analysis*, 33:522–554, 1996.

[25] B. Cockburn, S. Hou, and C. W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. *Math. Comp.*, 54:545–581, 1990.

[26] B. Cockburn, G. E. Karniadakis, and C. W. Shu. *Discontinuous Galerkin Methods Theory, Computation and Applications, Lecture Notes in Computational Science and Engineering*, volume 11. Springer, Berlin, 2000.

[27] B. Cockburn, S. Y. Lin, and C. W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin methods of scalar conservation laws III: One dimensional systems. *Journal of Computational Physics*, 84:90–113, 1989.

[28] B. Cockburn and C. W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws ii: General framework. *Math. Comp.*, 52:411–435, 1989.

[29] B. Cockburn and C. W. Shu. The Runge-Kutta local projection p1-discontinuous Galerkin method for scalar conservation laws. *RAIRO Modél. Math. Anal. Numér.*, 25:337–361, 1991.

[30] B. Cockburn and C. W. Shu. The local discontinuous Galerkin finite element method for convection-diffusion systems. *SIAM Journal on Numerical Analysis*, 35:2240–2463, 1998.

[31] B. Cockburn and C. W. Shu. The Runge-Kutta discontinuous Galerkin method for conservation laws v: Multidimensional systems. *Journal of Computational Physics*, 141(2):199–224, 1998.

[32] C. M. Dafermos. *Hyperbolic Conservation Laws in Continuum Physics*. Springer, 2000.

[33] M. Delfour, W. Hager, and F. Trochu. Discontinuous Galerkin methods for ordinary differential equation. *Math. Comp.*, 154:455–473, 1981.

[34] L. Demkowicz, Ph. Devloo, and J. T. Oden. On an h-type mesh refinement strategy based on minimization of interpolation errors. *Comput. Methods Appl. Mech. Engrg.*, 53:67–89, 1985.

[35] L. Demkowicz, J. T. Oden, and T. Strouboulis. Adaptive finite elements for flow problems with moving boundaries, part 1: Variational principles and a posteriori error estimates. *Comput. Methods Appl. Mech. Engrg.*, 46:217–251, 1984.

[36] L. Demkowicz, J. T. Oden, and T. Strouboulis. An adaptive p-version finite element method for transient flow problems with moving boundaries. In R.H. Gallagher, editor, *Finite elements in fluids VI*, pages 291–305. John Wiley, 1985.

[37] J. R. Dormand and P. J. Prince. A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980.

[38] A. Givental. *Linear Algebra and Differential Equations*. AMS Bookstore, 2001.

[39] G. H. Golub and C. F. Van Loan. *Matrix Computations*. JHU Press, 1996.

[40] R. Hartmann and P. Houston. Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservations laws. *SIAM J. Sci. Comput.*, 24:979–1004, 2002.

[41] N. J. Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2 edition, 2002.

[42] P. Houston, J. A. Mackenzie, E. Süli, and G. Warnecke. A posteriori error analysis for numerical approximations of Friedrichs systems. *Numerische Mathematik*, 82:433–470, 1999.

[43] P. Houston, D. Schötzau, and T. Wihler. Energy norm a posteriori error estimation of *hp*-adaptive discontinuous Galerkin methods for elliptic problems. *Math. Models Methods Appl. Sci.*, 17:33–62, 2007.

[44] C. Johnson. Adaptive finite element methods for diffusion and convection problems. *Computer Methods in Applied Mechanics and Engineering*, 82:301–322, 1990.

[45] C. Johnson and J. Hansbo. Adaptive finite element methods in computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, 101:143–181, 1992.

[46] C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comput.*, 46(173):1–26, 1986.

[47] C. Johnson and A. Szepessy. Adaptive finite element methods for conservation laws. *Commun Pure Appl. Math.*, 48:199–243, 1995.

[48] O. Karakashian and Ch. Makridakis. A space-time finite element method for the nonlinear Shrödinger equation: The discontinuous Galerkin method. Preprint #96-9, Department of Mathematics, University of Crete, 71409 Heraklion-Crete, Greece, 1996.

[49] L. Krivodonova and J. E. Flaherty. Error estimation for discontinuous Galerkin solutions of two-dimensional hyperbolic problems. *Advances in Computational Mathematics*, 19:57–71, 2003.

[50] S. N. Kružkov. First order quasilinear equations in several independent variables. *Math. U.S.S.R. Sbornik*, 10:217–243, 1970.

[51] P. Ladeveze and D. Leguillon. Error estimate procedure in the finite element method and applications. *SIAM J. Numer. Anal.*, 20:485509, 1983.

[52] M. Larson and T. Barth. A posteriori error estimation for adaptive discontinuous Galerkin approximation of hyperbolic systems. In B. Cockburn, G. E. Karniadakis, and C. W. Shu, editors, *Proc. International Symposium on Discontinuous Galerkin Methods Theory, Computation and Applications*, Berlin, 2000. Springer.

[53] P. LeSaint and P. Raviart. On a finite element method for solving the neutron transport equations. In C. de Boor, editor, *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 89–145, New York, 1974. Academic Press.

[54] Q. Lin and A.-H. Zhou. Convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *Acta Math. Sci.*, 13:207–210, 1993.

[55] J. Peraire, M. Vahdati, K. Morgan, and O. C. Zienkiewicz. Adaptive remeshing for compressible flow computations. *J. Comp. Phys.*, 72:449–466, 1987.

[56] T. E. Peterson. A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *SIAM Journal on Numerical Analysis*, 28(1):133–140, 1991.

[57] W. H. Reed and T. R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, 1973.

[58] G. Richter. An optimal-order error estimate for discontinuous Galerkin method. *Math. Comput.*, 50:75–88, 1988.

[59] B. Riviere and M. F. Wheeler. A posteriori error estimates for a discontinuous Galerkin method applied to elliptic problems. *Comput. Math. Appl.*, 46:143–163, 2003.

[60] S. Roman. *Advanced Linear Algebra*. Springer, 3 edition, 2007.

[61] L. A. Sadun. *Applied Linear Algebra; the decoupling principle*. AMS Bookstore, 2 edition, 2008.

[62] D. Schötzau and C. Schwab. An hp a-priori error analysis of the DG time-stepping method for initial value problems. *Calcolo*, 37:207–232, 2000.

[63] D. Schötzau and C. Schwab. Time discretization of parabolic problems by the hp-version of the discontinuous Galerkin finite element method. *SIAM Journal on Numerical Analysis*, 38:837–875, 2000.

[64] D. Serre. *Matrices: Theory and Applications*. Springer Verlag, New York, 2002.

[65] E. Süli. A posteriori error analysis and adaptivity for finite element approximations of hyperbolic problems. In D. Kröner, M. Ohlberger, and C. Rhode, editors, *An introduction to recent developments in theory and numerics for conservation laws, volume 5 of Lecture Notes in Computational Sciences and Engineering*, Berlin, 1999. Springer.

[66] E. Süli and P. Houston. Finite element methods for hyperbolic problems: a posteriori error analysis and adaptivity. In I. Duff and G.A. Watson, editors, *State of the Art in Numerical Analysis*, Oxford, 1997. Oxford University Press.

[67] B.A. Szabo. Estimation and control of error based on p convergence. In *Accuracy Estimates and Adaptive Refinements in Finite Element Computations [13]*, page 6170.

[68] F. Szabo. *An introduction Using Mathematica*. Academic Press, 2000.

[69] J. C. Tannehill, D. A. Anderson, and R. H. Pletcher. *Computational Fluid Mechanics and Heat Transfer*. Taylor & Francis, 2 edition, 1997.

[70] B. van Leer. Towards the ultimate conservation difference scheme, II. *J. Comput. Phys.*, 14:361–376, 1974.

[71] R. Verfürth. *A Review of a Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*. Teubner-Wiley, Teubner, 1996.

[72] O. C. Zienkiewicz and J. Z. Zhu. A simple error estimator and adaptive procedure for practical engineering analysis. *Int. J. Numer. Methods Engrg.*, 24:337357, 1987.