ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Spectral Discontinuous Galerkin Method for Hyperbolic Problems

Master Project
of
Benjamin Stamm

Directed by:
Dr. E. Burman
Prof. A. Quarteroni

Chair of Modelling and Scientific Computing, CMCS
Section of Mathematics, EPFL, Lausanne

February 18, 2005

Contents

Introduction

In this project we address the numerical approximation of hyperbolic equations and systems using the discontinuous Galerkin (DG) method in combination with higher order polynomial degrees. In short, this is called Spectral Discontinuous Galerkin (SDG) method. Our interest is to review the theoretical properties of the SDG-method, particularly for what concerns stability, convergence, dissipation and dispersion. Special emphases will be shed on the role of the two parameters, h (the grid-size) and N (the local polynomial degree). In this respect, we will carefully analyse the available theoretical results from the literature, then we extend some of them and implement several test cases with the purpose of assessing quantitatively the predicted theoretical properties.

In particular in chapter 1, the spectral discontinuous Galerkin method is studied for a time-dependent scalar transport equation. First, the SDG-approach is used for the space discretization. The presentation of the method is followed by a convergence analysis of the space discretization. The dispersive and dissipation behavior in space of the SDG-method is studied following the paper of Ainsworth [2]. Numerical tests confirm the theoretical results. Finally, the fully discrete space-time spectral discontinuous Galerkin method is presented. A convergence analysis for this method is then developed.

In chapter 2, the SDG-method is applied to a steady, linear hyperbolic system. First, two types of boundary conditions are presented. The stability of the problem is then studied for both types of boundary conditions. We present the SDG-method and develop the main result, a global convergence theorem, which is numerically confirmed by three test cases. Finally, the SDG-method for time-dependent linear hyperbolic systems is discussed briefly. A numerical example follows.

In chapter 3, we define the dual problem of a linear hyperbolic problem, develop its SDG-formulation and present a convergence result. An a posteriori estimation follows where we estimate the error of the outgoing characteristics on the boundary. Inspired by the article of Houston and Süli [11], we propose a theorem showing the convergence behavior of the estimated error. The theoretical investigations are concluded with a theorem describing the convergence behavior of the difference between the estimated error if a SDG-method is used for approximating the dual solution and the error using the exact solution. Finally three test cases illustrates the practical convergence behavior in the context of these two theorems.

Chapter 1

A Time-Dependent Scalar Transport Equation

1.1 Introduction

In this chapter, the spectral discontinuous Galerkin (SDG) method is studied in the context of a time-dependent scalar transport equation. First, the model problem is presented. Then we use the SDG method to semi discretize the problem in space. The space discretization leads to a system of ordinary differential equations (ODE) with respect to the time variable. The discretization of this system of ODE is not studied in this context and could be a subject of a future work. In the following section the convergence of the method is analysed combining the results of Houston, Schwab and Süli [10] and Burman and Ericsson [5]. The main result is Theorem 1.8 and its corollary, Corollary 1.9, proving convergence of the space discretization. Then the dispersive and dissipation behavior in space of the SDG-method is studied following the paper of Ainsworth [2]. The aim here is to understand the techniques of the proofs. The main result is Theorem 1.10. Finally, numerical tests of the dissipation and dispersion error confirm the theoretical results. Finally, the fully discrete space-time spectral discontinuous Galerkin method is presented. A convergence analysis for this method is then developed.

1.2 Model Problem

Let Ω denote an open and bounded Lipschitz polyhedral domain in $\mathbb{R}^d$, for $d = 2, 3$ and Γ its boundary. The following problem is considered:

find $u : \Omega \times (0, T) \to \mathbb{R}$ such that:

$$\begin{aligned} \partial_t u + \mu u + u_\beta &= f && \text{in } \Omega \times (0,T) \\ u &= g && \text{on } \Gamma_- \times (0,T) \\ u &= u_0 && \text{on } \Omega \times \{0\} \end{aligned} \qquad (1.2.1)$$

where $u_\beta = \beta \cdot \nabla u$ denotes the derivative in the β-direction, and $\beta :\to \mathbb{R}^d$ is a given vector function. The coefficient $\mu : \Omega \to \mathbb{R}$ is a function such that there exists $\mu_0 \in \mathbb{R}$ with $\mu(\mathbf{x}) \geq \mu_0 > 0$. We assume that $\mu \in L^\infty(\Omega)$, $\beta \in [W^{1,\infty}(\Omega)]^d$, $f \in H^1(\Omega \times (0, T))$ and $g \in L^2(\Gamma_- \times (0, T))$ where Γ_- is defined by

$$\Gamma_- = \big\{ \mathbf{x} \in \Gamma \;\big|\; \mathbf{n}(\mathbf{x}) \cdot \beta(\mathbf{x}) < 0 \big\}$$

where $\mathbf{n}(\mathbf{x})$ is the outward normal unit vector at the point $\mathbf{x}$. Analogously Γ_+ is defined by $\Gamma_+ = \{\mathbf{x} \in \Gamma \mid \mathbf{n}(\mathbf{x}) \cdot \beta(\mathbf{x}) \geq 0\}$.

1.3 Notations and Technical Results

Suppose that Ω is a bounded Lipschitz polyhedral domain in $\mathbb{R}^d$, $d \geq 1$ and τ_h a partition of Ω into elements so that $\bar{\Omega} = \cup_{K \in \tau_h} \bar{K}$ where every element is a parallelepiped. We assume that τ_h is shape-regular and that for each element K, there exists an affine transformation $F_K : \hat{K} \to K$ such that $F_K(\hat{K}) = K$ where $\hat{K}$ is the unit hypercube $(-1,1)^d$.
Let h_K be the diameter of the element K defined by $h_K = \max_{\mathbf{x},\mathbf{y} \in K} |\mathbf{x} - \mathbf{y}|$. Then h is defined by $h = \max_K h_K$.
Consider an element $K \in \tau_h$. Its boundary ∂K may be split into

$$\begin{array}{rcl} \partial K_- & = & \big\{ \mathbf{x} \in \partial K \,\big|\, \mathbf{n}(\mathbf{x}) \cdot \beta(\mathbf{x}) < 0 \big\} \\ \partial K_+ & = & \big\{ \mathbf{x} \in \partial K \,\big|\, \mathbf{n}(\mathbf{x}) \cdot \beta(\mathbf{x}) \geq 0 \big\} \end{array}$$

where $\mathbf{n}(\mathbf{x})$ denotes the outward unit normal. Consequently we have that $\partial K = \partial K_- \cup \partial K_+$. Additionally, let us define $\Gamma_-^K = \partial K \cap \Gamma_-$. We now define some limits associated with the boundary ∂K:

$$\begin{array}{rcl} v^- & = & \lim_{s \to 0^-} v\big(\mathbf{x} + s\beta(\mathbf{x})\big) \\ v^+ & = & \lim_{s \to 0^+} v\big(\mathbf{x} + s\beta(\mathbf{x})\big) \end{array}$$

$$\hat{v}(\mathbf{x}) = \begin{cases} v^+ & \text{if } \mathbf{x} \in \partial K_- \\ v^- & \text{if } \mathbf{x} \in \partial K_+ \end{cases}$$

and the jump term

$$[v] = v^+ - v^- .$$

Let $\mathcal{Q}_N(\hat{K})$ be the set of all tensor-product polynomials on $\hat{K}$ of maximum degree N in each direction. Using the affine transformation F_K for each element, this space can be extended for an arbitrary element $K \in \tau_h$:

$$\mathcal{Q}_N(K) = \big\{ v : K \to \mathbb{R} \,\big|\, v \circ F_K \in \mathcal{Q}_N(\hat{K}) \big\} .$$

Then we define the polynomial space

$$\mathcal{Q}_N(\tau_h) = \big\{ v : \Omega \to \mathbb{R} \,\big|\, v_{|K} \in \mathcal{Q}_N(K),\ \forall K \in \tau_h \big\} .$$

The symbol δ is a parameter describing the quality of the discretization and represents the couple (h, N), the mesh size h and the polynomial order N. Then we define the finite element space

$$V_\delta = \mathcal{Q}_N(\tau_h) .$$

Additionally, we define the local space

$$V_\delta^K = \mathcal{Q}_N(K) .$$

Since our finite element space V_δ will consist of discontinuous elements, it will not lie in $H^k(\Omega)$ but rather in the piecewise Sobolev space defined by

$$H^k(\tau_h) = \Big\{ v \in L^2(\Omega) \,\big|\, v_{|K} \in H^k(K),\ \ \forall K \in \tau_h \Big\}$$

for all integers $k \geq 0$. $H^k(\tau_h)$ is a Hilbert space with respect to the following scalar product:

$$(f, g)_{\tau_h, k} = \sum_{K \in \tau_h} \sum_{|\alpha| \leq k} \int_K (D^\alpha f)(D^\alpha g)$$

with the associated norm

$$\|f\|_{\tau_h,k} = \sqrt{(f,f)_{\tau_h,k}}$$

and semi-norm

$$|f|_{\tau_h,k} = \sqrt{\sum_{K\in\tau_h}\sum_{|\alpha|=k}\int_K |D^\alpha f|^2}\,.$$

Observe that all these definitions hold also for $L^2(\tau_h) = H^0(\tau_h)$ which can be defined analogously. But in either case $L^2(\tau_h) = L^2(\Omega)$. For the case $k = 0$ the index k is left out for the scalar product $(\cdot,\cdot)_{\tau_h} = (\cdot,\cdot)_{L^2(\Omega)}$ and the norm $\|\cdot\|_{\tau_h} = \|\cdot\|_{L^2(\Omega)}$.
Let us denote by P_δ the orthogonal projector in $L^2(\Omega)$ onto the finite element space V_δ. For a given $v \in L^2(\Omega)$, $P_\delta v$ is defined by

$$(v - P_\delta v, w)_{\tau_h} = 0 \qquad \forall w \in V_\delta\,.$$

Analogously, we define the orthogonal projector P_δ^K in $L^2(K)$ onto the local space V_δ^K by

$$(v - P_\delta^K v, w)_K = 0 \qquad \forall w \in V_\delta^K$$

where $(\cdot,\cdot)_K$ denotes the usual L^2-scalar product on K.
Next, we present two lemmas that will be used through the whole report. They are proven by Houston, Schwab and Süli, [10]. The first lemma is a consequence of Lemma 3.4 in [10] using the Stirling formula and the affine transformation F_K.

Lemma 1.1 (Lemma 3.4, [10], p.2140) *For any $K \in \tau_h$, let $v \in H^k(K)$ for some integer $k \geq 1$. Further, let P_δ^K be the $L^2(K)$-projection onto V_δ^K with $N_K \geq 1$; then, for any integer s, $0 \leq s \leq \min(N_K + 1, k)$, we have*

$$\|v - P_\delta^K v\|_{L^2(K)} \leq C_K(d)\frac{h_K^s}{N_K^s}|v|_{H^s(K)}$$

where C_K depends only on the spatial dimension d and the element K.

The next Lemma is an error estimate of the $L^2(K)$-projection onto V_δ^K on the boundary of the element K for functions smooth enough.

Lemma 1.2 (Lemma 3.9, [10], p.2144) *Let $K \in \tau_h$ and suppose that $v \in H^k(K)$ for some integer $k \geq 1$. Then, for any integer s, $0 \leq s \leq \min(N_K + 1, k)$ and $N_K \geq 0$, we have that*

$$\|v - P_\delta^K v\|_{L^2(\partial K)} \leq C_K^1(d)\Phi_1(s,p)h_K^{s-\frac{1}{2}}|v|_{H^s(K)}$$

where $\Phi_1(s,p) \leq C_K^2(s)(N_K + 1)^{-(s-\frac{1}{2})}$. The constant C_K^1 is only depending on d and the element K whereas C_K^2 depends only on s and K.

1.4 Semi-Discretization in Space

The model problem is first discretised in space by a SDG method which leads to a first order Ordinary Differential Equation (ODE) system with respect to the time variable. The ODE may be then solved by a Runge-Kutta method, see [17].

1.4.1 Spectral Discontinuous Galerkin Method in Space

We consider first the problem restricted to one element $K \in \mathcal{T}_h$. On K the following problem is considered:

find $u : K \times (0,T) \to \mathbb{R}$ such that

$$\begin{aligned} \partial_t u + \mu u + u_\beta &= f && \text{in } K \times (0,T) \\ u &= u^- && \text{on } \partial K_- \times (0,T) \\ u &= u_0 && \text{on } K \times \{0\} \,. \end{aligned} \tag{1.4.1}$$

Let W_K be the local functional space

$$W_K = \left\{ w \in L^2(K) \;\middle|\; \beta \cdot \nabla w \in L^2(K) \right\}$$

for all elements $K \in \mathcal{T}_h$ and let W_Ω be the global functional space

$$W_\Omega = \left\{ w \in L^2(\mathcal{T}_h) \;\middle|\; \beta \cdot \nabla w \in L^2(\Omega) \right\} .$$

Observe that functions in W_Ω have traces in $L^2(\partial\Omega; \beta \cdot \mathbf{n})$. Multiplying equation (1.4.1) by a sufficient regular test function $v \in W_K$ and integrating over the element K leads to the variational formulation of (1.4.1):

$\forall t > 0$, find $u(t) \in W_K$ such that:

$$\int_K \partial_t u(t) v + \int_K \mu u(t) v + \int_K \beta \cdot \nabla u(t) v = \int_K f(t) v \qquad \forall v \in W_K \,.$$

Using integration by parts and imposing the boundary condition in a weak sense gives:

$\forall t > 0$, find $u(t) \in W_K$ such that:

$$\begin{aligned} \big(\partial_t u(t), v\big)_K + \big(\mu u(t), v\big)_K - \big(u(t), \operatorname{div}(v\beta)\big)_K + \big(\beta \cdot \mathbf{n}\, u^-(t), v^-\big)_{\partial K_+} \\ = \big(f(t), v\big)_K + \big(|\beta \cdot \mathbf{n}|\, u^-(t), v^+\big)_{\partial K_-} \qquad \forall v \in W_K \,. \end{aligned}$$

Note that $u^-(t)$ is known on ∂K_- by the first boundary condition of (1.4.1) but not on ∂K_+ where $u^-(t)$ denotes the unknown solution on the element K. Now, a Galerkin approximation is used. This means that the space W_K is replaced by the finite dimensional space $V_\delta^K \subset W_K$:

$\forall t > 0$, find $u_{DG}(t) \in V_\delta^K$ such that:

$$\begin{aligned} &\big(\partial_t u_{DG}(t), v_\delta\big)_K + \big(\mu u_{DG}(t), v_\delta\big)_K - \big(u_{DG}(t), \operatorname{div}(v_\delta \beta)\big)_K \\ &+ \big(\beta \cdot \mathbf{n}\, u_{DG}^-(t), v_\delta^-\big)_{\partial K_+} = \big(f(t), v_\delta\big)_K + \big(|\beta \cdot \mathbf{n}|\, u_{DG}^-(t), v_\delta^+\big)_{\partial K_-} \quad \forall v_\delta \in V_\delta^K \end{aligned} \tag{1.4.2}$$

where $u_{DG}^-(\mathbf{x}, t)$ is an approximation of $u^-(\mathbf{x}, t)$ on ∂K_- for each $t \in (0,T)$. Counterintegrating by parts leads to

$\forall t > 0$, find $u_{DG}(t) \in V_\delta^K$ such that:

$$\big(\partial_t u_{DG}(t), v_\delta\big)_K + a_K\big(u_{DG}(t), v_\delta\big) = \big(f(t), v_\delta\big)_K \qquad \forall v_\delta \in V_\delta^K \tag{1.4.3}$$

where $u_{DG}^{-}(t)$ is given on the inflow boundary ∂K_{-} and $a_K : W_K \times W_K \to \mathbb{R}$ is defined by

$$a_K\big(w,v\big) = \big(\mu w + w_\beta, v\big)_K + \big(|\beta\cdot\mathbf{n}|\,[w], v^{+}\big)_{\partial K_-} \qquad \forall w,v \in W_K\ .$$

Local Approach

As basis of the finite element space V_δ^K the Legendre basis $\{\varphi_{\mathbf{i}}\}$ is chosen, where $\mathbf{i}$ is a multi-index in $\{1,..,N\}^d$. The basis $\varphi_{\mathbf{i}}$ is defined by $\varphi_{\mathbf{i}} = \hat{\varphi}_{\mathbf{i}} \circ F_K$ where $\hat{\varphi}_{\mathbf{i}}$ is the tensor product of Legendre polynomials of polynomial order i_j for each coordinate $\mathbf{x}_j$. $\hat{\varphi}_{\mathbf{i}}$ is defined by $\hat{\varphi}_{\mathbf{i}}(\mathbf{x}) = \Pi_{j=1}^{d} L_{i_j}(x_j)$ where L_{i_j} denotes the i_j-th order Legendre polynomial on $(-1,1)$.
Writing $u_{DG}(\mathbf{x},t) = \sum_{\mathbf{j}} u_{\mathbf{j}}(t)\varphi_{\mathbf{j}}(\mathbf{x})$ leads to a system of ODE with respect to the time variable:

$$M\dot{\mathbf{u}}(t) + A\mathbf{u}(t) = \mathbf{f}(t) \tag{1.4.4}$$

where

$$\begin{aligned} M_{i,j} &= (\varphi_{\mathbf{j}},\varphi_{\mathbf{i}})_K = C_{\mathbf{i}}\,\delta_{\mathbf{i},\mathbf{j}} \\ A_{i,j} &= (\varphi_{\mathbf{j},\beta} + \mu\varphi_{\mathbf{j}},\varphi_{\mathbf{i}})_K + (|\beta\cdot\mathbf{n}|\varphi_{\mathbf{j}}^{+},\varphi_{\mathbf{i}}^{+})_{\partial K_-} \\ \mathbf{f}_i &= (f,\varphi_{\mathbf{i}})_K + (|\beta\cdot\mathbf{n}|u^{-},\varphi_{\mathbf{i}}^{+})_{\partial K_-} \end{aligned}$$

Remark that i and j are single scalar indices such that there exists a bijection between i and $\mathbf{i}$ resp. j and $\mathbf{j}$.
This is the spectral discontinuous Galerkin formulation of the problem on the element K. To solve the problem on one element, one needs only to know the initial condition, the inflow boundary data and f. Given the inflow boundary data of the whole domain Ω, one can find an order of elements such that the problems can be solved element by element and for a given element K, the solution is already known on the inflow boundary.

Global Approach

The problem can also be formulated in a global way. Problem (1.2.1) is considered. The sum over (1.4.3) is taken, such that:

$\forall t > 0$, find $u_{DG}(t) \in V_\delta$ such that:

$$\begin{aligned} \big(\partial_t u_{DG}(t), v_\delta\big)_{\mathcal{T}_h} + a\big(u_{DG}(t), v_\delta\big) &= \big(f(t), v_\delta\big)_{\mathcal{T}_h} + \textstyle\sum_{K\in\mathcal{T}_h}\big(|\beta\cdot\mathbf{n}|\,g(t), v_\delta^{+}\big)_{\Gamma_-^K} \quad \forall v_\delta \in V_\delta \\ u_{DG}(0) &= P_\delta u_0 \qquad \text{for } t=0 \end{aligned} \tag{1.4.5}$$

where $a : W_\Omega \times W_\Omega \to \mathbb{R}$ is defined by

$$a\big(w,v\big) = \big(\mu w + w_\beta, v\big)_{\mathcal{T}_h} + \sum_{K\in\mathcal{T}_h}\Big\{\big(|\beta\cdot\mathbf{n}|\,[w], v^{+}\big)_{\partial K_-\setminus\Gamma} + \big(|\beta\cdot\mathbf{n}|\,w^{+}, v^{+}\big)_{\Gamma_-^K}\Big\}$$

for all $w,v \in W_\Omega$. As above, a Legendre basis $\{\Phi_{\mathbf{i},K}\}$ is chosen for the finite element space V_δ such that

$$\Phi_{\mathbf{i},K}(\mathbf{x}) = \begin{cases} \hat{\varphi}_{\mathbf{i}}(F_K^{-1}(\mathbf{x})) & \text{if } \mathbf{x}\in K \\ 0 & \text{if } \mathbf{x}\notin K \end{cases}.$$

Then, problem (1.4.5) is equivalent to solve the following ODE

$$M_\Omega\dot{\mathbf{u}}(t) + A_\Omega\mathbf{u}(t) = \mathbf{f}_\Omega(t) \tag{1.4.6}$$

where

$$
\begin{aligned}
(M_\Omega)_{i,j} &= (\Phi_{\mathbf{j}}, \Phi_{\mathbf{i}})_{\tau_h} = C_{\mathbf{i}}\, \delta_{\mathbf{i},\mathbf{j}} \\
(A_\Omega)_{i,j} &= a(\Phi_{\mathbf{j}}, \Phi_{\mathbf{i}}) \\
(\mathbf{f}_\Omega)_i &= (f, \Phi_{\mathbf{i}})_{\tau_h} + \Sigma_{K\in\tau_h} (|\beta\cdot\mathbf{n}|\, g, \Phi_{\mathbf{i}}^+)_{\Gamma_-^K}
\end{aligned}
$$

As above, i and j are mono dimensional indices such that there exists a bijection between i and $\mathbf{i}$ resp. j and $\mathbf{j}$.

1.4.2 Convergence Analysis of Space Discretisation

We analyse the discretization in space. The main result is Theorem 1.8 where we prove convergence of the space discretization combining the results of [5] and [10]. Suppose in this section that β is constant on each element.

Note that $a : W_\Omega \times W_\Omega \to \mathbb{R}$. We will now first consider the coercivity properties of $a(\cdot,\cdot)$. Then

$$
a(v,v) \;\geq\; (\mu_0 v + v_\beta, v)_{\tau_h} + \sum_{K\in\tau_h} \Big\{ (|\beta\cdot\mathbf{n}|[v], v^+)_{\partial K_-\setminus\Gamma} + (|\beta\cdot\mathbf{n}| v^+, v^+)_{\Gamma_-^K} \Big\}
$$

since $\mu(\mathbf{x}) \geq \mu_0 > 0$. Since β is constant on each element, observe that

$$
(v_\beta, v)_{\tau_h} = \sum_{K\in\tau_h} \frac{1}{2}\Big((|\beta\cdot\mathbf{n}| v^-, v^-)_{\partial K_+} - (|\beta\cdot\mathbf{n}| v^+, v^+)_{\partial K_-} \Big)
$$

by integrating by parts and consequently

$$
\begin{aligned}
a(v,v) \geq \|\mu_0^{\frac{1}{2}} v\|_{\tau_h}^2 + \sum_{K\in\tau_h} \Big\{ & (|\beta\cdot\mathbf{n}|[v], v^+)_{\partial K_-\setminus\Gamma} + (|\beta\cdot\mathbf{n}| v^+, v^+)_{\Gamma_-^K} \\
& + \frac{1}{2}(|\beta\cdot\mathbf{n}| v^-, v^-)_{\partial K_+} - \frac{1}{2}(|\beta\cdot\mathbf{n}| v^+, v^+)_{\partial K_-} \Big\}.
\end{aligned}
$$

Now, the following equality is used

$$
(a-b)a = \frac{1}{2}(a^2 + (a-b)^2 - b^2)
$$

so that

$$
(|\beta\cdot\mathbf{n}|[v], v^+)_{\partial K_-\setminus\Gamma} = \frac{1}{2}\int_{\partial K_-\setminus\Gamma} |\beta\cdot\mathbf{n}| \big((v^+)^2 + [v]^2 - (v^-)^2 \big) .
$$

Then

$$
\begin{aligned}
a(v,v) \;&\geq\; \|\mu_0^{\frac{1}{2}} v\|_{\tau_h}^2 + \frac{1}{2}\sum_{K\in\tau_h} \Big\{ \int_{\partial K_-\setminus\Gamma} |\beta\cdot\mathbf{n}| \Big((v^+)^2 + [v]^2 - (v^-)^2 \Big) \\
&\qquad + \int_{\partial K_-\setminus\Gamma} |\beta\cdot\mathbf{n}| (v^-)^2 - \int_{\partial K_-\setminus\Gamma} |\beta\cdot\mathbf{n}| (v^+)^2 + \int_{\Gamma\cap\partial K} |\beta\cdot\mathbf{n}| \hat{v}^2 \Big\} \\
&\geq\; \|\mu_0^{\frac{1}{2}} v\|_{\tau_h}^2 + \frac{1}{2}\sum_{K\in\tau_h} \Big\{ \int_{\partial K_-\setminus\Gamma} |\beta\cdot\mathbf{n}|\, [v]^2 + \int_{\Gamma\cap\partial K} |\beta\cdot\mathbf{n}|\, \hat{v}^2 \Big\} \qquad (1.4.7)
\end{aligned}
$$

This motivates us to define the following norm for W_Ω:

$$|||v|||^2 = \|\mu_0^{\frac{1}{2}} v\|^2_{\tau_h} + \frac{1}{2} \sum_{K \in \tau_h} \Big\{ \int_{\partial K_- \setminus \Gamma} |\beta \cdot \mathbf{n}| \, [v]^2 + \int_{\Gamma \cap \partial K} |\beta \cdot \mathbf{n}| \, \hat{v}^2 \Big\} \qquad \forall v \in W_\Omega \, .$$

The triple norm is a norm for $\mu_0 > 0$. In addition, we define a semi-norm $]|\cdot[|$ on W_Ω by:

$$]|v[| \;=\; \Big(\sum_K \int_{\partial K} |\beta \cdot \mathbf{n}| \, \hat{v}^2 \Big)^{\frac{1}{2}} .$$

In the following, note that C and C_K denotes generic constants taking different values. We will not pay attention to the explicit form of these constants but note that they are independent of h and N. Let for the whole analysis h and N be fixed, $u(t)$ the exact solution of (1.2.1) and $u_{DG}(t)$ the solution of (1.4.5) for any given $t \in (0,T)$. Now some intermediary results are presented, which will be used for the proof of Theorem 1.8.

Lemma 1.3 (Coercivity) *For all $v \in W_\Omega$:*

$$a(v,v) \geq |||v|||^2$$

Proof. By (1.4.7) and the definition of $|||\cdot|||$:

$$a(v,v) \geq \|\mu_0^{\frac{1}{2}} v\|^2_{\tau_h} + \frac{1}{2} \sum_{K \in \tau_h} \Big\{ \int_{\partial K_- \setminus \Gamma} |\beta \cdot \mathbf{n}| [v]^2 + \int_{\Gamma \cap \partial K} |\beta \cdot \mathbf{n}| \hat{v}^2 \Big\} = |||v|||^2$$

□ **Lemma** (1.3)

Lemma 1.4 (Galerkin Orthogonality) *If the exact solution of problem (1.2.1) satisfies $u(t) \in W_\Omega$, $\forall t \in (0,T)$, then for all $t \in (0,T)$*

$$\big(\dot{u}_{DG}(t) - \dot{u}(t), v_\delta\big)_{\tau_h} + a\big(u_{DG}(t) - u(t), v_\delta\big) = 0 \qquad \forall v_\delta \in V_\delta$$

where $u_{DG}(t)$ the solution of (1.4.5).

Proof. First, since u_{DG} is solution of (1.4.5), it satisfies for all $t \in (0,T)$:

$$\big(\dot{u}_{DG}(t), v_\delta\big)_{\tau_h} + a\big(u_{DG}(t), v_\delta\big) = \big(f(t), v_\delta\big)_\Omega + \sum_{K \in \tau_h} \big(|\beta \cdot \mathbf{n}| \, g(t), v_\delta^+\big)_{\Gamma_-^K} \qquad \forall v_\delta \in \; V_\delta \, . \tag{1.4.8}$$

Secondly, with u the exact solution, we have for all $t \in (0,T)$:

$$\begin{aligned} \big(\dot{u}(t), v_\delta\big)_{\tau_h} + \big(\mu u(t) + u_\beta(t), v_\delta\big)_{\tau_h} &= \big(f(t), v_\delta\big)_{\tau_h} \quad && \forall v_\delta \in \; V_\delta \\ \textstyle\sum_{K \in \tau_h} \big(u(t) - g(t), v_\delta^+\big)_{\Gamma_-^K} &= 0 && \forall v_\delta \in \; V_\delta \, . \end{aligned}$$

In addition, since $u(t) \in W_\Omega$ the traces in $L^2(\partial K; \beta \cdot \mathbf{n})$ are well defined and hence

$$\int_{\partial K_-} |\beta \cdot \mathbf{n}| \, [u]^2(t) = 0 \, .$$

Injecting the exact solution into formulation (1.4.5) then gives

$$\big(\dot{u}(t), v_\delta\big)_{\tau_h} + a\big(u(t), v_\delta\big) = \big(f(t), v_\delta\big)_{\tau_h} + \sum_{K \in \tau_h} \big(|\beta \cdot \mathbf{n}| \, g(t), v_\delta^+\big)_{\Gamma_-^K} \qquad \forall v_\delta \in \; V_\delta \, . \tag{1.4.9}$$

Taking the difference between (1.4.8) and (1.4.9) leads to

$$\big(\dot{u}_{DG}(t)-\dot{u}(t),v_\delta\big)_{\tau_h}+a\big(u_{DG}(t)-u(t),v_\delta\big)=0 \qquad \forall\, v_\delta\ \in V_\delta .$$

□ **Lemma** (1.4)

Next, we introduce some notations to render the proofs more readable. $\eta(t)$ is defined by $\eta(t)=u(t)-P_\delta u(t)$ and $\xi(t)$ by $\xi(t)=u_{DG}(t)-P_\delta u(t)$ where $u(t)$ is the exact solution of (1.2.1) and $u_{DG}(t)$ the SDG-approximation defined by (1.4.5). Observe that $\eta(t)\in W_\Omega$ and $\xi(t)\in V_\delta$.

Lemma 1.5 *Let $\eta(t)$ and $\xi(t)$ be defined as above. If β is constant on each element and $\mu\equiv 1$, $\forall \mathbf{x}\in\Omega$; then*

$$a\big(\eta(t),\xi(t)\big)\le 8\,|]\eta(t)[|\cdot|||\xi(t)|||\ .$$

Proof. Let $t\in(0,T)$ be fixed. The bilinear form $a(\cdot,\cdot)$ is defined by

$$a\big(\eta(t),\xi(t)\big)=\big(\eta(t)+\eta_\beta(t),\xi(t)\big)_{\tau_h}+\sum_{K\in\tau_h}\Big\{\big(|\beta\cdot\mathbf{n}|[\eta(t)],\xi(t)^+\big)_{\partial K_-\backslash\Gamma}$$
$$+\big(|\beta\cdot\mathbf{n}|\,\eta(t)^+,\xi(t)^+\big)_{\Gamma^K_-}\Big\}\quad .$$

Observe that $\xi(t)\in V_\delta$ and therefore

$$\big(\eta(t),\xi(t)\big)_K=0 \qquad \forall K\in\tau_h$$

and consequently

$$\big(\eta(t),\xi(t)\big)_{\tau_h}=0\ .$$

Integrating by parts

$$\big(\eta_\beta(t),\xi(t)\big)_K=-\big(\eta(t),\xi_\beta(t)\big)_K+\big(|\beta\cdot\mathbf{n}|\,\hat{\eta}(t),\hat{\xi}(t)\big)_{\partial K} \qquad \forall K\in\tau_h$$

But by hypothesis $\xi_\beta(t)\in V_\delta$ since β is constant elementwise and consequently $\big(\eta(t),\xi_\beta(t)\big)_K=0$. Then by reassembling the terms we may write

$$\begin{aligned}
a\big(\eta(t),\xi(t)\big) &= \sum_{K\in\tau_h}\Big\{\big(|\beta\cdot\mathbf{n}|\,[\eta(t)],\xi(t)^+\big)_{\partial K_-\backslash\Gamma}+\big(|\beta\cdot\mathbf{n}|\eta(t)^+,\xi(t)^+\big)_{\Gamma^K_-}\\
&\qquad +\big(|\beta\cdot\mathbf{n}|\,\eta(t)^-,\xi(t)^-\big)_{\partial K_+}-\big(|\beta\cdot\mathbf{n}|\,\eta(t)^+,\xi(t)^+\big)_{\partial K_-}\Big\}\\
&= \sum_{K\in\tau_h}\Big\{\big(|\beta\cdot\mathbf{n}|\,\eta(t)^-,\xi(t)^-\big)_{\partial K_+}-\big(|\beta\cdot\mathbf{n}|\,\eta(t)^-,\xi(t)^+\big)_{\partial K_-\backslash\Gamma}\Big\}\\
&= \sum_{K\in\tau_h}\Big\{\big(|\beta\cdot\mathbf{n}|\,\eta(t)^-,\xi(t)^--\xi(t)^+\big)_{\partial K_-\backslash\Gamma}\\
&\qquad +\big(|\beta\cdot\mathbf{n}|\,\eta(t)^-,\xi(t)^-\big)_{\Gamma_+\cap\partial K}\Big\}=\mathrm{I}+\mathrm{II}\ .
\end{aligned}$$

Now applying the Cauchy-Schwarz inequality leads to

$$
\begin{aligned}
\mathrm{I} = \sum_{K\in\tau_h} & \left(|\beta\cdot\mathbf{n}|\,\eta^-,\xi^- - \xi^+\right)_{\partial K_-\setminus\Gamma} \\
\leq & \left(\sum_{K\in\tau_h}\int_{\partial K_-\setminus\Gamma}|\beta\cdot\mathbf{n}|(\eta^-)^2\right)^{\frac{1}{2}}\left(\sum_{K\in\tau_h}\int_{\partial K_-\setminus\Gamma}|\beta\cdot\mathbf{n}|\,[\xi]^2\right)^{\frac{1}{2}} \\
\leq & \; 2\left(\sum_{K\in\tau_h}\int_{\partial K}|\beta\cdot\mathbf{n}|\,\hat{\eta}^2\right)^{\frac{1}{2}}|||\xi||| \\
\leq & \; 4\,|]\eta[|\,|||\xi|||
\end{aligned}
$$

by the definition of the triple norm $|||\cdot|||$ and the semi-norm $|]\cdot[|$. Additionally

$$
\begin{aligned}
\mathrm{II} = \sum_{K\in\tau_h} & \left(|\beta\cdot\mathbf{n}|\,\eta^-,\xi^-\right)_{\Gamma_+\cap\partial K} \\
\leq & \left(\sum_{K\in\tau_h}\int_{\Gamma_+\cap\partial K}|\beta\cdot\mathbf{n}|\,(\eta^-)^2\right)^{\frac{1}{2}}\left(\sum_{K\in\tau_h}\int_{\Gamma_+\cap\partial K}|\beta\cdot\mathbf{n}|\,(\xi^-)^2\right)^{\frac{1}{2}} \\
\leq & \left(\sum_{K\in\tau_h}\int_{\partial K}|\beta\cdot\mathbf{n}|\,\hat{\eta}^2\right)^{\frac{1}{2}}\left(\sum_{K\in\tau_h}\int_{\Gamma\cap\partial K}|\beta\cdot\mathbf{n}|\,\hat{\xi}^2\right)^{\frac{1}{2}} \\
\leq & \; 4\,|]\eta[|\,|||\xi|||
\end{aligned}
$$

and the result follows immediately.

□ **Lemma** (1.5)

The next lemma estimates the projection error of the triple norm $|||\cdot|||$.

Lemma 1.6 (Global Projection) *Suppose that $v\in H^k(\tau_h)$ for some integer $k\geq 1$. Then, for any integer $s, 1\leq s\leq min(N+1,k)$ and $N\geq 1$:*

$$|||v-P_\delta v|||\leq C(d,s,\beta)\frac{h^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}}|v|_{\tau_h,s}$$

where C is a positive constant depending on d, s and β.

Proof. Let $v\in H^k(\tau_h)$ and let us denote $\rho = v - P_\delta v$ for simplicity. By definition of the triple norm

$$|||\rho|||^2 = ||\mu_0^{\frac{1}{2}}\rho||^2_{\tau_h}+\frac{1}{2}\sum_{K\in\tau_h}\left\{\int_{\partial K_-\setminus\Gamma}|\beta\cdot\mathbf{n}|[\rho]^2+\int_{\Gamma\cap\partial K}|\beta\cdot\mathbf{n}|\,\hat{\rho}^2\right\}.$$

Considering each term of this sum:

- The first term is bounded by

$$
\begin{aligned}
||\mu_0^{\frac{1}{2}}\rho||^2_{\tau_h} & \leq ||\mu_0||_{L^\infty(\Omega)}||\rho||^2_{\tau_h}\leq ||\mu_0||_{L^\infty(\Omega)}\sum_{K\in\tau_h}||\rho||^2_{L^2(K)} \\
& \leq ||\mu_0||_{L^\infty(\Omega)}\sum_{K\in\tau_h}\left(C_K(d)\frac{h_K^s}{N^s}|\rho|_{H^s(K)}\right)^2
\end{aligned}
$$

  applying Lemma (1.1).

- Using the inequality $(a-b)^2 \le 2(a^2+b^2)$ leads to

$$\begin{aligned}
\frac{1}{2}\sum_{K\in\tau_h}\int_{\partial K_-\setminus\Gamma}|\beta\cdot\mathbf{n}|\,[\rho]^2 &\le \sum_{K\in\tau_h}\int_{\partial K_-\setminus\Gamma}|\beta\cdot\mathbf{n}|\big((\rho^+)^2+(\rho^-)^2\big)\\
&\le 2\sum_{K\in\tau_h}\int_{\partial K}|\beta\cdot\mathbf{n}|\,\hat{\rho}^2 \le 2\|\beta\|_{[L^\infty(\Omega)]^d}\sum_{K\in\tau_h}\|\hat{\rho}\|^2_{L^2(\partial K)}\\
&\le 2\,\|\beta\|_{[L^\infty(\Omega)]^d}\sum_{K\in\tau_h}\Big(C_K(d,s)\frac{h_K^{s-\frac{1}{2}}}{(N+1)^{s-\frac{1}{2}}}|v|_{H^s(K)}\Big)^2
\end{aligned}$$

  applying Lemma (1.2).

- The last term is bounded by

$$\begin{aligned}
\frac{1}{2}\sum_{K\in\tau_h}\int_{\partial K\cap\Gamma}|\beta\cdot\mathbf{n}|\,\hat{\rho}^2 &\le \frac{1}{2}\sum_{K\in\tau_h}\int_{\partial K}|\beta\cdot\mathbf{n}|\,\hat{\rho}^2\\
&\le \|\beta\|_{[L^\infty(\Omega)]^d}\sum_{K\in\tau_h}\|\hat{\rho}\|^2_{L^2(\partial K)}\\
&\le \|\beta\|_{[L^\infty(\Omega)]^d}\sum_{K\in\tau_h}\Big(C_K(d,s)\frac{h_K^{s-\frac{1}{2}}}{(N+1)^{s-\frac{1}{2}}}|v|_{H^s(K)}\Big)^2
\end{aligned}$$

  applying Lemma (1.2).

Considering the bounds of all three terms leads to

$$\begin{aligned}
|||\rho|||^2 &\le \sum_{K\in\tau_h}\|\mu_0\|_{L^\infty(\Omega)}C_K^2(d)\Big(\frac{h_K^s}{N^s}|v|_{H^s(K)}\Big)^2\\
&\quad+\sum_{K\in\tau_h}3\|\beta\|_{L^\infty(\Omega)}C_K^2(d,s)\Big(\frac{h_K^{s-\frac{1}{2}}}{(N+1)^{s-\frac{1}{2}}}|v|_{H^s(K)}\Big)^2\\
&\le \sum_{K\in\tau_h}\|\mu_0\|_{L^\infty(\Omega)}C_K^2(d)\Big(\frac{h_K^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}}|v|_{H^s(K)}\Big)^2\\
&\quad+\sum_{K\in\tau_h}3\|\beta\|_{L^\infty(\Omega)}C_K^2(d,s)\Big(\frac{h_K^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}}|v|_{H^s(K)}\Big)^2 .
\end{aligned}$$

Let $C_K^2(d,s,\beta) = \frac{1}{2}\max\big(\|\mu_0\|_{L^\infty(\Omega)}C_K^2(d), 3\|\beta\|_{L^\infty(\Omega)}C_K^2(d,s)\big)$. Then

$$|||\rho|||^2 \le \sum_{K\in\tau_h}C_K^2(d,s,\beta)\Big(\frac{h_K^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}}|v|_{H^s(K)}\Big)^2 .$$

Let us define $C(d,s,\beta) = \max_{K\in\tau_h}C_K(d,s,\beta)$ and conclude

$$|||\rho||| \le C(d,s,\beta)\frac{h^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}}|v|_{\tau_h,s} .$$

□ **Lemma** (1.6)

Lemma 1.7 *Let $v \in H^k(\tau_h)$ for some integer $k \geq 1$. Then, for any integer $s, 1 \leq s \leq min(N+1, k)$ and $N \geq 1$:*

$$]|v - P_\delta^K v[| \leq C(d,s,\beta)\frac{h^{s-\frac{1}{2}}}{(N+1)^{s-\frac{1}{2}}}|v|_{\tau_h,s} .$$

Proof. As in the previous proof, we denote $\rho = v - P_\delta v$ for simplicity and let $v \in H^k(\tau_h)$. Then by the definition of the semi-norm $]| \cdot [|$:

$$]|\rho[|^2 = \sum_K \int_{\partial K} |\beta \cdot \mathbf{n}|\, \hat{\rho}^2 \quad \leq \quad \sum_K \|\beta\|_{[L^\infty(\Omega)]^d} \|\hat{\rho}\|^2_{L^2(\partial K)} .$$

Applying Lemma (1.2) leads to

$$]|\rho[|^2 \quad \leq \quad \sum_K \Big(\|\beta\|^{1/2}_{[L^\infty(\Omega)]^d} C_K(d,s) \frac{h_K^{s-\frac{1}{2}}}{(N+1)^{s-\frac{1}{2}}} |v|_{H^s(K)} \Big)^2 .$$

We define $C(d,s,\beta) = \max_{K\in\tau_h} \|\beta\|^{1/2}_{[L^\infty(\Omega)]^d} C_K(d,s)$, then

$$]|\rho[| \quad \leq \quad C(d,s,\beta)\frac{h^{s-\frac{1}{2}}}{(N+1)^{s-\frac{1}{2}}}|v|_{\tau_h,s} .$$

□ **Lemma** (1.7)

Theorem 1.8 *Suppose that $u(t) \in H^k(\tau_h) \cap W_\Omega$ for some integer $k \geq 1$ and $\forall t > 0$. If β is constant on each element and if $\mu \equiv 1$, then for any integer s, $1 \leq s \leq min(N+1, k)$ and $N \geq 1$, there exists a positive constant C, only depending on d, s and β, such that*

$$\|\xi(t)\|^2_{\tau_h} + \int_0^t |||\xi(\tau)|||^2 d\tau \leq C\frac{h^{2s-1}}{(N+1)^{2s-1}}\|u\|^2_{L^2(0,t;H^s(\tau_h))} . \qquad (1.4.10)$$

Proof. Since $u_{DG}(\cdot, 0) = P_\delta u_0$, we have that $\xi(0) = u_{DG}(\cdot,0) - P_\delta u_0 = 0$.
Observe that

$$\begin{aligned}\int_0^t \big(\dot{\xi}(\tau), \xi(\tau)\big)_{\tau_h} d\tau &= \frac{1}{2}\sum_{K\in\tau_h}\int_K\int_0^t \frac{d}{d\tau}(\xi^2)d\tau = \frac{1}{2}\sum_{K\in\tau_h}\int_K \big(\xi^2(t) - \xi^2(0)\big) \\ &= \frac{1}{2}\,\|\xi(t)\|^2_{\tau_h} .\end{aligned}$$

Then, using the coercivity (Lemma 1.3) yields

$$\frac{1}{2}\,\|\xi(t)\|^2_{\tau_h} + \int_0^t |||\xi(\tau)|||^2 d\tau \leq \int_0^t \Big[\big(\dot{\xi}(\tau), \xi(\tau)\big)_{\tau_h} + a\big(\xi(\tau), \xi(\tau)\big)\Big] d\tau .$$

Using the Galerkin orthogonality (Lemma 1.4)

$$\big(\dot{\eta}(t) - \dot{\xi}(t), \xi(t)\big)_{\tau_h} + a\big(\eta(t) - \xi(t), \xi(t)\big) = 0,$$

leads to

$$\frac{1}{2}\,\|\xi(t)\|^2_{\tau_h} + \int_0^t |||\xi(\tau)|||^2 d\tau = \int_0^t \Big[\big(\dot{\eta}(\tau), \xi(\tau)\big)_{\tau_h} + a\big(\eta(\tau), \xi(\tau)\big)\Big] d\tau .$$

Additionally, observe that

$$\int_0^t \big((\dot{\eta}(\tau), \xi(\tau)\big)_{\tau_h} d\tau = \big(\eta(t), \xi(t)\big)_{\tau_h} - \int_0^t \big(\eta(\tau), \dot{\xi}(\tau)\big)_{\tau_h} d\tau = 0$$

since $\xi(0) = 0$ and note that $\xi(t), \dot{\xi}(t) \in V_\delta$. By the definition of the projection P_δ:

$$\begin{array}{rcll} \big(\eta(t), \xi(t)\big)_{\tau_h} & = & 0 & \forall t \in (0, T) \\ \big(\eta(t), \dot{\xi}(t)\big)_{\tau_h} & = & 0 & \forall t \in (0, T) \,. \end{array}$$

So we get by Lemma 1.5 and the Cauchy-Schwarz inequality followed by a Young inequality with $\epsilon = 1/16$ that

$$\begin{array}{rcl} \frac{1}{2}\|\xi(t)\|_{\tau_h}^2 + \int_0^t |||\xi(\tau)|||^2 d\tau & = & \int_0^t a\big(\eta(\tau), \xi(\tau)\big) d\tau \\ & \leq & 8 \int_0^t |]\eta(\tau)[| \; |||\xi(\tau)||| d\tau \\ & \leq & 8 \Big(\int_0^t |]\eta(\tau)[|^2 d\tau \Big)^{\frac{1}{2}} \Big(\int_0^t |||\xi(\tau)|||^2 d\tau \Big)^{\frac{1}{2}} \\ & \leq & 32 \int_0^t |]\eta(\tau)[|^2 d\tau + \frac{1}{2} \int_0^t |||\xi(\tau)|||^2 d\tau \end{array}$$

and consequently

$$\|\xi(t)\|_{\tau_h}^2 + \int_0^t |||\xi(\tau)|||^2 d\tau \leq C \int_0^t |]\eta(\tau)[|^2 d\tau$$

with $C = 64$. Applying finally Lemma 1.7 leads to the result

$$\begin{array}{rcl} \|\xi(t)\|_{\tau_h}^2 + \int_0^t |||\xi(\tau)|||^2 d\tau & \leq & C^2(d, s, \beta) \frac{h^{2s-1}}{(N+1)^{2s-1}} \int_0^t |u(\tau)|_{\tau_h, s}^2 d\tau \\ & \leq & C \frac{h^{2s-1}}{(N+1)^{2s-1}} \|u\|_{L^2(0,t;H^s(\tau_h))}^2 \,. \end{array}$$

□ **Theorem** (1.8)

Corollary 1.9 *Let us assume the same conditions than in Theorem 1.8, then*

$$\|u(t) - u_{DG}(t)\|_{\tau_h} \leq C \frac{h^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}} \Big(|u(t)|_{\tau_h, s} + \|u\|_{L^2(0,t;H^s(\tau_h))} \Big) \,.$$

Proof. By the local projection error we have that

$$\|u - P_\delta^K u\|_K \leq C_K \frac{h_K^s}{N^s} |u|_{H^s(K)}$$

so that for the whole domain

$$\|u - P_\delta u\|_{\tau_h} \leq C \frac{h^s}{N^s} |u|_{\tau_h, s}$$

and finally, owing to (1.4.10),

$$\begin{array}{rcl} \|u(t) - u_{DG}(t)\|_{\tau_h} & \leq & \|u(t) - P_\delta u(t)\|_{\tau_h} + \|\xi(t)\|_{\tau_h} \\ & \leq & C \frac{h^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}} \Big(|u(t)|_{\tau_h, s} + \|u\|_{L^2(0,t;H^s(\tau_h))} \Big) \,. \end{array}$$

□ **Corollary** (1.9)

1.5 Dispersive and Dissipative behavior of the SDG-method

1.5.1 Introduction

The dispersive and dissipative behavior of the spectral discontinuous Galerkin (SDG) method in space is analysed. All the results of this section are quoted from [2]. The aim is to understand the techniques of the proofs, especially the proof of Theorem 1.10. This result is showing that the SDG-approximation is also satisfying a Bloch-Wave condition, as the exact solution, but with a discrete wave vector. This wave vector is close to the one of the exact solution and given in an explicit form.

In [2] the general numerical flux

$$\tilde{\sigma}_\gamma(\mathbf{n}, u_{DG}) = \frac{1}{2}(\mathbf{n}\cdot\beta \pm \gamma|\mathbf{n}\cdot\beta|)\, u_{DG}^{+} + \frac{1}{2}(\mathbf{n}\cdot\beta \mp \gamma|\mathbf{n}\cdot\beta|)\, u_{DG}^{-} \qquad \text{on} \quad \partial K \cap \partial K'$$

for $\gamma \in [-1, 1]$ is used whereas here only the special case of $\gamma = 1$, that is

$$\tilde{\sigma}_1(\mathbf{n}, u_{DG}) = \frac{1}{2}(\mathbf{n}\cdot\beta \pm |\mathbf{n}\cdot\beta|)\, u_{DG}^{+} + \frac{1}{2}(\mathbf{n}\cdot\beta \mp |\mathbf{n}\cdot\beta|)\, u_{DG}^{-} \qquad \text{on} \quad \partial K \cap \partial K'$$

is treated.

1.5.2 Short introduction to Padé Approximant

The analysis relies on the Padé approximant and we first introduce some notations. The theory of Padé approximants is developed in [13]. The following results and definitions are quoted from [16].

The confluent hyperbolic function ${}_1F_1$ is defined by

$${}_1F_1(a; b; z) = 1 + \frac{a}{b}z + \frac{a}{b}\frac{a+1}{b+1}\frac{z^2}{2!} + \dots .$$

If the Pochhammer's notation $(a)_0 = 1$ and $(a)_m = a(a+1)\dots(a+m-1)$ is used, the confluent hyperbolic function can be noted as

$${}_1F_1(a; b; z) = \sum_{m=0}^{\infty} \frac{(a)_m}{(b)_m}\frac{z^m}{m!} .$$

Let p and q be two non-negative integers. The padé approximant is an approximation of e^z which is defined by

$$\left[\frac{p}{q}\right]_{e^z} = \frac{{}_1F_1(-p; -p-q; z)}{{}_1F_1(-q; -p-q; -z)} .$$

The error of the approximation takes the form

$$e^z - \left[\frac{p}{q}\right]_{e^z} = e^z \frac{z^{p+q+1}\int_0^1 e^{-tz} t^p (t-1)^q dt}{(p+q)!\; {}_1F_1(-q; -p-q; -z)} .$$

1.5.3 Dispersion and Dissipation Error Results

The dispersion and dissipation error results are presented in this section. Let us take $\Omega = \mathbb{R}^d$ and assume a rectangular grid $h\mathbb{Z}^d$ of mesh size h. Consider the problem:

find $u : \mathbb{R}^d \times (0,T) \to \mathbb{R}$ such that

$$\begin{array}{rcll} \partial_t u + u_\beta &=& 0 & \text{in } \mathbb{R}^d \times (0,T) \\ u(\mathbf{x},0) &=& Ce^{i\,\mathbf{k}\cdot\mathbf{x}} & \text{for } t=0 \end{array}$$

where β and $\mathbf{k}$ are supposed to be constant vectors. This problem is equivalent to problem (1.2.1) on $\mathbb{R}^d$ with $f=0$ and $\mu=0$. The unique solution $u(\mathbf{x},t) = Ce^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}$, with $\omega = \mathbf{k}\cdot\beta$, is satisfying the Bloch-Wave condition:

$$u(\mathbf{x}+h\mathbf{m}, t+\tau) = e^{i(h\mathbf{k}\cdot\mathbf{m}-\omega\tau)}u(\mathbf{x},t)$$

for all $\mathbf{m} \in \mathbb{Z}^d$ and $\tau \in \mathbb{R}$. The main result, Theorem 1.10, affirm that the unique solution of the SDG-scheme is also satisfying the Bloch-Wave condition, but with a discrete wave vector $\mathbf{k}_\delta$ which is close to $\mathbf{k}$. Recall that δ represents the two refinement parameters h and N, $\delta \simeq (h,N)$.

Theorem 1.10 ([2], Theorem 1, p.5) *Let $h>0$ be the mesh size of a rectangular grid $h\mathbb{Z}^d$ and $N \in \mathbb{N}$ the polynomial order of the SDG-method. If $\omega \in \mathbb{R}$ and $\mathbf{k} \in \mathbb{R}^d$ satisfy $\omega = \beta\cdot\mathbf{k}$, then there exists a solution u_{DG} of (1.4.3), satisfying:*
$\forall \mathbf{m} \in \mathbb{Z}^d$, $\tau \in \mathbb{R}$

$$u_{DG}(\mathbf{x}+h\mathbf{m}, t+\tau) = e^{i(h\mathbf{k}_\delta\cdot\mathbf{m}-\omega\tau)}u_{DG}(\mathbf{x},t) \tag{1.5.1}$$

where $e^{ihk_{\delta,l}} = [\frac{N}{N+1}]_{e^{ihk_l}}$, for all $l \in \{1,..,d\}$.

Proof. Let by hypothesis $\omega = \beta\cdot\mathbf{k}$ and let $K = (a_1,b_1)\times..\times(a_d,b_d)$ be a master element of the grid.
Let us first construct the SDG-approximation on this master element K. It will be the tensor product of solutions for each coordinate. Consider the following eigenvalue problem:

find $\phi \in \mathbb{P}_N(-1,1)$ and $\lambda \in \mathbb{C}$ such that for given $\Theta \in \mathbb{R}$

$$(\phi',v) + \big(\phi(-1) - \lambda^{-1}\phi(1)\big)v(-1) = \frac{1}{2}i\Theta(\phi,v) \qquad \forall v \in \mathbb{P}_N(-1,1) \tag{1.5.2}$$

where (u,v) denotes the L^2 inner product on $(-1,1)$. This eigenvalue problem has a non-trivial solution according to the following lemma that we state here. The proof is given later.

Lemma 1.11 ([2], Lemma 3, p.18) *Let $P_n^{(a,b)}$ be the Jacobi polynomials of order n and parameters a and b. If $\lambda = [\frac{N}{N+1}]_{e^{i\Theta}}$, then the eigenvalue problem (1.5.2) admits a non-trivial solution $\phi_N \in \mathbb{P}_N(-1,1)$ of the form:*

$$\phi_N(s) = \sum_{m=0}^{N}(i\Theta)^m\,\frac{(2N-m+1)!}{(2N+1)!}\,P_m^{(N-m,N-m+1)}(s)$$

Proof. The proof of this Lemma is given at the end of the current proof.

Fix $l \in \{1,..,d\}$ and let $(\phi_l,\ \lambda_l)$ be the non-trivial solution and the associated eigenvalue of (1.5.2) corresponding to $\Theta = hk_l$, where k_l is the l-th component of the wave vector $\mathbf{k}$. Let $x_l \in (a_l,b_l)$ and define $u_{\delta,l}(x_l) = \phi_l(s)$, where $s = \frac{2x_l-a_l-b_l}{h}$ is a variable transformation from (a_l,b_l) to $(-1,1)$.
Then, we make the ansatz

$$u_{DG}(\mathbf{x},t) = Ce^{-i\omega t}\prod_{l=1}^{d}u_{\delta,l}(x_l) \qquad \text{for } \mathbf{x} \in K$$

where $\mathbf{x} = (x_1, .., x_l, .., x_d)^T$. Now the SDG-approximation u_{DG} is defined on the master element K.
For an other element $K' \neq K$, let $h\mathbf{m} \in h\mathbb{Z}^d$ be the position vector of the centroid of K' relative to the centroid of K. If $\mathbf{y} \in K'$ such that $\mathbf{x} = \mathbf{y} - h\mathbf{m} \in K$, u_{DG} is defined on K' as follows:

$$u_{DG}(\mathbf{y}, t) = \prod_{l=1}^{d} \lambda_l^{m_l}\, u_{DG}(\mathbf{x}, t) \qquad \text{for } \mathbf{y} \in K' \tag{1.5.3}$$

where λ_l is given by $[\frac{N}{N+1}]_{e^{ihk_l}}$ (Lemma 1.11). Therefore $k_{\delta,l}$ is defined by $e^{ihk_{\delta,l}} = \lambda_l = [\frac{N}{N+1}]_{e^{ihk_l}}$.
Observe that u_{DG} satisfies (1.5.1) by construction:

$$\begin{aligned} u(\mathbf{x} + h\mathbf{m}, t+\tau) &= \prod_{l=1}^{d} \lambda_l^{m_l} u_{DG}(\mathbf{x}, t+\tau) = \prod_{l=1}^{d} \lambda_l^{m_l}\, Ce^{-i\omega(t+\tau)} \prod_{l=1}^{d} u_{\delta,l}(x_l) \\ &= e^{i(h\mathbf{m}\cdot\mathbf{k}_\delta - \omega\tau)} u_{DG}(\mathbf{x}, t) \end{aligned}$$

Finally let us prove that u_{DG} satisfies the SDG-scheme (1.4.3). The following lemma shows that (1.4.3) only has to be proved on the master element K.

Lemma 1.12 *The two following statements are equivalent:*

- *u_{DG} satisfies*
  $\big(\partial_t u_{DG} + \beta \cdot \nabla u_{DG}, v_\delta\big)_K + \big(|\beta \cdot \mathbf{n}|\, [u_{DG}], v_\delta^+\big)_{\partial K_-} = 0 \qquad \forall v_\delta \in V_\delta^K$
  on the master element K.

- *u_{DG} satisfies*
  $\big(\partial_t u_{DG} + \beta \cdot \nabla u_{DG}, w_\delta\big)_{K'} + \big(|\beta \cdot \mathbf{n}|\, [u_{DG}], w_\delta^+\big)_{\partial K'_-} = 0 \qquad \forall w_\delta \in V_\delta^{K'}$
  on any element $K' \in \tau_h$.

Remark 1.13 *The vector β is supposed to be constant, then div$(\beta v) = \beta \cdot \nabla v$.*

Proof Lemma (1.12). We obtain the result by replacing $u_{DG}(\mathbf{y}, t)$ by its definition, (1.5.3), in the second statement and performing a variable transformation. Note that the Jacobian (from the reference element K' to the master element K) is equal to one, due to the uniform mesh size. Hence

$$\begin{aligned} &\int_{K'} \big(\partial_t u_{DG}\,(\mathbf{y}, t) + \beta \cdot \nabla u_{DG}(\mathbf{y}, t)\big)\, w_\delta(\mathbf{y})\, d\mathbf{y} \\ &\qquad + \int_{\partial K'_-} |\beta \cdot \mathbf{n}|\, [u_{DG}](\mathbf{y}, t)\, w_\delta^+(\mathbf{y})\, d\mathbf{y} = 0 \quad \forall w_\delta \in V_\delta^{K'} \end{aligned}$$

$$\Leftrightarrow$$

$$\begin{aligned} \prod_{l=1}^{d} \lambda_l^{m_l} \cdot \Big[&\int_K \big(\partial_t u_{DG}(\mathbf{x}, t) + \beta \cdot \nabla u_{DG}(\mathbf{x}, t)\big)\, v_\delta(\mathbf{x})\, d\mathbf{x} \\ &+ \int_{\partial K_-} |\beta \cdot \mathbf{n}|\, [u_{DG}](\mathbf{x}, t)\, v_\delta^+(\mathbf{x})\, d\mathbf{x} \Big] = 0 \quad \forall v_\delta \in V_\delta^K \end{aligned}$$

and the result follows.

□ **Lemma** (1.12)

So let us only verify that u_{DG} satisfy the SDG-scheme (1.4.3) on the master element K. For this, fix $l \in \{1, .., d\}$ and take the eigenvalue problem (1.5.2) corresponding to $\Theta = hk_l$. Observe that by a variable transformation:

- $(\partial_s \phi_l, v) = \frac{2}{h}(\frac{h}{2}\partial_l u_{DG}, \tilde{v})_l = (\partial_l u_{DG}, \tilde{v})_l \qquad$ where $\tilde{v}(x_l) = v(s)$
- $(\phi_{x_l}, v) = \frac{2}{h}(u_{DG}, \tilde{v})_l$

where $(\cdot,\cdot)_l$ denotes the L^2 inner product on the interval (a_l, b_l), $(\cdot,\cdot)$ the L^2 inner product on $(-1,1)$ and ∂_l the partial derivative with respect to the x_l-coordinate. With this substitution of variables we get:

$$\begin{array}{rcll} (\partial_s \phi_l, v) + \big(\phi_l(-1) - \lambda_l^{-1}\phi_l(1)\big)v(-1) & = & \frac{1}{2} ihk_l(\phi_l, v) & \forall v \in \mathbb{P}_N(-1,1) \\ & \Leftrightarrow & & \\ (\partial_{x_l} u_{\delta,l}, v)_l + \big(u_{\delta,l}(a_l) - \lambda_l^{-1} u_{\delta,l}(b_l)\big)v(a_l) & = & ik_l(u_{\delta,l}, v)_l & \forall v \in \mathbb{P}_N(a_l, b_l) \end{array} \tag{1.5.4}$$

Let $\mathbf{x} \in K$, $\mathbf{m} = m\mathbf{e_l}$ and $\tau = 0$, where $\mathbf{e_l}$ is the l-th unit vector in $\mathbb{R}^d$ and $l \in \{1, .., d\}$. Then

$$\begin{array}{rcl} u_{DG}(\mathbf{x} + hm\mathbf{e}_l, t) = e^{i(hm\mathbf{k}_\delta \cdot \mathbf{e}_l)} u_{DG}(\mathbf{x}, t) & = & e^{ihmk_{\delta,l}} u_{DG}(\mathbf{x}, t) \\ & \Leftrightarrow & \\ Ce^{-i\omega t}\, u_{\delta,l}(x_l + hm) \displaystyle\prod_{j=1, j\neq l}^{d} u_{\delta,j}(x_j) & = & Ce^{-i\omega t}\lambda_l^m \displaystyle\prod_{j=1}^{d} u_{\delta,j}(x_j) \\ & \Leftrightarrow & \\ u_{\delta,l}(x_l + hm) & = & \lambda_l^{m_l} u_{\delta,l}(x_l) \qquad , \quad x_l \in (a_l, b_l) \end{array}$$

and this last term is evaluated at $x_l = b_l^-$ with $m = -1$, we obtain

$$u_{\delta,l}(b_l^- - h) = u_{\delta,l}(a_l^-) = \frac{1}{\lambda_l} u_{\delta,l}(b_l^-) \, .$$

Inserting this in (1.5.4) leads the eigenvalue problem equivalent to:

$$\big(\partial_{x_l} u_{\delta,l}, v_\delta\big)_l + \big(u_{\delta,l}(a_l^+) - u_{\delta,l}(a_l^-)\big)v_\delta(a_l^+) = ik_l\big(u_{\delta,l}, v_\delta\big)_l \quad \forall v_\delta \in \mathbb{P}_N(a_l, b_l) \, . \tag{1.5.5}$$

This result will be used later. Now, take equation (1.4.3), and let $v_\delta(\mathbf{x}) = \Pi_{l=1}^d v_{\delta,l}(x_l)$ where $v_l \in \mathbb{P}_N(a_l, b_l)$. Then, using that

- $\partial_t u_{DG} = -i\omega\, u_{DG}$
- $\beta \cdot \nabla u_{DG} = \Big(\sum_{j=1}^d \beta_j \frac{\partial_j u_{\delta,j}}{u_{\delta,j}}\Big) u_{DG}$

yields

$$\Big(\partial_t u_{DG} + \beta \cdot \nabla u_{DG}, \Pi_{l=1}^d v_{\delta,l}\Big)_K + \Big(|\beta \cdot \mathbf{n}|[u_{DG}], (\Pi_{l=1}^d v_{\delta,l})^-\Big)_{\partial K_-} = 0$$

$$\Leftrightarrow$$

$$\int_K Ce^{-i\omega t}\, \Pi_{l=1}^d v_{\delta,l}(x_l)\, \Pi_{l=1}^d u_{\delta,l}(x_l) \Big(-i\omega + \sum_{j=1}^d \beta_j \frac{\partial_j u_{\delta,j}}{u_{\delta,j}}\Big)\, d\mathbf{x} \\ + \Big(|\beta \cdot \mathbf{n}|\, Ce^{-i\omega t}\, [\Pi_{l=1}^d u_{\delta,l}], (\Pi_{l=1}^d v_{\delta,l})^-\Big)_{\partial K_-} = 0 \, .$$

Eliminating the factor $Ce^{-i\omega t}$ of u_{DG} leads to the equivalent formulation

$$\begin{aligned}
&\int_K \Pi_{l=1}^d v_{\delta,l}(x_l) u_{\delta,l}(x_l) \Big(-i\omega + \sum_{j=1}^d \beta_j \frac{\partial_j u_{\delta,j}}{u_{\delta,j}} \Big)\, d\mathbf{x} \\
&\qquad\qquad + \Big(|\beta\cdot\mathbf{n}|\, [\Pi_{l=1}^d u_{\delta,l}], (\Pi_{l=1}^d v_{\delta,l})^- \Big)_{\partial K_-} \quad = \quad 0 \\
&\qquad\qquad\qquad\qquad \Leftrightarrow \\
i\omega\, \Pi_{l=1}^d \big(v_{\delta,l}, u_{\delta,l}\big)_l \quad &= \quad \textstyle\sum_{j=1}^d \Big(\big(v_{\delta,j}, \beta_j u_{\delta,j}\big)_j\, \Pi_{l\neq j} \big(v_{\delta,l}, \partial_j u_{\delta,l}\big)_l \Big) \\
&\qquad + \Big(|\beta\cdot\mathbf{n}|\, [\Pi_{l=1}^d u_{\delta,l}], (\Pi_{l=1}^d v_{\delta,l})^- \Big)_{\partial K_-} .
\end{aligned} \tag{1.5.6}$$

For simplicity, suppose that $\beta_j > 0$, $\forall j \in \{1, .., d\}$. Then, an inflow face Γ_{j-} of K can be parameterized by

$$\Gamma_{j-} = \{\mathbf{x} \in K \mid x_j = a_j\}$$

and $\partial K_- = \cup_{j=1}^d \Gamma_{j-}$. Consider the jump $[\Pi_{l=1}^d u_{\delta,l}]$ evaluated at $\hat{\mathbf{x}} = (x_1, .., x_{l-1}, a_j, x_{l+1}, .., x_d) \in \Gamma_{j-}$ and let K' be such that $K \cap K' = \Gamma_{j-}$. Observe that

$$u_{DG}(\hat{x}^-) = u_{\delta,j}(a_j^-)\, \Pi_{l\neq j} u_{\delta,l}(x_l)$$

and

$$u_{DG}(\hat{x}^+) = u_{\delta,j}(a_j^+)\, \Pi_{l\neq j} u_{\delta,l}(x_l)$$

where in the first case u_{DG} is defined by (1.5.3):

$$u_{DG}(\mathbf{y}, t) = \lambda_j^{-1} u_{DG}(\mathbf{x}, t) = \lambda_j^{-1} \Pi_{l=1}^d u_{\delta,l}(x_j) .$$

Then

$$[\Pi_{l=1}^d u_{\delta,l}](\hat{\mathbf{x}}) = u_{DG}(\hat{x}^+) - u_{DG}(\hat{x}^-) = \Pi_{l\neq j}\, u_{\delta,l}(x_l) \big(u_{\delta,j}(a_l^+) - u_{\delta,j}(a_l^-) \big)$$

and

$$\begin{aligned}
\Big(|\beta\cdot\mathbf{n}|\, [\Pi_{l=1}^d u_{\delta,l}], (\Pi_{l=1}^d v_{\delta,l})^- \Big)_{\partial K_-} &= \sum_{j=1}^d \beta_j \int_{\Gamma_{j-}} \Pi_{l=1}^d v_{\delta,l}\, [\Pi_{l=1}^d u_{\delta,l}]\, d\mathbf{x} \\
&= \sum_{j=1}^d \beta_j\, v_{\delta,j}(a_j^+) \big(u_{\delta,j}(a_j^+) - u_{\delta,j}(a_j^-) \big)\, \Pi_{l\neq j} (u_{\delta,l}, v_{\delta,l})_l .
\end{aligned} \tag{1.5.7}$$

Using (1.5.6) and (1.5.7) we have, recalling (1.5.5),

$$\begin{aligned}
&i\omega \Pi_{l=1}^d (v_{\delta,l}, u_{\delta,l})_l \\
&\quad = \textstyle\sum_{j=1}^d \beta_j \Pi_{l\neq j} (v_{\delta,l}, u_{\delta,l})_l \Big((v_{\delta,j}, \partial_{x_j} u_{\delta,j}) + v_{\delta,j}(a_j^+) \big(u_{\delta,j}(a_j^+) - u_{\delta,j}(a_j^-) \big) \Big) \\
&\quad = \textstyle\sum_{j=1}^d \beta_j \Pi_{l\neq j} (v_{\delta,l}, u_{\delta,l})_l\; i k_j (u_{\delta,j}, v_{\delta,j}) \\
&\quad = i\, \beta\cdot\mathbf{k}\;\; \Pi_{l=1}^d (v_{\delta,l}, u_{\delta,l})_l .
\end{aligned}$$

Consequently, the condition that u_{DG} satisfies (1.4.3) is equivalent to $\omega = \beta\cdot\mathbf{k}$, which is true by hypothesis. Therefore (1.4.3) is satisfied.

□ **Theorem** (1.10)

Proof Lemma (1.11). The idea of this proof is that we give the explicit form of ϕ_N and show that under the assumption that $\lambda = [\frac{N}{N+1}]_{e^{i\Theta}}$, ϕ_N satisfies the eigenvalue problem.

Let $\mathcal{L}$ be the operator defined by $\mathcal{L}v = -\frac{1}{2}i\Theta v + v'$. Then we use the following relation for Jacobi polynomials, see $(8.961)_4$ of [9]:

$$\frac{d}{ds}P_m^{(N-m,N-m+1)}(s) = \frac{1}{2}\frac{(2N-m+2)!}{(2N-m+1)!}\,P_{m-1}^{(N-m+1,N-m+2)}(s)\,.$$

Using the explicit form of ϕ_N yields

$$\begin{aligned}
\mathcal{L}\phi_N &= -\frac{1}{2}i\Theta\Big(\sum_{m=0}^{N}(i\Theta)^m\frac{(2N-m+1)!}{(2N+1)!}\,P_m^{(N-m,N-m+1)}(s)\Big)\\
&\quad+\sum_{m=1}^{N}\frac{(i\Theta)^m}{2}\frac{(2N-m+1)!}{(2N+1)!}\frac{(2N-m+2)!}{(2N-m+1)!}\,P_{m-1}^{(N-m+1,N-m+2)}(s)\\
&= -\sum_{m=0}^{N}\frac{(i\Theta)^{m+1}}{2}\frac{(2N-m+1)!}{(2N+1)!}\,P_m^{(N-m,N-m+1)}(s)\\
&\quad+\sum_{l=0}^{N-1}\frac{(i\Theta)^{l+1}}{2}\frac{(2N-l+1)!}{(2N+1)!}\,P_l^{(N-l,N-l+1)}(s)\\
&= -\frac{(i\Theta)^{N+1}}{2}\frac{(N+1)!}{(2N+1)!}\,P_N^{(0,1)}(s)\,.
\end{aligned}$$

Let v be of the form $v(s) = (1+s)w(s)$ with $w \in \mathbb{P}_{N-1}$ and insert v in the eigenvalue problem (1.5.2). ϕ_N is a solution of the eigenvalue problem for the test function $v = (1+s)w$ if and only if

$$\big(\mathcal{L}\phi_N,(1+s)w\big)_{(-1,1)} = 0 \quad\Leftrightarrow\quad \big(P_N^{(0,1)},(1+s)w\big)_{(-1,1)} = 0\,.$$

The last equation is true without any condition on λ since by the orthogonality of $P_N^{(0,1)}$ there holds

$$\int_{-1}^{1}P_n^{(0,1)}P_m^{(0,1)}(1+s)ds = \delta_{nm}$$

Because $\mathbb{P}_N = span\big(1,(1+s)\mathbb{P}_{N-1}\big)$, it remains to prove (1.5.2) for $v=1$.
In the following, we use some properties of the Jacobi polynomials. They are not proven in the context of this work. Observe that

- Thanks to $(8.960)_2$ of [9], $P_m^{(N-m,N-m+1)}(1) = \left(\begin{array}{c} N \\ m \end{array}\right)$

  so that

$$\begin{aligned}
\phi_N(1) &= \sum_{m=0}^{N}(i\Theta)^m\frac{(2N-m+1)!}{(2N+1)!}\frac{N!}{m!(N-m)!}\\
&= \sum_{m=0}^{N}\frac{(-N)_m(i\Theta)^m}{(-2N-1)_m\,m!} = {}_1F_1(-N;-2N-1;i\Theta)\,.
\end{aligned}$$

- In addition, we use that $(P_N^{(0,1)},1) = (-1)^N\frac{2}{N+1}$. In fact $(P_N^{(1,0)},1) = \frac{2}{N+1}$ thanks to $(7.371)_4$ of [9] and then by $(8.961)_1$ of [9], $P_N^{(1,0)}(-s) = (-1)^N P_N^{(0,1)}(s)$, and we obtain

$(P_N^{(0,1)}, 1) = (-1)^N \frac{2}{N+1}$. Hence

$$(\mathcal{L}\phi_N, 1) \quad = \quad \frac{(-i\Theta)^{(N+1)} N!}{(2N+1)!} = \frac{(-N-1)_{N+1}(-i\Theta)^{N+1}}{(-2N-1)_{N+1}(N+1)!} \, .$$

- Thanks to $(8.960)_2$ and $(8.961)_1$ of [9],

$$P_m^{(N-m,N-m+1)}(-1) = (-1)^m \begin{pmatrix} N+1 \\ m \end{pmatrix}$$

so that

$$\begin{aligned} \phi(-1) &= \sum_{m=0}^{N} (i\Theta)^m \frac{(2N-m+1)!}{(2N+1)!} (-1)^m \begin{pmatrix} N+1 \\ m \end{pmatrix} \\ &= \sum_{m=0}^{N} \frac{(-i\Theta)^m}{m!} \frac{(2N-m+1)!}{(2N+1)!} \frac{(N+1)!}{(N-m)!} \\ &= \sum_{m=0}^{N} \frac{(-N-1)(-i\Theta)^m}{(-2N-1)_m \, m!} \, . \end{aligned}$$

The eigenvalue problem (1.5.2) for $v = 1$ is equivalent to

$$\begin{aligned} (\mathcal{L}\phi_N, 1) \quad &+ \quad \big(\phi_N(-1) - \frac{1}{\lambda}\phi_N(1)\big) = 0 \\ &\Leftrightarrow \\ \lambda \quad &= \quad \frac{\phi(1)}{(\mathcal{L}\phi_N, 1) + \phi_N(-1)} \, . \end{aligned}$$

Observe that

$$(\mathcal{L}\phi_N, 1) + \phi_N(-1) \quad = \quad \sum_{m=0}^{N+1} \frac{(-N-1)_m(-i\Theta)^m}{(-2N-1)_m \, m!} = {}_1F_1(-N-1; -2N-1; -i\Theta)$$

and consequently

$$\lambda = \frac{{}_1F_1(-N-; -2N-1; i\Theta)}{{}_1F_1(-N-1; -2N-1; -i\Theta)} = \Big[\frac{N}{N+1}\Big]_{e^{i\Theta}} \, .$$

Note that this λ is the same as the one proposed in the Lemma.

□ **Lemma** (1.11)

1.5.4 Small Wave Number

The evolution of the dissipative and dispersive error is analysed for a small wave number k_l under a refinement in h such that $hk_l \ll 1$. Let us denote $k = k_l$ and $k_\delta = k_{\delta,l}$ in this section. This means that we only analyse one component of the wave vector. Let ρ_N be defined by

$$\rho_N = \frac{e^{ihk} - e^{ihk_\delta}}{e^{ihk}} \, . \tag{1.5.8}$$

In this context, we use a result of the analysis of the Padé approximant. For more details, see [9] and [2]. The following lemma from [2], p. 13, is quoted:

Lemma 1.14 ([2], Corollary 1, p.13) *Let $N \in \mathbb{N}$ and suppose $\Theta \in \mathbb{R}$ is small. Then*

$$e^{i\Theta} - \left[\frac{N+1}{N}\right]_{e^{i\Theta}} = -\Theta^{2N+2}\frac{e^{i\Theta}}{2}\Big(\frac{N!}{(2N+1)!}\Big)^2\Big(1 - \frac{2i\Theta(N+1)}{(2N+1)(2N+3)} + \mathcal{O}(\Theta^2)\Big) .$$

Thanks to this result, we can state and proof the following theorem.

Theorem 1.15 ([2], Theorem 2, p.6) *Let $N \in \mathbb{N}$ and suppose $hk \ll 1$, then*

$$\rho_N \approx \frac{(hk)^{2N+2}}{2}\Big(\frac{N!}{(2N+1)!}\Big)^2\Big(1 + 2ihk\frac{N+1}{(2N+1)(2N+3)}\Big) .$$

Proof. Let ε_N be the relative error of the Padé Approximant of $e^{i\Theta}$ defined by

$$\varepsilon_N = \frac{e^{i\Theta} - [\frac{N+1}{N}]_{e^{i\Theta}}}{e^{i\Theta}} .$$

Let $*$ denotes the complex conjugate and observe that

$$\begin{aligned} (e^{i\Theta})^* &= e^{-i\Theta} \\ \left[\frac{N+1}{N}\right]^*_{e^{i\Theta}} &= \frac{{}_1F_1(-N-1;-2N-1;i\Theta)^*}{{}_1F_1(-N;-2N-1;-i\Theta)^*} \end{aligned}$$

and that

$$\begin{aligned} {}_1F_1(-N;-2N-1;-i\Theta)^* &= \sum_{m=0}^{N}\frac{(-N)_m(-i\Theta)^{m*}}{(-2N-1)_m\, m!} = \sum_{m=0}^{N}\frac{(-N)_m(i\Theta)^m}{(-2N-1)_m\, m!} \\ &= {}_1F_1(-N;-2N-1;i\Theta) \end{aligned}$$

as well as

$${}_1F_1(-N-1;-2N-1;i\Theta)^* = {}_1F_1(-N-1;-2N-1;-i\Theta) .$$

Therefore

$$\left[\frac{N+1}{N}\right]^*_{e^{i\Theta}} = \left[\frac{N+1}{N}\right]_{e^{-i\Theta}}$$

and consequently

$$\varepsilon_N^* = \frac{e^{-i\Theta} - [\frac{N+1}{N}]_{e^{-i\Theta}}}{e^{-i\Theta}} .$$

After some algebraic operations, we get that

$$\left[\frac{N+1}{N}\right]_{e^{-i\Theta}} = e^{-i\Theta}(1-\varepsilon_N^*) .$$

Introducing this in (1.5.8) leads to

$$\rho_N = \frac{e^{i\Theta} - \frac{e^{i\Theta}}{1-\varepsilon_N^*}}{e^{i\Theta}} = -\frac{\varepsilon_N^*}{1-\varepsilon_N^*} .$$

Then for small ε_N^* the following approximation holds

$$\rho_N = -\frac{\varepsilon_N^*}{1-\varepsilon_N^*} = -\sum_{k=1}^{\infty}{\varepsilon_N^*}^k = -\varepsilon_N^* + \mathcal{O}(|\varepsilon_N^*|^2)$$

and consequently

$$\rho_N \approx -\varepsilon_N^* .$$

Finally from Lemma 1.14 with $\theta = hk$ we conclude that

$$\rho_N \approx \frac{(hk)^{2N+2}}{2}\Big(\frac{N!}{(2N+1)!}\Big)^2\Big(1+2ihk\frac{N+1}{(2N+1)(2N+3)}\Big)\,.$$

□ **Theorem** (1.15)

We define the dispersion and dissipation error as follows

$$\begin{aligned}\varepsilon_{DISP} &= |\Re(hk)-\Re(hk_\delta)|\\ \varepsilon_{DISSI} &= |\Im(hk_\delta)|\end{aligned}$$

where $\Re$ and $\Im$ denotes the real and imaginary parts. Observe that $\Im(hk)=0$. Then, if $hk \ll 1$ one can write

$$\rho_N = \frac{e^{ihk}-e^{ihk_\delta}}{e^{ihk}} \approx ih(k-k_\delta) = \mathcal{I}(hk_\delta)+i\big(\mathcal{R}(hk)-\mathcal{R}(hk_\delta)\big)$$

and consequently

$$\begin{aligned}\varepsilon_{DISP} &\approx |\Im(\rho_N)|\\ \varepsilon_{DISSI} &\approx |\Re(\rho_N)|\,.\end{aligned}$$

This leads to the following corollary

Corollary 1.16 (Dispersion and Dissipation Error) *Let be $N \in \mathbb{N}$ and suppose $hk \ll 1$, then*

$$\begin{aligned}\varepsilon_{DISP} &\approx (hk)^{2N+3}\Big(\frac{N!}{(2N+1)!}\Big)^2\Big(\frac{N+1}{(2N+1)(2N+3)}\Big)\\ \varepsilon_{DISSI} &\approx \frac{(hk)^{2N+2}}{2}\Big(\frac{N!}{(2N+1)!}\Big)^2\,.\end{aligned}$$

1.5.5 Large Wave Number

We assume a large wave number k_l and a constant mesh size h. The evolution of the dispersion and dissipation error is analysed under refinement in N. As in the previous section we note $k = k_l$ and $k_\delta = k_{\delta,l}$. The next theorem shows three phases of convergence depending on N. It also gives an estimate of the thresholds between the different phases of convergence.

Theorem 1.17 ([2],Theorem 3,p.7) *Let $N \in \mathbb{N}$. As the order N is increased relative to hk, the relative error ρ_N passes through three distinct phases*

1. *if $2N+1 < hk - C(hk)^{\frac{1}{3}}$, then ρ_N oscillates but does not decay as N is increased.*

2. *if $hk-o(hk)^{\frac{1}{3}} < 2N+1 < hk+o(hk)^{\frac{1}{3}}$, then ρ_N decays algebraically at a rate $\mathcal{O}(N^{-\frac{1}{3}})$.*

3. *if $2N+1 \gg hk$, then ρ_N decays at a super-exponential rate as $N \to \infty$,*

$$\rho_N \approx \Big(1+\frac{ihk}{2N+3}\Big)\Big(\frac{ehk}{2\sqrt{(2N+1)(2N+3)}}\Big)^{2N+2}\,.$$

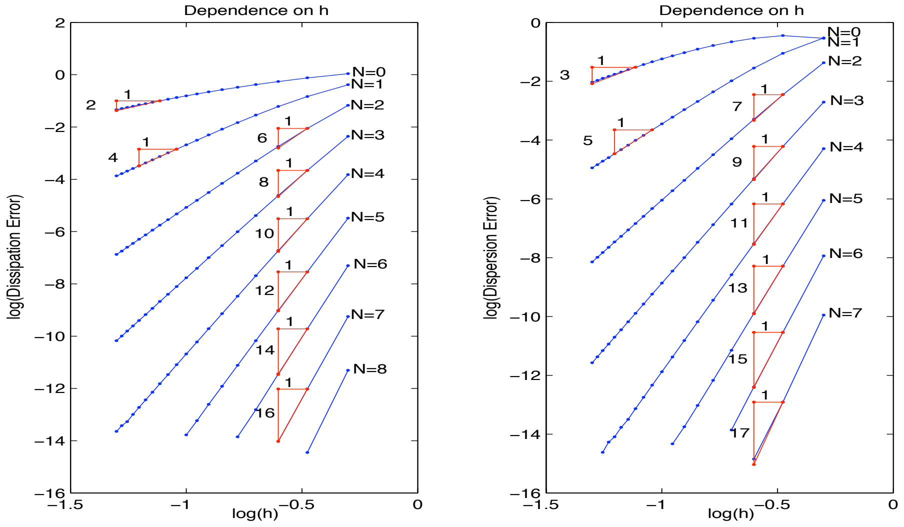

Figure 1.1: *h-refinement of the dissipation and dispersion error for different polynomial orders N and $\omega = k = 2\pi$.*

1.5.6 Numerical Results

We verify quantitatively by numerical experiments the main theoretical results reviewed in the previous section, namely Corollary 1.16 and Theorem 1.17. Let us take as domain Ω the interval $I = (0, 1)$ and the exact solution $u : I \times (0, T) \to \mathbb{C}$:

$$u(x, t) = e^{i(kx-\omega t)}$$

with corresponding initial and Dirichlet boundary conditions. Using the Bloch-wave condition of Theorem 1.10, the function u can be reduced to the initial condition:

$$u(x, t) = e^{-i\omega t}u(x, 0) = e^{-i\omega t}u_0(x) .$$

Let $t \in (0, T)$ be fixed. Solving the problem

$$\begin{array}{rcll} \partial_t u + \partial_x u & = & 0 & \text{in } I \times \{t\} \\ u & = & e^{-i\omega t} & \text{at } (0, t) \end{array}$$

is equivalent to solving the problem

$$\begin{array}{rcll} -i\omega u_0 + \partial_x u_0 & = & 0 & \text{in } I \\ u_0 & = & 1 & \text{at } x = 0 \,. \end{array}$$

Separating the real and imaginary part of this equation leads to the following linear hyperbolic system. For more details about linear hyperbolic systems, see chapter 2. Let $u_0(x) = u_0^r(x) + iu_0^i(x)$, we may then write the problem of the form:

find $\mathbf{u} = (u_0^r \; u_0^i)^T : I \to \mathbb{R}^2$ such that

$$B\mathbf{u} + A\partial_x \mathbf{u} = \mathbf{0}$$

with

$$B = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix}, \quad A = \beta \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{u} = \begin{pmatrix} u_0^r \\ u_0^i \end{pmatrix}$$

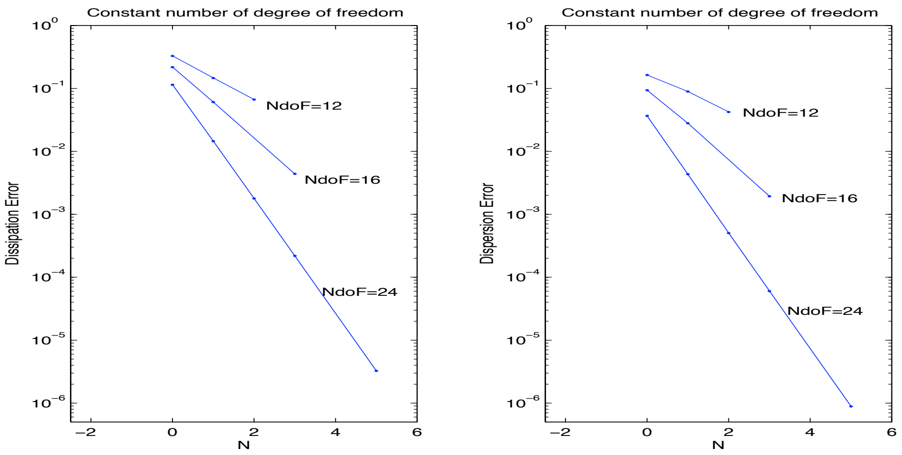

Figure 1.2: *Dissipation and dispersion error for a constant number of freedom with different polynomial order N and corresponding mesh size h where $\omega = k = 2\pi$.*

where $\beta = \omega/k$.

The dispersion and dissipation error is measured as follows. By the Bloch-wave condition of Theorem 1.10

$$e^{ik_\delta h} = \frac{u_{DG}(x+h)}{u_{DG}(x)}$$

so that

$$\begin{aligned} e^{ik_\delta h} \quad &= \quad \frac{u^r_{DG}(x+h)\, u^r_{DG}(x) + u^i_{DG}(x+h)\, u^i_{DG}(x)}{|u^r_{DG}(x)|^2 + |u^i_{DG}(x)|^2} \\ & \qquad + i\, \frac{u^i_{DG}(x+h)\, u^r_{DG}(x) + u^r_{DG}(x+h)\, u^i_{DG}(x)}{|u^r_{DG}(x)|^2 + |u^i_{DG}(x)|^2} \, . \end{aligned}$$

Then we can measure the relative error

$$\rho_N = \frac{e^{ihk} - e^{ihk_\delta}}{e^{ihk}} \, .$$

We recall the formulas for the dispersion and dissipation error

$$\begin{aligned} \varepsilon_{DISP} \quad &\approx \quad |\Im(\rho_N)| \\ \varepsilon_{DISSI} \quad &\approx \quad |\Re(\rho_N)| \, . \end{aligned}$$

Since ρ_n is now known, we can measure the dispersion and dissipation error.

Large Wave Number

For the case of a small wave number, k is chosen to be 2π, ω to 2π and consequently β to 1. Figure 1.1 shows the logarithm of the real and imaginary part of ρ_N against the logarithm of h. The different lines denotes different polynomial degrees of the space V_δ. A modelled error of $E = Ch^p$ is in the log-log diagram expressed by a straight line with slope p. The triangles in Figure 1.1 illustrates the corresponding convergence rates according to Corollary 1.16 of $2N+2$ respectively $2N+3$.

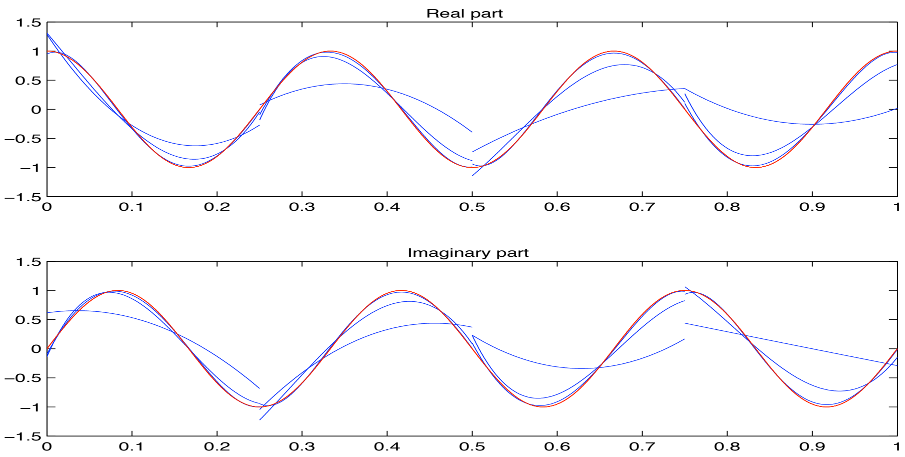

Figure 1.3: *Real and imaginary part of the SDG-approximation of order $N = 2, 3, 4$ and the exact solution $\omega = k = 4\pi$.*

Figure 1.2 shows an approximation of the dissipation and dispersion error for a constant number of freedom which takes the value 12, 16 and 24. The mesh size h is chosen such that number of freedom stays constant for the different polynomial orders N. One can clearly observe that the approximation using an increased N is much more precise than using small mesh size h.

Finally, Figure 1.3 shows the real and imaginary part of the solution of the spectral discontinuous Galerkin method for different polynomial orders $N = 2, 3, 4$ as well the exact solution with $\omega = k = 4\pi$.

Large Wave Number

In the case of a large wave number, k is chosen to be 200. Figure 1.4 shows the dissipation and dispersion error for $h = 1/8$ and $h = 1/4$ depending on the polynomial order N. One can observe the threshold of $N = \frac{hk-1}{2}$, according to Theorem 1.17, at which the convergence starts. In the first phase the relative error of ρ_N does not decay whereas in the second phase an exponential convergence can be observed. The intermediary phase can not be observed.

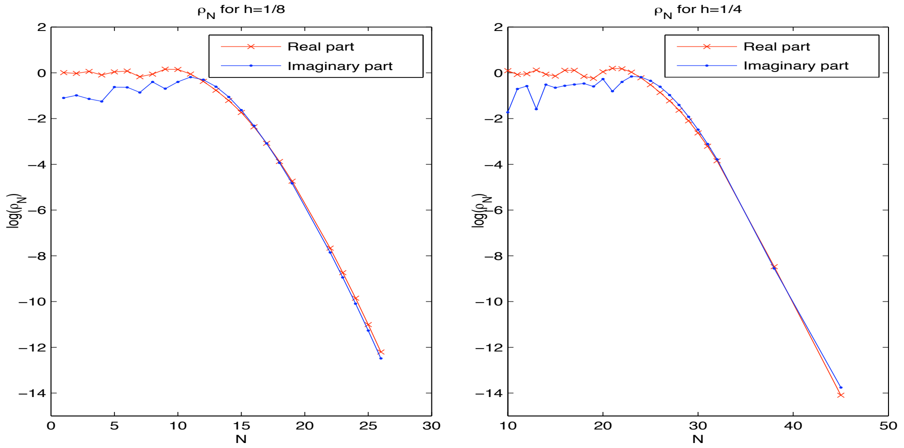

Figure 1.4: *Dissipation and dispersion error for $h = 1/8$ and $h = 1/4$ with $\omega = k = 200$. The real part represents the dissipation error and the imaginary part the dispersion error.*

1.6 Space-Time Discontinuous Galerkin Method

1.6.1 The Method

A fully space-time spectral discontinuous Galerkin scheme is considered. Starting from problem (1.2.1) we define $\tilde{\Omega} = \Omega \times (0,T)$ and $\tilde{\mathbf{x}} = (\mathbf{x},t)$. Then (1.2.1) is equivalent to

$$\begin{aligned} \tilde{\beta} \cdot \nabla u(\tilde{\mathbf{x}}) &= f(\tilde{\mathbf{x}}) \quad \text{in} \quad \tilde{\Omega} \\ u(\tilde{\mathbf{x}}) &= \tilde{g}(\tilde{\mathbf{x}}) \quad \text{on} \quad \tilde{\Gamma}_- = \big(\Omega \times \{0\}\big) \cup \big(\Gamma_- \times (0,T)\big) \end{aligned} \tag{1.6.1}$$

where ∇ denotes the space-time gradient $(\partial_{x_1}, .., \partial_{x_d}, \partial_t)$, $\tilde{\beta} = [\beta^T \; 1]^T$ and

$$\tilde{g}(\tilde{\mathbf{x}}) = \tilde{g}\big((\mathbf{x},t)\big) = \begin{cases} u_0(\mathbf{x}) & \text{on} \quad \Omega \times \{0\} \\ g(\mathbf{x},t) & \text{on} \quad \Gamma_- \times (0,T) \end{cases} .$$

Let us define $\tilde{\Gamma} = \partial\tilde{\Omega}$ and observe that $\tilde{\Gamma}_-$ can also be defined by

$$\tilde{\Gamma}_- = \big\{ \tilde{\mathbf{x}} \in \tilde{\Gamma} \;\big|\; \tilde{\beta}(\tilde{\mathbf{x}}) \cdot \tilde{\mathbf{n}}(\tilde{\mathbf{x}}) < 0 \big\}$$

where $\tilde{\mathbf{n}}(\tilde{\mathbf{x}})$ is the outward unit vector at $\tilde{\mathbf{x}} \in \partial\tilde{\Omega}$.

This problem can be considered as a steady-state problem. Observe that $\tilde{\Omega} \subset \mathbb{R}^{d+1}$ is a space-time cylinder. For constructing the partition $\tilde{\tau}_h$ of $\tilde{\Omega}$ let us first discretise the time interval $(0,T)$. Let $t^n = n\Delta t$ and $I_n = (t^n, t^{n+1})$. Fixing t^n, we consider $I_n \times \Omega$ and look for a partitioning of $I_n \times \Omega$ in elements $\tilde{K}$ which we choose to be rectangles of diameter $\tilde{h}_K$ such that $\tilde{K} \cap \big(\Omega \times \{t^n\}\big) \neq \emptyset$ and $\tilde{K} \cap \big(\Omega \times \{t^{n+1}\}\big) \neq \emptyset$. Then there exists for each element $\tilde{K}$ an affine transformation $F_{\tilde{K}}$ such that $\tilde{K}$ is the image of the master element $\hat{K}_{d+1} = (-1,1)^{d+1}$. This guarantees for each t^n an approximation in the whole domain Ω without using interpolation in time. Figure 1.5 illustrates a possible space-time grid for a one dimensional space domain Ω. Next, let us introduce the finite element space $\tilde{V}_\delta$. Let $\hat{K}_{d+1} = (-1,1)^{d+1}$ be the unit hypercube in $\mathbb{R}^{d+1}$ and let us define the two spaces

$$\tilde{V}_\delta = \big\{ v \in L^2(\tilde{\Omega}) \;\big|\; v_{|_{\tilde{K}}} \circ F_{\tilde{K}} \in \mathcal{Q}_N(\hat{K}_{d+1}), \quad \forall \tilde{K} \in \tilde{\tau}_h \big\}$$

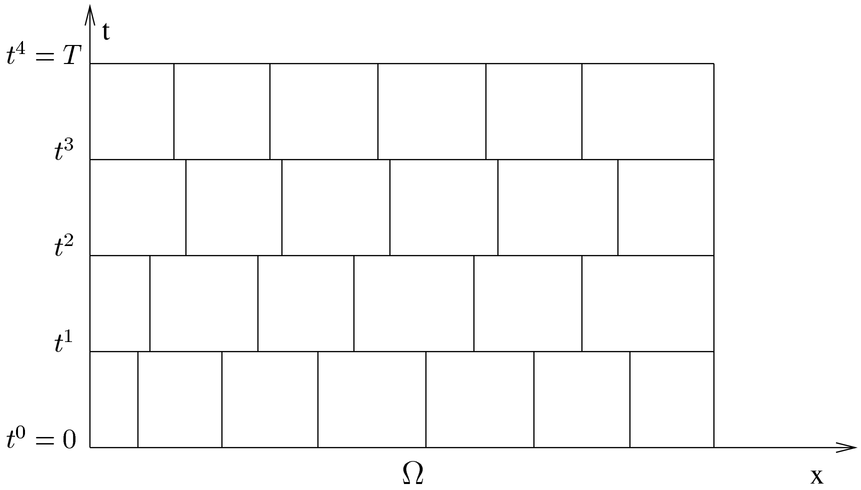

Figure 1.5: *Example of a space-time grid where every element is an affine image of the master element.*

and

$$\tilde{V}_\delta^{\tilde{K}} = \left\{ v \in L^2(\tilde{K}) \,\middle|\, v_{|_{\tilde{K}}} \circ F_{\tilde{K}} \in \mathcal{Q}_N(\hat{K}_{d+1}) \right\} .$$

Then, the SDG-scheme reads:

find $u_{DG} \in \tilde{V}_\delta$ such that

$$\tilde{a}(u_{DG}, v_\delta) = \tilde{F}(v_\delta) \qquad \forall v_\delta \in \tilde{V}_\delta \tag{1.6.2}$$

where

$$\begin{aligned} \tilde{a}(u,v) &= \left(\tilde{\beta} \cdot \nabla u, v\right)_{\tilde{\tau}_h} + \textstyle\sum_{\tilde{K} \in \tilde{\tau}_h} \left\{ \left(|\tilde{\beta} \cdot \tilde{\mathbf{n}}| \, [u], v^+\right)_{\partial \tilde{K}_- \setminus \Gamma} + \left(|\tilde{\beta} \cdot \tilde{\mathbf{n}}| \, u^+, v^+\right)_{\tilde{\Gamma}_-^{\tilde{K}}} \right\} \\ \tilde{F}(v) &= \left(f, v\right)_{\tilde{\tau}_h} + \textstyle\sum_{\tilde{K} \in \tilde{\tau}_h} \left(|\tilde{\beta} \cdot \tilde{\mathbf{n}}| \, \tilde{g}, v^+\right)_{\tilde{\Gamma}_-^{\tilde{K}}} \end{aligned}$$

for all $u, v \in W_{\tilde{\Omega}}$ and where $\tilde{\Gamma}_-^{\tilde{K}} = \tilde{\Gamma}_- \cap \tilde{K}$. Remark that $\tilde{\mathbf{n}}(\tilde{\mathbf{x}})$ is the outward normal unit vector at the point $\tilde{\mathbf{x}} = (\mathbf{x}, t) \in \partial \tilde{K}$.

1.6.2 Convergence Analysis

We present a convergence result for the space-time SDG-method. Some parts of the convergence analysis of section 1.4.2 are used some others has to be developed.
Let $W_{\tilde{\Omega}}$ denote the space

$$W_{\tilde{\Omega}} = \left\{ w \in L^2(\tilde{\Omega}) \,\middle|\, \tilde{\beta} \cdot \tilde{\nabla} w \in L^2(\tilde{\Omega}) \right\} .$$

Analogous to section 1.4.2, we define the triple norm for $W_{\tilde{\Omega}}$:

$$|||v|||^2 = ||\mu_0^{\frac{1}{2}} v||_{\tilde{\tau}_h}^2 + \frac{1}{2} \sum_{\tilde{K} \in \tilde{\tau}_h} \left\{ \int_{\partial \tilde{K}_- \setminus \partial \tilde{\Gamma}} |\tilde{\beta} \cdot \tilde{\mathbf{n}}| \, [v]^2 + \int_{\tilde{\Gamma} \cap \partial \tilde{K}} |\tilde{\beta} \cdot \mathbf{n}| \, \hat{v}^2 \right\}$$

and the semi norm:

$$|]v[| = \Big(\sum_{\tilde{K}} \int_{\partial \tilde{K}} |\tilde{\beta} \cdot \tilde{\mathbf{n}}|\, \hat{v}^2\Big)^{\frac{1}{2}} .$$

Let u denote the exact solution of (1.2.1) and u_{DG} the solution of the fully discrete time-space SDG-scheme. The Galerkin orthogonality holds also in this case:

Lemma 1.18 (Galerkin Orthogonality) *If the exact solution of problem (1.2.1) satisfies $u \in H^1(\Omega \times I)$, then*

$$\tilde{a}\big(u_{DG} - u, v_\delta\big) = 0 \qquad \forall v_\delta \in \tilde{V}_\delta$$

where $u_{DG}(t)$ the solution of (1.6.2).

Proof. First, since u_{DG} is solution of (1.6.2), it satisfies:

$$\tilde{a}\big(u_{DG}, v_\delta\big) = \big(f, v_\delta\big)_{\tilde{\tau}_h} + \sum_{\tilde{K} \in \tilde{\tau}_h} \big(|\tilde{\beta} \cdot \tilde{\mathbf{n}}|\, \tilde{g}, v_\delta^+\big)_{\tilde{\Gamma}_-^{\tilde{K}}} \qquad \forall v_\delta \in\ \tilde{V}_\delta\, . \tag{1.6.3}$$

Secondly, with u the exact solution:

$$\begin{aligned} \big(\mu u + u_{\tilde{\beta}}, v_\delta\big)_{\tilde{\tau}_h} &= \big(f, v_\delta\big)_{\tilde{\tau}_h} & \forall v_\delta \in\ \tilde{V}_\delta \\ \textstyle\sum_{\tilde{K} \in \tilde{\tau}_h} \big(u - \tilde{g}, v_\delta^+\big)_{\tilde{\Gamma}_-^{\tilde{K}}} &= 0 & \forall v_\delta \in\ \tilde{V}_\delta \end{aligned}$$

In addition, $u \in H^1(\tilde{\Omega})$ and consequently $u \in H^1(\tilde{K}), \forall \tilde{K} \in \tilde{\tau}_h$. So that the trace is well defined and hence

$$\int_{\partial \tilde{K}_-} [u]^2 = 0$$

such that

$$\frac{1}{2} \sum_{\tilde{K} \in \tilde{\tau}_h} \int_{\partial \tilde{K}_- \setminus \tilde{\Gamma}} |\tilde{\beta} \cdot \tilde{\mathbf{n}}| [u]^2 \le \frac{1}{2} \|\tilde{\beta}\|_{[L^\infty(\Omega)]^{d+1}} \sum_{\tilde{K} \in \tilde{\tau}_h} \int_{\partial \tilde{K}_-} [u]^2 = 0\, .$$

Therefore

$$\tilde{a}\big(u, v_\delta\big) = \big(f, v_\delta\big)_{\tilde{\tau}_h} + \sum_{\tilde{K} \in \tilde{\tau}_h} \big(|\tilde{\beta} \cdot \tilde{\mathbf{n}}|\, \tilde{g}, v_\delta^+\big)_{\tilde{\Gamma}_-^{\tilde{K}}} \qquad \forall v_\delta \in\ \tilde{V}_\delta \tag{1.6.4}$$

Taking the difference between (1.6.3) and (1.6.4) leads to

$$\tilde{a}\big(u_{DG} - u, v_\delta\big) = 0 \qquad \forall\ v_\delta\ \in \tilde{V}_\delta.$$

□ **Lemma** (1.18)

Let us denote by $\tilde{P}_\delta$ the orthogonal projector in $L^2(\tilde{\Omega})$ onto the finite element space $\tilde{V}_\delta$. For a given $v \in L^2(\tilde{\Omega})$, $\tilde{P}_\delta v$ is defined by

$$(v - \tilde{P}_\delta v, w)_{\tilde{\Omega}} = 0 \qquad \forall w \in \tilde{V}_\delta$$

where $(u, v)_{\tilde{\Omega}}$ denotes the L^2-scalar product in $L^2(\tilde{\Omega})$. Analogously, we define the orthogonal projector $\tilde{P}_\delta^{\tilde{K}}$ in $L^2(\tilde{K})$ onto the local space $\tilde{V}_\delta^{\tilde{K}}$ by

$$(v - \tilde{P}_\delta^{\tilde{K}} v, w)_{\tilde{K}} = 0 \qquad \forall w \in \tilde{V}_\delta^{\tilde{K}}\ .$$

Observe that Lemma 1.3, 1.5, 1.6 and 1.7 are also valid in this case, this means with the above defined norm and semi-norm for $W_{\tilde{\Omega}}$, the spaces $\tilde{V}_\delta$, $\tilde{V}_\delta^{\tilde{K}}$ and the corresponding orthogonal projectors $\tilde{P}_\delta$ and $\tilde{P}_\delta^{\tilde{K}}$. Thanks to these results, the following theorem can easily be proved.

Theorem 1.19 *Suppose that $u \in H^k(\tilde{\tau}_h) \cap H^1(\tilde{\Omega})$ for some integer $k \geq 1$. Then, for any integer s, $1 \leq s \leq \min(N+1, k)$ and $N \geq 1$, we have that*

$$|||u - u_{DG}||| \leq C(d+1, s, \tilde{\beta}) \frac{\tilde{h}^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}} |u|_{\tilde{\tau}_h, s}$$

where and $\tilde{h} = \max_{\tilde{K}}(\tilde{h}_{\tilde{K}})$

Proof. Let $\eta = \tilde{P}_\delta u - u$ and $\xi = \tilde{P}_\delta u - u_{DG}$. Then

$$|||u - u_{DG}||| \leq |||u - \tilde{P}_\delta u||| + |||\tilde{P}_\delta u - u_{DG}||| = |||\eta||| + |||\xi||| \, .$$

Firstly, using coercivity (Lemma 1.3), the Galerkin orthogonality (Lemma 1.18) and Lemma 1.5 yields

$$\begin{aligned} |||\xi|||^2 &\leq \tilde{a}(\tilde{P}_\delta u - u_{DG}, \tilde{P}_\delta u - u_{DG}) + \tilde{a}(u_{DG} - u, \tilde{P}_\delta u - u_{DG}) = \tilde{a}(\eta, \xi) \\ &\leq \,]|\eta|[\cdot |||\xi||| \end{aligned}$$

and hence

$$|||\xi||| \leq \,]|\eta|[\, .$$

Using Lemma 1.6 and 1.7 leads to

$$\begin{aligned} |||u - u_{DG}||| &\leq \textstyle\sum_{\tilde{K}} C^1_{\tilde{K}}(d+1, s, \tilde{\beta}) \frac{\tilde{h}_{\tilde{K}}^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}} |v|_{H^s(\tilde{K})} + \sum_{\tilde{K}} C^2_{\tilde{K}}(d+1, s, \tilde{\beta}) \frac{\tilde{h}_K^{s-\frac{1}{2}}}{(N+1)^{s-\frac{1}{2}}} |v|_{H^s(\tilde{K})} \\ &\leq \textstyle\sum_{\tilde{K}} C^1_{\tilde{K}}(d+1, s, \tilde{\beta}) \frac{\tilde{h}_{\tilde{K}}^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}} |v|_{H^s(\tilde{K})} + \sum_{\tilde{K}} C^2_{\tilde{K}}(d+1, s, \tilde{\beta}) \frac{\tilde{h}_K^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}} |v|_{H^s(\tilde{K})} \\ &\leq C(d+1, s, \tilde{\beta}) \frac{\tilde{h}^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}} |v|_{\tilde{\tau}_h, s} \, . \end{aligned}$$

□ **Theorem** (1.19)

1.7 Conclusion

A time-dependent scalar transport equation was considered. Compared to the standard continuous Galerkin (CG) method, or to a stabilized CG-method, the derivation and formulation of the SDG-method is slightly more technical. This complexity is due to the jumps across the interior faces of the grid. But as in the case of the CG-method, the space-discretization leads to a systems of ordinary differential equations (ODE) with respect to the time variable. This ordinary differential equation can be solved using a Runge-Kutta method, see Zhang and Shu [17], resulting in the Runge-Kutta discontinuous Galerkin method (RKDG).

A convergence result was then developed combining the results of Houston, Schwab and Süli [10] and Burman and Ericsson [5]. The convergence analysis is based on two results, error estimates of the L^2-projection on an element and also on its boundary. These two results are proved in [10]. The convergence for the space-discretization was considered with respect to the refinement parameters h (mesh size) and N (polynomial order).

The dispersion and dissipation error analysis follows the paper of Ainsworth [2]. The aim was to understand the techniques of the proofs. In a particular case of the model problem, the SDG-solution is constructed explicitly. The explicit formula of the SDG-solution allows a dispersion and dissipation error analysis either for a small or for a large wave number. We found that numerical results confirm the theory in both cases.

Another approach to fully discretize the time-dependent scalar transport equation is to use a space-time spectral discontinuous Galerkin method. Its formulation was derived as well as a convergence analysis.

Chapter 2

Linear Symmetric Hyperbolic Systems

2.1 Introduction

In this chapter, the spectral discontinuous Galerkin (SDG) method is developed for symmetric linear hyperbolic systems, also called Friedrichs' systems. Practical problems like the acoustic wave equation can be formulated in the form of a Friedrichs' systems.

First a stability result is presented. Stability is guaranteed not only for boundary conditions imposed on the incoming characteristics but also for linear combinations of the physical variables under a certain condition. Then, the SDG-method is developed for linear symmetric hyperbolic systems. The SDG-method leads to an algebraical linear system. A convergence analysis follows. The accuracy is studied with respect to the two refinement parameters h_K (the local mesh size) and N_K (the local polynomial order). Three numerical test cases are implemented with the purpose of assessing quantitatively the predicted theoretical properties. Finally, the SDG-formulation for a time-dependent linear hyperbolic system is presented, it leads to an ordinary differential equation (ODE) with respect to the time variable. A numerical example illustrates the SDG-solution where the ODE is solved by a Runge-Kutta method.

2.2 Model Problem

Suppose that Ω is a bounded Lipschitz polyhedral domain in $\mathbb{R}^d$, $d \geq 1$. We look for a m-vector function $\mathbf{u} : \Omega \to \mathbb{R}^m$ which satisfies the hyperbolic system

$$\sum_{j=1}^{d} A_j \partial_j \mathbf{u} + B\mathbf{u} = \mathbf{f} \qquad \text{in} \quad \Omega \tag{2.2.1}$$

where ∂_j denotes the partial derivative with respect to the x_j-coordinate. Assume that the matrices A_j and B are $m \times m$ real matrices in $[L^\infty(\Omega)]^{m \times m}$. The matrices A_j are supposed to be symmetric, B positive definite and $\mathbf{f}$ in $V_\Omega = [L^2(\Omega)]^m$.
Let $\mathbf{x} \in \Gamma = \partial\Omega$ and denote $\mathbf{n} = (n_1, .., n_d)^T$ the outward normal. Then let us define $D(\mathbf{x})$ as:

$$D(\mathbf{x}) = \sum_{j=1}^{d} n_j(\mathbf{x}) A_j(\mathbf{x}) \; .$$

For the sake of simplicity we note $D = D(\mathbf{x})$. $D\mathbf{u}$ plays the role of a flux across Γ in the direction $\mathbf{n}$. D is also symmetric. This implies that there exists S and Λ such that $D = S\Lambda S^T$, where Λ is the diagonal matrix having the eigenvalues of D on its diagonal. The rows of S are the right eigenvectors of D. Let us decompose Λ in Λ_+ and Λ_-, its positive and non-positive eigenvalues, such that $\Lambda = \Lambda_+ + \Lambda_-$. Then let us define

$$\begin{aligned} D^+ &= M^+ = S\,\Lambda_+\,S^T, \\ D^- &= -M^- = S\,\Lambda_-\,S^T, \\ M &= M^+ + M^- = S\,|\Lambda|\,S^T \end{aligned}$$

where $|\Lambda| = \mathrm{diag}\big(|\lambda_1|, .., |\lambda_m|\big)$. Observe that this splitting varies with the different points $\mathbf{x} \in \partial\Omega$.

2.3 Boundary Conditions

We introduce the boundary conditions for the model problem (2.2.1). Let us introduce the characteristic variables $\mathbf{z} = S^T\mathbf{u}$. According to the decomposition of D in D^+ and D^-, we split the characteristic variables in the incoming and outgoing parts $\mathbf{z} = (\mathbf{z}^+ \;\; \mathbf{z}^-)^T$. We assume for simplicity that the incoming and outgoing parts of $\mathbf{z}$ are well separated. Remark that this splitting depends on the point $\mathbf{x} \in \partial\Omega$. Suppose that $\mathbf{z}^+$ is a vector of p and $\mathbf{z}^-$ of $m-p$ components. In addition, let λ^+ and λ^- be the $p \times p$ respectively the $m-p \times m-p$ matrix blocs such that

$$\Lambda^+ = \begin{pmatrix} \lambda^+ & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \Lambda^- = \begin{pmatrix} 0 & 0 \\ 0 & \lambda^- \end{pmatrix}.$$

A natural way to impose the boundary conditions would be

$$\mathbf{z}^- = \mathbf{g} \qquad \text{on } \Gamma$$

where $\mathbf{g} : \Gamma \to \mathbb{R}^p$. This means that the p incoming characteristic variables are determined on the boundary Γ. But often, the boundary conditions are only given for some physical variables. Suppose that the physical variables are a linear combination of the variables $\mathbf{u}$, say

$$B_\Gamma \mathbf{u} = \mathbf{g} \qquad \text{on } \Gamma \tag{2.3.1}$$

where $B_\Gamma \in \mathbb{R}^{p\times m}$ and $\mathbf{g} : \Gamma \to \mathbb{R}^p$. Then we can write

$$B_\Gamma \mathbf{u} = C\mathbf{z} = \mathbf{g} \qquad \text{on } \Gamma$$

with $C = B_\Gamma S$.

2.4 Friedrichs' Theorem

Let us consider problem (2.2.1). We give necessary and sufficient assumptions for $\Sigma_{j=1}^d A_j\partial_j + B$ to be an isomorphism and we recall a fundamental wellposedness result.

We define the Hilbert space

$$W_\Omega = \big\{\mathbf{u} \in V_\Omega \;\big|\; \Sigma_{j=1}^d A_j \partial_j \mathbf{u} \in V_\Omega\big\}$$

where $V_\Omega = [L^2(\Omega)]^m$. Owing to the Riesz-Fréchet Theorem, we identify W_Ω and its dual W'_Ω as well as V_Ω with V'_Ω. The norm of W_Ω is defined by

$$||\mathbf{u}||^2_{W_\Omega} = ||\Sigma_{j=1}^d A_j \partial_j \mathbf{u}||^2_\Omega + ||\mathbf{u}||^2_\Omega$$

where $\|\cdot\|_\Omega$ denotes the usual $[L^2(\Omega)]^m$-norm. $\|\cdot\|_{W_\Omega}$ is called the graph norm. In addition, observe that $\Sigma_{j=1}^d A_j\partial_j + B \in \mathcal{L}(W_\Omega; V_\Omega)$.
Let us consider problem (2.2.1). Assume that there exists a matrix function $N \in [L^\infty(\Gamma)]^{m\times m}$ and a positive constant σ_0 such that

(F1) A_j is symmetric for $j = 1, .., d$
(F2) $B + B^T - \sum_{j=1}^d \partial_j A_j \geq 2\sigma_0 I > 0$ a.e. on Ω
(F3) $N + N^T \geq 0$ a.e. on $\partial\Omega$
(F4) $\text{Ker}(D - N) + \text{Ker}(D + N) = \mathbb{R}^m$ a.e. on $\partial\Omega$

Then, we assume the following boundary conditions:

$$(N - D)\mathbf{u}_{|\Gamma} = 0 \,.$$

We refer to [7] for the following theorem:

Theorem 2.1 (Courant-Friedrichs, [7], Theorem 5.7, p.228) *Assume (F1)-(F4) and define*

$$V = \left\{\mathbf{u} \in W_\Omega \,\middle|\, (N - D)\mathbf{u} = \mathbf{0}\right\} .$$

Then, $\Sigma_{j=1}^d A_j\partial_j + B : V \to L$ is an isomorphism.

Remark 2.2 *If we choose the particular case $N = M$, the boundary condition becomes*

$$(M - D)\mathbf{u} = 2M^-\mathbf{u} = 0$$

and therefore

$$\Lambda^-\mathbf{z} = 0 \qquad \Rightarrow \qquad \mathbf{z}^- = 0$$

such that the boundary conditions are imposed on the incoming characteristics.
Observe that we assume that the matrices A_j are symmetric and thus the first condition (F1) is satisfied. Condition (F2) has still to be assumed whereas condition (F3) is satisfied with this particular choice of $N = M$. For the last condition (F4), observe that $D - M = -2M^-$ and $D + M = 2M^+$ and hence (F4) is equivalent to

$$\mathit{Ker}(M^-) + \mathit{Ker}(M^+) = \mathbb{R}^m$$

which is satisfied by the definition of M^- and M^+.

2.5 Stability

In this section, a stability result for the Friedrichs' system is presented. First, let us split S in the parts of the eigenvectors corresponding to the positive and non-positive eigenvalues such that $\left(S^+ \; S^-\right)$, where S^+ are the columns of S corresponding to the positive eigenvalues of D and S^- the columns corresponding to the negative ones. Then, associated with the matrix B_Γ of the boundary term (2.3.1) let us define $C^+ = B_\Gamma S^+ \in \mathbb{R}^{p\times(m-p)}$ and $C^- = B_\Gamma S^- \in \mathbb{R}^{p\times p}$ and finally $\bar{C} = -(C^-)^{-1}C^+ \in \mathbb{R}^{p\times p}$.

Lemma 2.3 *Assume that the matrices A_j are constant and symmetric, B positive definite.*

- *If the boundary conditions are imposed on the incoming characteristics, that is $\mathbf{z}^- = \mathbf{g}$, then there exist two constants $\mu_0 > 0$ and $\mu_1 > 0$ such that*

$$\mu_1\|\mathbf{z}^+\|^2_{\partial\Omega} + \mu_0\|\mathbf{u}\|^2_\Omega \leq \frac{1}{\mu_0}\|\mathbf{f}\|^2_\Omega + (\mathbf{g}, |\lambda^-|\mathbf{g})_{\partial\Omega} \,.$$

- *If general boundary conditions are imposed, that is* $B_\Gamma \mathbf{u} = \mathbf{g}$ *and if the matrix* $\Lambda^+ + \bar{C}^T \Lambda^- \bar{C}$ *is positive definite, then there exists two constants* $\mu_0 > 0$ *and* $\mu_1 > 0$ *such that*

$$\frac{\mu_1}{2}\|\mathbf{z}^+\|^2_{\partial\Omega} + \mu_0\|\mathbf{u}\|^2_{\Omega} \leq \frac{1}{\mu_0}\|\mathbf{f}\|^2_{\Omega} + \big(\mathbf{g}_C, |\lambda^-|\mathbf{g}_C\big)_{\partial\Omega} + \frac{1}{\mu_1}\big\||\lambda^-|\mathbf{g}_C\big\|^2_{\partial\Omega} .$$

Proof. First, multiply equation (2.2.1) by $\mathbf{u}$ and integrate over the domain Ω leads to

$$\big(\Sigma_{j=1}^d A_j \partial_j \mathbf{u}, \mathbf{u}\big)_\Omega + \big(B\mathbf{u}, \mathbf{u}\big)_\Omega = \big(\mathbf{f}, \mathbf{u}\big)_\Omega .$$

Using the following integration by parts

$$\big(\Sigma_{j=1}^d A_j \partial_j \mathbf{u}, \mathbf{u}\big)_\Omega = \frac{1}{2}\big(D\mathbf{u}, \mathbf{u}\big)_{\partial\Omega}$$

and writing the system in terms of the characteristic variables $\mathbf{z} = S^T\mathbf{u}$ yields

$$\frac{1}{2}\big(\Lambda\mathbf{z}, \mathbf{z}\big)_{\partial\Omega} + \big(\bar{B}\mathbf{z}, \mathbf{z}\big)_\Omega = \big(\bar{\mathbf{f}}, \mathbf{z}\big)_\Omega$$

where $\bar{B} = S^T B S$ and $\bar{\mathbf{f}} = S^T\mathbf{f}$. Splitting Λ into $\Lambda^+ + \Lambda^-$ leads to

$$\frac{1}{2}\big(\Lambda^+\mathbf{z}, \mathbf{z}\big)_{\partial\Omega} + \big(\bar{B}\mathbf{z}, \mathbf{z}\big)_\Omega = \big(\bar{\mathbf{f}}, \mathbf{z}\big)_\Omega - \frac{1}{2}\big(\Lambda^-\mathbf{z}, \mathbf{z}\big)_{\partial\Omega} .$$

Additionally, we have that

$$\mathbf{z}^T \Lambda^+ \mathbf{z} = (\mathbf{z}^+)^T \, \lambda^+ \, \mathbf{z}^+$$

and analogously

$$\mathbf{z}^T \Lambda^- \mathbf{z} = (\mathbf{z}^-)^T \, \lambda^- \, \mathbf{z}^- .$$

This leads to

$$\frac{1}{2}\big(\lambda^+\mathbf{z}^+, \mathbf{z}^+\big)_{\partial\Omega} + \big(\bar{B}\mathbf{z}, \mathbf{z}\big)_\Omega = \big(\bar{\mathbf{f}}, \mathbf{z}\big)_\Omega - \frac{1}{2}\big(\lambda^-\mathbf{z}^-, \mathbf{z}^-\big)_{\partial\Omega} . \qquad (2.5.1)$$

First case:
Imposing the boundary conditions on the incoming characteristics, $\mathbf{z}^- = \mathbf{g}$, yields

$$\frac{1}{2}\big(\lambda^+\mathbf{z}^+, \mathbf{z}^+\big)_{\partial\Omega} + \big(\bar{B}\mathbf{z}, \mathbf{z}\big)_\Omega = \big(\mathbf{f}, S\mathbf{z}\big)_\Omega - \frac{1}{2}\big(\lambda^-\mathbf{g}, \mathbf{g}\big)_{\partial\Omega}$$

Since B and λ^+ are both positive definite there exists two constants $\mu_0 > 0$ and $\mu_1 > 0$ such that

$$\frac{\mu_1}{2}\|\mathbf{z}^+\|^2_{\partial\Omega} + \mu_0\|S\mathbf{z}\|^2_\Omega \leq \|\mathbf{f}\|_\Omega \, \|S\mathbf{z}\|_\Omega + \frac{1}{2}\big(\mathbf{g}, |\lambda^-|\mathbf{g}\big)_{\partial\Omega}$$

where $|\lambda^-|$ denotes $|\lambda^-| = diag(|\lambda_1^-|, .., |\lambda_{m-p}^-|)$. Applying a Young equality with $\varepsilon = \frac{\mu_0}{2}$ yields

$$\frac{\mu_1}{2}\|\mathbf{z}^+\|^2_{\partial\Omega} + \frac{\mu_0}{2}\|S\mathbf{z}\|^2_\Omega \leq \frac{1}{2\mu_0}\|\mathbf{f}\|^2_\Omega + \frac{1}{2}\big(\mathbf{g}, |\lambda^-|\mathbf{g}\big)_{\partial\Omega}$$

The result follows by multiplying the last inequality by 2 and observing that $S\mathbf{z} = \mathbf{u}$.

Second case:
Consider the boundary condition $B_\Gamma\mathbf{u} = \mathbf{g}$ and develop

$$\mathbf{g} = B_\Gamma\mathbf{u} = B_\Gamma S\mathbf{z} = B_\Gamma\big[S^+ \; S^-\big]\begin{bmatrix} \mathbf{z}^+ \\ \mathbf{z}^- \end{bmatrix} = B_\Gamma S^+\mathbf{z}^+ + B_\Gamma S^-\mathbf{z}^- = C^+\mathbf{z}^+ + C^-\mathbf{z}^-$$

and hence

$$\mathbf{z}^- = -(C^-)^{-1}C^+\mathbf{z}^+ + (C^-)^{-1}\mathbf{g} = \bar{C}\mathbf{z}^+ + \mathbf{g}_C$$

where $\bar{C} = -(C^-)^{-1}C^+ \in \mathbb{R}^{(m-p)\times p}$ and $\mathbf{g}_C = (C^-)^{-1}\mathbf{g}$. Introducing this into (2.5.1) leads to

$$\big(\mathbf{z}^+, \lambda^+\mathbf{z}^+\big)_{\partial\Omega} + 2\big(\mathbf{z}, \bar{B}\mathbf{z}\big)_{\Omega} = 2\big(\bar{\mathbf{f}}, \mathbf{z}\big)_{\Omega} - \big(\bar{C}\mathbf{z}^+ + \mathbf{g}_C, \lambda^-(\bar{C}\mathbf{z}^+ + \mathbf{g}_C)\big)_{\partial\Omega}$$

and

$$\big(\mathbf{z}^+, (\lambda^+ + \bar{C}^T\lambda^-\bar{C})\mathbf{z}^+\big)_{\partial\Omega} + 2\big(\mathbf{z}, \bar{B}\mathbf{z}\big)_{\Omega} = 2\big(\bar{\mathbf{f}}, \mathbf{z}\big)_{\Omega} + \big(\mathbf{g}_C, |\lambda^-|\mathbf{g}_C\big)_{\partial\Omega} + 2\big(\mathbf{g}_C, |\lambda^-|\bar{C}\mathbf{z}^+\big)_{\partial\Omega}$$

since $|\lambda^-|$ is a symmetric matrix. Then using that B and $\Lambda^+ + \bar{C}^T\Lambda^-\bar{C}$ are positive definite, there exists some constants $\mu_0 > 0$ and $\mu_1 > 0$ such that

$$\mu_1\|\mathbf{z}^+\|^2_{\partial\Omega} + 2\mu_0\|S\mathbf{z}\|^2_{\Omega} \le 2\|\mathbf{f}\|_{\Omega}\,\|S\mathbf{z}\|_{\Omega} + \big(\mathbf{g}_C, |\lambda^-|\mathbf{g}_C\big)_{\partial\Omega} + 2\big\||\lambda^-|\mathbf{g}_C\big\|_{\partial\Omega}\,\|\mathbf{z}^+\|_{\partial\Omega}\,.$$

Applying two times a Young inequality yields

$$\frac{\mu_1}{2}\|\mathbf{z}^+\|^2_{\partial\Omega} + \mu_0\|S\mathbf{z}\|^2_{\Omega} \le \frac{1}{\mu_0}\|\mathbf{f}\|^2_{\Omega} + \big(\mathbf{g}_C, |\lambda^-|\mathbf{g}_C\big)_{\partial\Omega} + \frac{1}{\mu_1}\big\||\lambda^-|\mathbf{g}_C\big\|^2_{\partial\Omega}\,.$$

Observe that $S\mathbf{z} = \mathbf{u}$ and therefore the result.

□ **Lemma** (2.3)

2.6 Notations and Technical Results

We present some notations, definitions and technical results which will be used through chapter 2.

Suppose that Ω is a bounded Lipschitz polyhedral domain in $\mathbb{R}^d$, $d \ge 1$ and τ_h a partition of Ω into elements such that $\bar{\Omega} = \cup_{K\in\tau_h}\bar{K}$. Assume that each element is a parallelepiped and that τ_h is shape-regular. Suppose also that for each element $K \in \tau_h$, there exists an affine transformation $F_K : \hat{K} \to K$ such that $F_K(\hat{K}) = K$ where $\hat{K}$ is the unit hypercube $(-1,1)^d$.
In the context of the SDG-method, we extend the definition of D to the set $\cup_{K\in\tau_h}\partial K$.
The vector $\mathbf{h} = (h_{K_1}, h_{K_2}, ..) \in \mathbb{R}^{card(\tau_h)}$ denotes the vector of the diameters of the elements. Let $K \in \tau_h$ be a fixed element, then the diameter h_K of the element K is defined by $h_K = \max_{\mathbf{x},\mathbf{y}\in K}|\mathbf{x}-\mathbf{y}|$. The scalar quantity h is defined by $h = \max_K h_K$.
Let $\mathcal{Q}_N(\hat{K})$ be the set of all tensor-product polynomials on $\hat{K}$ of maximum degree N in each coordinate and let $\mathbf{N} \in \mathbb{N}^{card(\tau_h)}$ be the vector of all polynomial orders varying from element to element. Using the affine transformation F_K for each element, this space can be extended for an arbitrary element $K \in \tau_h$:

$$\mathcal{Q}_N(K) = \big\{\mathbf{v} : K \to \mathbb{R}^m \;\big|\; \mathbf{v}\circ F_K \in [\mathcal{Q}_N(\hat{K})]^m\big\}\,.$$

Then we define the polynomial space

$$\mathcal{Q}_{\mathbf{N}}(\tau_h) = \big\{\mathbf{v} : \Omega \to \mathbb{R}^m \;\big|\; \mathbf{v}_{|_K} \in \mathcal{Q}_{N_K}(K), \quad \forall K \in \tau_h\big\}$$

The parameter δ describs the quality of discretization and represents the couple $(\mathbf{h}, \mathbf{N})$. Then we define the finite element space

$$V_\delta = \mathcal{Q}_{\mathbf{N}}(\tau_h)\,.$$

Additionally, let us define the local space

$$V_\delta^K = \mathcal{Q}_{N_K}(K)\,.$$

Since our finite element space V_δ will consist of discontinuous elements, it will not lie in $[H^k(\Omega)]^m$ but rather in the piecewise Sobolev space defined by

$$\mathbf{H^k}(\tau_h) = \left\{\mathbf{v} \in [L^2(\Omega)]^m \;\middle|\; \mathbf{v}_{|_K} \in [H^{k_K}(K)]^m, \quad \forall K \in \tau_h\right\}$$

where $\mathbf{k} \in \mathbb{N}^{card(\tau_h)}$ denotes the vector of the regularities for each element $K \in \tau_h$. $\mathbf{H^k}(\tau_h)$ is a Hilbert space with respect to the following scalar product:

$$(\mathbf{f},\mathbf{g})_{\tau_h,\mathbf{k}} = \sum_{K\in\tau_h}\sum_{|\alpha|\le k_K}\int_K (D^\alpha \mathbf{f})\cdot(D^\alpha \mathbf{g})$$

with associated norm

$$\|\mathbf{f}\|_{\tau_h,\mathbf{k}} = \sqrt{(\mathbf{f},\mathbf{f})_{\tau_h,\mathbf{k}}}$$

and semi-norm

$$|\mathbf{f}|_{\tau_h,\mathbf{k}} = \sqrt{\sum_{K\in\tau_h}\sum_{|\alpha|= k_K}\int_K (D^\alpha \mathbf{f})\cdot(D^\alpha \mathbf{f})}\,.$$

Observe that all these definitions holds also for $[L^2(\tau_h)]^m = \mathbf{H^0}(\tau_h)$ which can be defined analogously. But in either case $[L^2(\Omega)]^m = [L^2(\tau_h)]^m$. In the case $k=0$ the index k is left out for the scalar product $(\cdot,\cdot)_{\tau_h} = (\cdot,\cdot)_{[L^2(\Omega)]^m}$ and the norm $\|\cdot\|_{\tau_h} = \|\cdot\|_{[L^2(\Omega)]^m}$.
Additionally, for sake of simplicity, let us denote $|\cdot|_{K,s}$ the Sobolev semi-norm $|\cdot|_{[H^s(K)]^m}$ and $\|\cdot\|_K, \|\cdot\|_{\partial K}$ the usual $[L^2(K)]^m$ resp. $[L^2(\partial K)]^m$-norm.
Let us denote $\mathbf{P}_\delta$ the orthogonal projector in $[L^2(\Omega)]^m$ onto the finite element space V_δ. For a given $\mathbf{v} \in [L^2(\Omega)]^m$, $\mathbf{P}_\delta\mathbf{v}$ is defined by

$$(\mathbf{v}-\mathbf{P}_\delta\mathbf{v},\mathbf{w})_{\tau_h} = 0 \qquad \forall \mathbf{w}\in V_\delta\,.$$

Analogously, we define the orthogonal projector $\mathbf{P}_\delta^K$ in $[L^2(K)]^m$ onto the local space V_δ^K by

$$(\mathbf{v}-\mathbf{P}_\delta^K\mathbf{v},\mathbf{w})_K = 0 \qquad \forall \mathbf{w}\in V_\delta^K$$

where $(\mathbf{v},\mathbf{w})_K = \sum_{i=1}^m (\mathbf{v}_i,\mathbf{w}_i)_K$ denotes the usual $[L^2(K)]^m$-scalar product on K.
For $r\ge 0$ let us define the following spaces

$$W^{r,\infty}(\tau_h) = \left\{v \;\middle|\; v_{|_K} \in W^{r,\infty}(K),\ \forall K\in\tau_h\right\}.$$

Then we consider $[W^{r,\infty}(\tau_h)]^{m\times m}$ and denote its norm as $\|\cdot\|_{r,\infty}$. For $k=0$, the norm is defined by

$$\|A\|_{0,\infty} = \max_{K\in\tau_h}\|A\|_{[L^\infty(K)]^{d\times d}}$$

whereas for $k=1$, $\|\cdot\|_{1,\infty}$ is defined by:

$$\|A\|_{1,\infty} = \|A\|_{0,\infty} + \max_{l=1,..,d}\|\partial_l A\|_{0,\infty}\,.$$

Next, we present two lemmas which are directly derived from Lemma 1.1 and Lemma 1.2. They are just a generalization to estimate the projection error on K and ∂K in the context of the above defined multidimensional projector $\mathbf{P}_\delta^K$ in $[L^2(K)]^m$ onto V_δ^K.

Lemma 2.4 *For any $K \in \tau_h$, let $\mathbf{v} \in [H^k(K)]^m$ for some integer $k \geq 1$. Further, let $\mathbf{P}_\delta^K$ be the $[L^2(K)]^m$-projection onto V_δ^K with $N_K \geq 1$; then, for any integer s, $0 \leq s \leq \min(N_K + 1, k)$, we have*

$$\|\mathbf{v} - \mathbf{P}_\delta^K \mathbf{v}\|_K \leq C_K(d) \frac{h_K^s}{N_K^s} |\mathbf{v}|_{K,s}$$

where C_K depends only on the spatial dimension and the element K.

Proof. Applying Lemma 1.1 yields

$$\begin{aligned} \|\mathbf{v} - \mathbf{P}_\delta^K \mathbf{v}\|_K^2 &= \sum_{j=1}^m \|(\mathbf{v} - \mathbf{P}_\delta^K \mathbf{v})_j\|_{L^2(K)}^2 \leq \sum_{j=1}^m C_K^2(d) \frac{h_K^{2s}}{N_K^{2s}} |\mathbf{v}_j|_{H^s(K)}^2 \\ &\leq C_K^2(d) \frac{h_K^{2s}}{N_K^{2s}} \sum_{j=1}^m |\mathbf{v}_j|_{H^s(K)}^2 = C_K^2(d) \frac{h_K^{2s}}{N_K^{2s}} |\mathbf{v}|_{K,s}^2 \,. \end{aligned}$$

□ **Lemma** (2.4)

Lemma 2.5 *Let $K \in \tau_h$ and suppose that $\mathbf{v} \in [H^k(K)]^m$ for some integer $k \geq 1$. Then, for any integer s, $0 \leq s \leq \min(N_K + 1, k)$ and $N_K \geq 0$, we have that*

$$\|\mathbf{v} - \mathbf{P}_\delta^K \mathbf{v}\|_{\partial K} \leq C_K(d, s) \frac{h_K^{s-\frac{1}{2}}}{(N_K + 1)^{s-\frac{1}{2}}} |\mathbf{v}|_{K,s} \,.$$

The constant C_K is only depending on s and the element K.

Proof. Applying Lemma 1.2 yields

$$\begin{aligned} \|\mathbf{v} - \mathbf{P}_\delta^K \mathbf{v}\|_{\partial K}^2 &= \sum_{j=1}^m \|(\mathbf{v} - \mathbf{P}_\delta^K \mathbf{v})_j\|_{L^2(\partial K)}^2 \leq \sum_{j=1}^m C_K^2(d,s) \frac{h_K^{2s-1}}{(N_K+1)^{2s-1}} |\mathbf{v}_j|_{H^s(K)}^2 \\ &\leq C_K^2(d,s) \frac{h_K^{2s-1}}{(N_K+1)^{2s-1}} \sum_{j=1}^m |\mathbf{v}_j|_{H^s(K)}^2 = C_K^2(d,s) \frac{h_K^{2s-1}}{(N_K+1)^{2s-1}} |\mathbf{v}|_{K,s}^2 \,. \end{aligned}$$

□ **Lemma** (2.5)

2.7 Spectral Discontinuous Galerkin Method

In this section, the spectral discontinuous Galerkin method is discussed for problem (2.2.1). Additionally to the conditions on B and A_j assumed in section 2.2, let us suppose that the condition (F2) of section 2.4 holds. This means that there exists $\sigma_0 > 0$ such that

$$B + B^T - \sum_{j=1}^m \partial_j A_j \geq 2\sigma_0 > 0 \,.$$

Let us recall that in section 2.2 we assumed that B is positive definite and the matrices A_j are symmetric.

2.7.1 Remark on the Boundary Conditions

Here, we show that imposing the boundary condition on the incoming characteristics is a particular case of the boundary condition of type

$$B_\Gamma \mathbf{u} = C\mathbf{z} = \mathbf{g}$$

Let C be of the form $C = (0\ I)$, then $C\mathbf{z} = \mathbf{z}^- = \mathbf{g}$. As $B_\Gamma = CS^T$, we have $B_\Gamma = S^{-T}$. As consequence

$$\begin{aligned} C^+ &= B_\Gamma S^+ = S^{-T} S^+ = 0 \\ C^- &= B_\Gamma S^- = S^{-T} S^- = I \end{aligned}$$

due to of the orthogonality of S. Finally, we conclude that $\bar{C} = 0$. Observe also that $\mathbf{g} = \mathbf{g}_C$ in this case.

2.7.2 The Method

Consider the following problem on an element $K \in \mathcal{T}_h$:

find $\mathbf{u} : K \to \mathbb{R}^m$, such that

$$\begin{aligned} \textstyle\sum_{j=1}^d A_j \partial_j \mathbf{u} + B\mathbf{u} &= \mathbf{f} && \text{in } K \\ (S^-)^T \mathbf{u} &= (S^-)^T \mathbf{u}_{ext} && \text{on } \Gamma_i^K \\ B_\Gamma \mathbf{u} &= \mathbf{g} && \text{on } \Gamma_e^K \end{aligned} \qquad (2.7.1)$$

where $\Gamma_i^K = \partial K \backslash \Gamma$ and $\Gamma_e^K = \partial K \cap \Gamma$. This is problem (2.2.1) restricted to one element $K \in \mathcal{T}_h$. The boundary condition $(S^-)^T \mathbf{u} = (S^-)^T \mathbf{u}_{ext}$ on Γ_i^K means that on an interior face, the boundary conditions are imposed on the incoming characteristics since $(S^-)^T \mathbf{u} = \mathbf{z}^-$.
Let us define the local space W_K by

$$W_K = \left\{ \mathbf{v} \in [L^2(K)]^m \,\middle|\, \Sigma_{j=1}^d A_j \partial_j \mathbf{v} \in [L^2(K)]^m \right\}$$

and the recall the definition of the global space W_Ω

$$W_\Omega = \left\{ \mathbf{v} \in V_\Omega \,\middle|\, \Sigma_{j=1}^d A_j \partial_j \mathbf{v} \in V_\Omega \right\}$$

where $V_\Omega = [L^2(\Omega)]^m$. Let $\mathbf{v} \in W_K$ be a test function. Then, multiplying equation (2.7.1) by the test function $\mathbf{v}$ and integrating over K leads to

find $\mathbf{u} \in W_K$ such that

$$\begin{aligned} \left(\Sigma_{j=1}^d A_j \partial_j \mathbf{u} + B\mathbf{u}, \mathbf{v}\right)_K &= (\mathbf{f}, \mathbf{v})_K && \forall\, \mathbf{v} \in W_K \\ (S^-)^T \mathbf{u} &= (S^-)^T \mathbf{u}_{ext} && \text{on } \Gamma_i^K \\ B_\Gamma \mathbf{u} &= \mathbf{g} && \text{on } \Gamma_e^K\ . \end{aligned}$$

Integration by parts

$$(A_j \partial_j \mathbf{u}, \mathbf{v})_K = -\big(\mathbf{u}, \partial_j (A_j \mathbf{v})\big)_K + (n_j A_j \mathbf{u}, \mathbf{v})_{\partial K}$$

is used so that

$$\sum_{j=1}^d (A_j \partial_j \mathbf{u}, \mathbf{v})_K = -\sum_{j=1}^d \big(\mathbf{u}, \partial_j (A_j \mathbf{v})\big)_K + (D\mathbf{u}, \mathbf{v})_{\partial K}\ .$$

Then the problem becomes:

find $\mathbf{u} \in W_K$ such that

$$\begin{array}{rcll} (B\mathbf{u},\mathbf{v})_K - \sum_{j=1}^{d} \big(\mathbf{u}, \partial_j(A_j\mathbf{v})\big)_K + (D\mathbf{u},\mathbf{v})_{\partial K} & = & (\mathbf{f},\mathbf{v})_K & \forall\, \mathbf{v} \in W_K \\ (S^-)^T\mathbf{u} & = & (S^-)^T\mathbf{u}_{ext} & \text{on } \Gamma_i^K \\ B_\Gamma \mathbf{u} & = & \mathbf{g} & \text{on } \Gamma_e^K\ . \end{array}$$

Splitting D in $M^+ - M^-$ leads to

$$(B\mathbf{u},\mathbf{v})_K - \sum_{j=1}^{d} \big(\mathbf{u}, \partial_j(A_j\mathbf{v})\big)_K + (M^+\mathbf{u},\mathbf{v})_{\partial K} = (\mathbf{f},\mathbf{v})_K + (M^-\mathbf{u},\mathbf{v})_{\partial K}$$

and splitting ∂K in $\Gamma_i^K \cup \Gamma_e^K$ yields

$$(B\mathbf{u},\mathbf{v})_K - \sum_{j=1}^{d} \big(\mathbf{u}, \partial_j(A_j\mathbf{v})\big)_K + (M^+\mathbf{u},\mathbf{v})_{\partial K} - (M^-\mathbf{u},\mathbf{v})_{\Gamma_i^K} = (\mathbf{f},\mathbf{v})_K + (M^-\mathbf{u},\mathbf{v})_{\Gamma_e^K}\ .$$

Let us write the boundary terms of the previous equation in terms of the characteristic variables $\mathbf{z} = S^T\mathbf{u}$. For this let $\mathbf{y} = S^T\mathbf{v}$. The hybrid formulation is:

find $\mathbf{u} \in W_K$ such that

$$\begin{array}{rcll} (B\mathbf{u},\mathbf{v})_K - \sum_{j=1}^{d} \big(\mathbf{u}, \partial_j(A_j\mathbf{v})\big)_K + (\Lambda^+\mathbf{z},\mathbf{y})_{\partial K} + (\Lambda^-\mathbf{z},\mathbf{y})_{\Gamma_i^K} & = & (\mathbf{f},\mathbf{v})_K - (\Lambda^-\mathbf{z},\mathbf{y})_{\Gamma_e^K} & \\ & & \forall\, \mathbf{v} \in W_K & \\ \mathbf{z}^- & = & \mathbf{z}_{ext}^+ \ \ \text{on } \Gamma_i^K & \\ C\mathbf{z} & = & \mathbf{g} \ \ \text{on } \Gamma_e^K & \end{array}$$

where $\mathbf{z}_{ext}^+ = (S^-)^T\mathbf{u}_{ext}$. Then, the boundary conditions are imposed in a weak sense. As in section 2.5 we can write

$$\mathbf{z}^- = \bar{C}\mathbf{z}^+ + \mathbf{g}_C$$

and

$$(\Lambda^-\mathbf{z},\mathbf{y})_{\Gamma_e^K} = (\lambda^-\mathbf{z}^-,\mathbf{y}^-)_{\Gamma_e^K}\ .$$

Hence

$$(\Lambda^-\mathbf{z},\mathbf{y})_{\Gamma_e^K} = \big(\lambda^-(\bar{C}\mathbf{z}^+ + \mathbf{g}_C),\mathbf{y}^-\big)_{\Gamma_e^K}$$

Additionally

$$(\Lambda^-\mathbf{z},\mathbf{y})_{\Gamma_i^K} = (\lambda^-\mathbf{z}^-,\mathbf{y}^-)_{\Gamma_i^K} = (\lambda^-\mathbf{z}_{ext}^+,\mathbf{y}^-)_{\Gamma_i^K} = (\Lambda^-\mathbf{z}_{ext},\mathbf{y})_{\Gamma_i^K}$$

and the problem becomes:

find $\mathbf{u} \in W_K$ such that

$$\begin{array}{l} (B\mathbf{u},\mathbf{v})_K - \sum_{j=1}^{d} \big(\mathbf{u}, \partial_j(A_j\mathbf{v})\big)_K + (\Lambda^+\mathbf{z},\mathbf{y})_{\partial K} + (\Lambda^-\mathbf{z}_{ext},\mathbf{y})_{\Gamma_i^K} \\ \qquad\qquad = (\mathbf{f},\mathbf{v})_K - \big(\lambda^-(\bar{C}\mathbf{z}^+ + \mathbf{g}_C),\mathbf{y}^-\big)_{\Gamma_e^K} \quad \forall\, \mathbf{v} \in W_K \end{array}$$

where $\mathbf{z} = S^T\mathbf{u}$, $\mathbf{y} = S^T\mathbf{v}$ and $\mathbf{z}_{ext} = S^T\mathbf{u}_{ext}$. Let us define

$$\bar{\bar{C}} = \begin{pmatrix} 0 & 0 \\ \bar{C} & 0 \end{pmatrix} \qquad \text{and} \qquad \bar{\mathbf{g}} = (0 \quad \mathbf{g}_C^T)^T$$

and observe that

$$\big(\lambda^- \bar{C}\mathbf{z}^+, \mathbf{y}^-\big)_{\Gamma_e^K} = \big(\Lambda^-\, \bar{\bar{C}}\mathbf{z}, \mathbf{y}\big)_{\Gamma_e^K}$$

and

$$-\big(\lambda^- \mathbf{g}_C, \mathbf{y}^-\big)_{\Gamma_e^K} = \big(|\Lambda^-|\; \bar{\mathbf{g}}, \mathbf{y}\big)_{\Gamma_e^K} .$$

The problem becomes:

find $\mathbf{u} \in W_K$ such that

$$\begin{aligned}(B\mathbf{u}, \mathbf{v})_K - \textstyle\sum_{j=1}^d \big(\mathbf{u}, \partial_j(A_j\mathbf{v})\big)_K + (\Lambda^+\mathbf{z}, \mathbf{y})_{\partial K} + (\Lambda^-\mathbf{z}_{ext}, \mathbf{y})_{\Gamma_i^K}& \\ +\big(\Lambda^-\, \bar{\bar{C}}\mathbf{z}, \mathbf{y}\big)_{\Gamma_e^K} = (\mathbf{f}, \mathbf{v})_K + \big(|\Lambda^-|\; \bar{\mathbf{g}}, \mathbf{y}\big)_{\Gamma_e^K} \quad \forall\, \mathbf{v} \in W_K .&\end{aligned}$$

Let us rewrite the equation in terms of the physical variables $\mathbf{u}$ and $\mathbf{v}$:

find $\mathbf{u} \in W_K$ such that

$$\begin{aligned}(B\mathbf{u}, \mathbf{v})_K - \textstyle\sum_{j=1}^d \big(\mathbf{u}, \partial_j(A_j\mathbf{v})\big)_K + (M^+\mathbf{u}, \mathbf{v})_{\partial K} - (M^-\mathbf{u}_{ext}, \mathbf{v})_{\Gamma_i^K}& \\ -\big(M^-\; S\,\bar{\bar{C}}S^T\mathbf{u}, \mathbf{v}\big)_{\Gamma_e^K} = (\mathbf{f}, \mathbf{v})_K + \big(S|\Lambda^-|\bar{\mathbf{g}}, \mathbf{v}\big)_{\Gamma_e^K} \quad \forall \mathbf{v} \in W_K .&\end{aligned}$$

Next, a Galerkin approximation is used. This means that the functional space W_K is replaced by the finite dimensional space $V_\delta^K \subset W_K$. In addition, counterintegrating by parts leads to the problem:

find $\mathbf{u}_{DG} \in V_\delta^K$ such that

$$\begin{aligned}&(B\mathbf{u}_{DG}, \mathbf{v}_\delta)_K + \textstyle\sum_{j=1}^d (A_j\partial_j\mathbf{u}_{DG}, \mathbf{v}_\delta)_K + \big(M^-(\mathbf{u}_{DG} - \mathbf{u}_{ext}), \mathbf{v}_\delta\big)_{\Gamma_i^K} \\ &\quad +(M^-\;\mathbf{u}_{DG}, \mathbf{v}_\delta)_{\Gamma_e^K} - (M^-\; S\,\bar{\bar{C}}S^T\mathbf{u}_{DG}, \mathbf{v}_\delta)_{\Gamma_e^K} \\ &\quad = (\mathbf{f}, \mathbf{v}_\delta)_K + (S|\Lambda^-|\bar{\mathbf{g}}, \mathbf{v}_\delta)_{\Gamma_e^K} \qquad\qquad \forall\, \mathbf{v}_\delta \in V_\delta^K .\end{aligned}$$

Let ∂K be an interior face of K and K' be a neighboring element. On $\partial K \cap \partial K'$, we define

$$[\mathbf{u}] = \mathbf{u}_{DG} - \mathbf{u}_{ext}$$

where $\mathbf{u}_{ext}$ is for the local approach on K given, but for the global problem $\mathbf{u}_{ext}$ denotes the solution on the neighboring element K'.

Local Approach:
The bilinear form $a_K : W_K \times W_K \to \mathbb{R}$ as well as the linear form $F_K : W_K \to \mathbb{R}$ are defined by

$$\begin{aligned}a_K(\mathbf{v}, \mathbf{w}) &= (B\mathbf{v}, \mathbf{w})_K + \textstyle\sum_{j=1}^d (A_j\partial_j\mathbf{v}, \mathbf{w})_K + \big(M^-\;[\mathbf{v}], \mathbf{w}\big)_{\Gamma_i^K} \\ &\quad +\big(M^-(I - S\bar{\bar{C}}S^T)\;\mathbf{v}, \mathbf{w}\big)_{\Gamma_e^K} \\ F_K(\mathbf{v}) &= (\mathbf{f}, \mathbf{v})_K + \big(S|\Lambda^-|\bar{\mathbf{g}}, \mathbf{v}\big)_{\Gamma_e^K}\end{aligned}$$

for all $\mathbf{v}, \mathbf{w} \in W^K$. The problem can be formulated as:

find $\mathbf{u}_{DG} \in V_\delta^K$ such that

$$a_K(\mathbf{u}_{DG}, \mathbf{v}_\delta) = F_K(\mathbf{v}_\delta) \qquad \forall\, \mathbf{v}_\delta \in V_\delta^K \tag{2.7.2}$$

Global Approach:
Now, the problem on the whole domain Ω is considered. Taking the sum over all elements $K \in \mathcal{T}_h$ leads to the global problem:

find $\mathbf{u}_{DG} \in V_\delta$ such that

$$a(\mathbf{u}_{DG}, \mathbf{v}_\delta) = F(\mathbf{v}_\delta) \qquad \forall\, \mathbf{v}_\delta \in V_\delta \tag{2.7.3}$$

where $a : W_\Omega \times W_\Omega \to \mathbb{R}$, $F : W_\Omega \to \mathbb{R}$ and

$$\begin{aligned} a(\mathbf{v}, \mathbf{w}) &= \textstyle\sum_{K \in \mathcal{T}_h} a_K(\mathbf{v}, \mathbf{w}) = (B\mathbf{v}, \mathbf{w})_{\mathcal{T}_h} + \sum_{j=1}^{d} (A_j \partial_j \mathbf{v}, \mathbf{w})_{\mathcal{T}_h} \\ &\quad + \textstyle\sum_{K \in \mathcal{T}_h} \left\{ \left(M^- \, [\mathbf{v}], \mathbf{w}\right)_{\Gamma_i^K} + \left(M^-(I - S\bar{\bar{C}}S^T)\, \mathbf{v}, \mathbf{w}\right)_{\Gamma_e^K} \right\} \\ F(\mathbf{v}) &= \textstyle\sum_{K \in \mathcal{T}_h} F_K(\mathbf{v}) = (\mathbf{f}, \mathbf{v})_\Omega + (S|\Lambda^-|\bar{\mathbf{g}}, \mathbf{v})_\Gamma \end{aligned} \tag{2.7.4}$$

for all $\mathbf{v}$, $\mathbf{w} \in W_\Omega$.

Remark 2.6 *In the case of boundary conditions imposed on the incoming characteristics, that is* $\mathbf{z}^- = \mathbf{g}$*, the above definitions still hold with* $\bar{\bar{C}} = 0$ *and* $\mathbf{g}_C = \mathbf{g}$*. Thus* $\bar{\mathbf{g}} = (0 \;\; \mathbf{g}^T)^T$*.*

2.8 Convergence Analysis

The convergence rate depending on the local mesh size h_K and the local polynomial order N_K of the spectral discontinuous Galerkin method for linear hyperbolic systems is studied.

We assume boundary conditions imposed on the incoming characteristics which implies that $\bar{\bar{C}} = 0$ in the context of the previous section. Let us denote $\tilde{C}_K(d)$ the constant of Lemma 2.4 and $C_K(d, s) = C_K(d) \cdot C_K(s)$ the constant of Lemma 2.5. Note that the main result of this section is Theorem 2.14 which is based on Lemma 3.4 and 3.9 of [10] as all intermediate Lemmas. Lemma 3.4 and 3.9 of [10] are generalized for the case of hyperbolic systems (and not scalar hyperbolic equations) and are presented in this section as Lemma 2.4 and Lemma 2.5. Lemma 2.8-2.13 and Theorem 2.14 are entirely worked out in this project.

Let us first define a norm for W_Ω. For this, the term $a(\mathbf{v}, \mathbf{v})$ is first developed for a $\mathbf{v} \in W_\Omega$. Since all matrices A_j are symmetric

$$\begin{aligned} \sum_{j=1}^{d} (A_j \partial_j \mathbf{v}, \mathbf{v})_{\partial K} &= (D\mathbf{v}, \mathbf{v})_{\partial K} - \sum_{j=1}^{d} \big(\partial_j (A_j \mathbf{v}, \mathbf{v})\big)_K \\ &= (D\mathbf{v}, \mathbf{v})_{\partial K} - \sum_{j=1}^{d} \big(\partial_j A_j \mathbf{v}, \mathbf{v}\big)_K - \sum_{j=1}^{d} \big(A_j \partial_j \mathbf{v}, \mathbf{v}\big)_K \end{aligned}$$

and hence

$$\begin{aligned} \sum_{j=1}^{d} (A_j \partial_j \mathbf{v}, \mathbf{v})_{\partial K} &= \frac{1}{2}(D\mathbf{v}, \mathbf{v})_{\partial K} - \frac{1}{2} \sum_{j=1}^{d} (\partial_j A_j \mathbf{v}, \mathbf{v})_K \\ &= \frac{1}{2}(M^+\mathbf{v}, \mathbf{v})_{\partial K} - \frac{1}{2}(M^-\mathbf{v}, \mathbf{v})_{\partial K} - \frac{1}{2} \sum_{j=1}^{d} (\partial_j A_j \mathbf{v}, \mathbf{v})_K \,. \end{aligned}$$

In addition, observe that M^- is symmetric and consequently

$$(\mathbf{a}-\mathbf{b})^T M^-\,\mathbf{a} = \frac{1}{2}\Big(\mathbf{a}^T M^-\,\mathbf{a} + (\mathbf{a}-\mathbf{b})^T M^-(\mathbf{a}-\mathbf{b}) - \mathbf{b}^T M^-\,\mathbf{b}\Big)\,.$$

Then

$$\begin{aligned} a(\mathbf{v},\mathbf{v}) &= \big((B-\frac{1}{2}\Sigma_{j=1}^{d}\partial_j A_j)\mathbf{v},\mathbf{v}\big)_{\tau_h} \\ &\quad +\frac{1}{2}\sum_{K\in\tau_h}\Big\{\big(M^+\,\mathbf{v},\mathbf{v}\big)_{\partial K} - \big(M^-\,\mathbf{v},\mathbf{v}\big)_{\partial K} + \big(M^-\,\mathbf{v},\mathbf{v}\big)_{\Gamma_i^K} \\ &\quad +\big(M^-\,[\mathbf{v}],[\mathbf{v}]\big)_{\Gamma_i^K} - \big(M^-\,\mathbf{v}_{ext},\mathbf{v}_{ext}\big)_{\Gamma_i^K} + 2(M^-\,\mathbf{v},\mathbf{v})_{\Gamma_e^K}\Big\}\,. \end{aligned}$$

Let $\tilde{B} = B + B^T - \sum_{j=1}^{d}\partial_j A_j$

$$\begin{aligned} a(\mathbf{v},\mathbf{v}) &= \frac{1}{2}(\tilde{B}\mathbf{v},\mathbf{v})_{\tau_h} + \frac{1}{2}\sum_{K\in\tau_h}\Big\{(M^+\,\mathbf{v},\mathbf{v})_{\Gamma_i^K} - (M^-\,\mathbf{v},\mathbf{v})_{\Gamma_i^K} \\ &\quad +(M^+\,\mathbf{v},\mathbf{v})_{\Gamma_e^K} - (M^-\,\mathbf{v},\mathbf{v})_{\Gamma_e^K} + (M^-\,\mathbf{v},\mathbf{v})_{\Gamma_i^K} \\ &\quad +\big(M^-\,[\mathbf{v}],[\mathbf{v}]\big)_{\Gamma_i^K} - (M^-\,\mathbf{v}_{ext},\mathbf{v}_{ext})_{\Gamma_i^K} + 2(M^-\,\mathbf{v},\mathbf{v})_{\Gamma_e^K}\Big\} \\ &= \frac{1}{2}(\tilde{B}\mathbf{v},\mathbf{v})_{\tau_h} + \frac{1}{2}\sum_{K\in\tau_h}\Big\{\big(M^-\,[\mathbf{v}],[\mathbf{v}]\big)_{\Gamma_i^K} + (M^+\,\mathbf{v},\mathbf{v})_{\Gamma_i^K} \\ &\quad -(M^-\,\mathbf{v}_{ext},\mathbf{v}_{ext})_{\Gamma_i^K} + (M^+\,\mathbf{v},\mathbf{v})_{\Gamma_e^K} + (M^-\,\mathbf{v},\mathbf{v})_{\Gamma_e^K}\Big\}\,. \end{aligned}$$

Let K and K' be two adjoint elements such that $\Sigma = \partial K \cap \partial K'$ is a face. D_K resp. $D_{K'}$ denotes the matrix D corresponding to K resp. K' implies

$$D_K + D_{K'} = 0$$

so that $S\Lambda_K S^T = -S\Lambda_{K'}S^T$ and $\Lambda_K = -\Lambda_{K'}$. Considering the positive and negative definite parts of this equality. Then

$$\begin{aligned} \Lambda_K^+ &= -\Lambda_{K'}^- \\ \Lambda_K^- &= -\Lambda_{K'}^+ \end{aligned}$$

and consequently $D_K^- + D_{K'}^+ = 0$ and $M_K^- = M_{K'}^+$.
Then,

$$\frac{1}{2}\sum_{K\in\tau_h}(M_K^-\,\mathbf{v}_{ext},\mathbf{v}_{ext})_{\Gamma_i^K} = \frac{1}{2}\sum_{K\in\tau_h}(M_{K'}^+\,\mathbf{v}_{ext},\mathbf{v}_{ext})_{\Gamma_i^K} = \frac{1}{2}\sum_{K'\in\tau_h}(M_{K'}^+\,\mathbf{v},\mathbf{v})_{\Gamma_i^{K'}} \qquad (2.8.1)$$

and

$$a(\mathbf{v},\mathbf{v}) = \frac{1}{2}(\tilde{B}\mathbf{v},\mathbf{v})_{\tau_h} + \frac{1}{2}\sum_{K\in\tau_h}\Big\{\big(M^-\,[\mathbf{v}],[\mathbf{v}]\big)_{\Gamma_i^K} + (M\mathbf{v},\mathbf{v})_{\Gamma_e^K}\Big\}\,.$$

Therefore the norm $|||\cdot|||$ for W_Ω is defined by

$$|||\mathbf{v}|||^2 = \|\sigma_0^{\frac{1}{2}}\mathbf{v}\|_{\tau_h}^2 + \frac{1}{2}\sum_{K\in\tau_h}\Big\{\big(M^-\,[\mathbf{v}],[\mathbf{v}]\big)_{\Gamma_i^K} + (M\mathbf{v},\mathbf{v})_{\Gamma_e^K}\Big\} \qquad \text{for} \quad \mathbf{v}\in W_\Omega\,.$$

Lemma 2.7 (Coercivity) *For all $v \in W_\Omega$:*

$$a(\mathbf{v},\mathbf{v}) \geq |||\mathbf{v}|||^2$$

Proof. Let us recall condition (F2) of section 2.4:

$$\tilde{B} = B + B^T - \sum_{j=1}^{d} \partial_j A_j \geq 2\sigma_0 I > 0 \qquad \text{a.e. on } \Omega$$

Then one can conclude that

$$\frac{1}{2}(\tilde{B}\mathbf{v},\mathbf{v})_{\tau_h} \geq \|\sigma_0^{\frac{1}{2}}\mathbf{v}\|^2_{\tau_h}$$

and hence

$$a(\mathbf{v},\mathbf{v}) \geq |||\mathbf{v}|||^2 \,.$$

□ **Lemma** (2.7)

Now, the Galerkin Orthogonality holds also for linear hyperbolic systems.

Lemma 2.8 (Galerkin Orthogonality) *Let $\mathbf{u}$ be the exact solution of problem (2.2.1) and $\mathbf{u}_{DG}$ the solution of (2.7.3). If $\mathbf{u} \in [H^1(\Omega)]^m$, then*

$$a(\mathbf{u}_{DG} - \mathbf{u}, \mathbf{v}_\delta) = 0 \qquad \forall\, \mathbf{v}_\delta \in V_\delta$$

Proof. $\mathbf{u}_{DG}$ is satisfying (2.7.3), this means

$$a(\mathbf{u}_{DG}, \mathbf{v}_\delta) = F(\mathbf{v}_\delta) \qquad \forall \mathbf{v}_\delta \in V_\delta \,. \tag{2.8.2}$$

Additionally, $\mathbf{u}$ being the exact solution of (2.2.1) implies that

$$\left(\Sigma_{j=1}^{d} A_j \partial_j \mathbf{u} + B\mathbf{u}, \mathbf{v}_\delta\right)_{\tau_h} = \left(\mathbf{f}, \mathbf{v}_\delta\right)_{\tau_h} \qquad \forall \mathbf{v}_\delta \in V_\delta$$

$\mathbf{u}$ being in $[H^1(\Omega)]^m$ yields $\mathbf{u} \in [H^1(K)]^d, \forall K \in \tau_h$. Then the trace on every element K is well defined. Hence

$$\sum_{K \in \tau_h} \left(M^- [\mathbf{u}], [\mathbf{u}]\right)_{\Gamma_i^K} = 0$$

and the boundary condition $\mathbf{z}^- = \mathbf{g}$ implies

$$\begin{aligned} \left(M^-\mathbf{u}, \mathbf{v}_\delta\right)_{\Gamma_e^K} &= \left(|\Lambda^-|\, \mathbf{z}, \mathbf{y}_\delta\right)_{\Gamma_e^K} = \left(|\lambda^-|\, \mathbf{z}^-, \mathbf{y}_\delta^-\right)_{\Gamma_e^K} \\ &= \left(|\lambda^-|\, \mathbf{g}, \mathbf{y}_\delta^-\right)_{\Gamma_e^K} = \left(|\Lambda^-|\, \bar{\mathbf{g}}, \mathbf{y}_\delta\right)_{\Gamma_e^K} \\ &= \left(S|\Lambda^-|\, \bar{\mathbf{g}}, \mathbf{v}_\delta\right)_{\Gamma_e^K} \qquad \forall \mathbf{v}_\delta \in V_\delta \end{aligned}$$

where $\mathbf{z}$ and $\mathbf{y}_\delta$ are the characteristic variables of $\mathbf{u}$ and $\mathbf{v}_\delta$. Then we conclude that

$$a(\mathbf{u}, \mathbf{v}_\delta) = F(\mathbf{v}_\delta) \qquad \forall \mathbf{v}_\delta \in V_\delta \tag{2.8.3}$$

and taking the difference between (2.8.2) and (2.8.3) leads to the result.

□ **Lemma** (2.8)

In addition, a semi-norm for W_Ω is defined, which we will only use for intermediate results.

$$|]\mathbf{v}[|^2 = \sum_{K\in\tau_h} (M\mathbf{v},\mathbf{v})_{\partial K} \qquad \text{for} \quad \mathbf{v}\in W_\Omega$$

In the following, η is defined by $\eta = \mathbf{u} - \mathbf{P}_\delta \mathbf{u}$ and $\xi = \mathbf{u}_{DG} - \mathbf{P}_\delta\mathbf{u}$ where $\mathbf{u}$ denotes the exact solution of (2.2.1) and $\mathbf{u}_{DG}$ its SDG-approximation defined by (2.7.3). Note that $\xi,\eta\in W_\Omega$. The next lemma is a continuity result on $a(\cdot,\cdot)$ that will be used in the proof of Theorem 2.14.

Lemma 2.9 *If $A_j \in [W^{1,\infty}(\tau_h)]^{m\times m}$ for all $j = 1,..,d$, then*

$$a(\eta,\xi) \;\; \leq \;\; \Big[C_0\|\eta\|_{\tau_h} + C_A d\Big(\sum_{K\in\tau_h} N_K^4 \|\eta\|_K^2\Big)^{\frac{1}{2}} + 4\,|]\eta[|\Big]\;|||\xi|||$$

Proof. By definition of the bilinear form $a : W_\Omega\times W_\Omega \to \mathbb{R}$:

$$a(\eta,\xi) = (B\eta,\xi)_{\tau_h} + \sum_{j=1}^{d}(A_j\partial_j\eta,\xi)_{\tau_h} + \sum_{K\in\tau_h}\Big\{\big(M^-\,[\eta],\xi\big)_{\Gamma_i^K} + (M^-\,\eta,\xi)_{\Gamma_e^K}\Big\}$$

Integrating the second term of the right hand side by parts yields

$$\begin{aligned} a(\eta,\xi) \;\; &= \;\; (\eta,\bar{B}\xi)_{\tau_h} - \sum_{j=1}^{d}(\eta,A_j\partial_j\xi)_{\tau_h} \\ &\quad + \sum_{K\in\tau_h}\Big\{(D\eta,\xi)_{\partial K} + (M^-\,[\eta],\xi)_{\Gamma_i^K} + (M^-\,\eta,\xi)_{\Gamma_e^K}\Big\} \end{aligned}$$

where $\bar{B} = B^T - \sum_{j=1}^d \partial_j A_j$.
Let us define the following three terms and treat them separately.

$$\begin{aligned} \mathcal{I}_1 \;\; &= \;\; (\eta,\bar{B}\xi)_{\tau_h} \\ \mathcal{I}_2 \;\; &= \;\; -\sum_{j=1}^{d}(\eta,A_j\partial_j\xi)_{\tau_h} \\ \mathcal{I}_3 \;\; &= \;\; \sum_{K\in\tau_h}\Big\{(D\eta,\xi)_{\partial K} + \big(M^-\,[\eta],\xi\big)_{\Gamma_i^K} + (M^-\,\eta,\xi)_{\Gamma_e^K}\Big\} \end{aligned}$$

Then, we bound each term:

- Let us develop a lemma for bounding the term $\mathcal{I}_1$:

  Lemma 2.10 *Let B be a matrix such that $B\in[L^\infty(\tau_h)]^{m\times m}$, then*

  $$(\eta,B\xi)_{\tau_h} \leq C_0\|\eta\|_{\tau_h}\;|||\xi|||$$

  where

  $$C_0 = \frac{\|B\|_{0,\infty}}{\sigma_0^{1/2}}\,.$$

Proof Lemma (2.10).

$$\begin{aligned}(\eta, B\xi)_{\tau_h} &= \sum_{K\in\tau_h}(\eta, B\xi)_K = \sum_{K\in\tau_h}\int_K \eta^T B\xi \\ &= \sum_{K\in\tau_h}\int_K \sum_{i,k=1}^{m}\eta_i B_{i,k}\xi_k \le \|B\|_{0,\infty}\sum_{K\in\tau_h}\sum_{i,k=1}^{m}\int_K |\eta_i|\,|\xi_k| \\ &\le \|B\|_{0,\infty}\Big(\sum_{K\in\tau_h}\sum_{i,k=1}^{m}\|\eta_i\|^2_{L^2(K)}\Big)^{\frac{1}{2}}\Big(\sum_{K\in\tau_h}\sum_{i,k=1}^{m}\|\xi_k\|^2_{L^2(K)}\Big)^{\frac{1}{2}} \\ &\le \frac{\|B\|_{0,\infty}}{\sigma_0^{1/2}}\Big(\sum_{K\in\tau_h}\|\eta\|^2_K\Big)^{\frac{1}{2}}\Big(\sum_{K\in\tau_h}\|\sigma_0^{1/2}\xi\|^2_K\Big)^{\frac{1}{2}}\end{aligned}$$

Let

$$C_0 = \frac{\|B\|_{0,\infty}}{\sigma_0^{1/2}}$$

Then

$$(\eta, B\xi)_{\tau_h} \le C_0\|\eta\|_{\tau_h}\ |||\xi|||$$

□ **Lemma** (2.10)

Since $\mathcal{I}_1 = (\eta, \bar{B}\xi)_{\tau_h}$ and $\bar{B} \in [L^\infty(\Omega)]^{m\times m}$ by hypothesis, we conclude that

$$\mathcal{I}_1 \le C_0\|\eta\|_{\tau_h}\ |||\xi|||$$

where

$$C_0 = \frac{\|\bar{B}\|_{0,\infty}}{\sigma_0^{1/2}}\,.$$

- For the second term $\mathcal{I}_2$ the estimated bound is also formulated in the form of a lemma.

Lemma 2.11 *If $A_j \in [W^{1,\infty}(\tau_h)]^{m\times m}$ for all $j = 1,..,d$, then*

$$\Big|\sum_{j=1}^{d}(\eta, A_j\partial_j\xi)_{\tau_h}\Big| \le C_A d\Big(\sum_{K\in\tau_h} N_K^4\|\eta\|^2_K\Big)^{\frac{1}{2}}|||\xi|||\,.$$

Proof Lemma (2.11). Let $\bar{A}^j_K = \frac{1}{|K|}\int_K A_j$ and observe that

$$\int_K \eta^T\sum_{j=1}^{d}\bar{A}^j_K\partial_j\xi = 0 \qquad (2.8.4)$$

since $\bar{A}^j_K$ is constant and $\partial_j\xi_{|K} \in V^K_\delta$. Let $\mathbf{y} \in K$, $\bar{i}$ and $\bar{k}$ be such that

$$\big\|\bar{A}^j_K - A_j\big\|_{[L^\infty(K)]^{m\times m}} = \big|(\bar{A}^j_K)_{\bar{i},\bar{k}} - (A_j)_{\bar{i},\bar{k}}(\mathbf{y})\big|$$

where $\bar{i}$ and $\bar{k}$ are two matrix indices. By continuity of $(A_j)_{\bar{i},\bar{k}}$, there exists $\bar{\mathbf{x}}$ such that $(A_j)_{\bar{i},\bar{k}}(\bar{\mathbf{x}}) = (\bar{A}^j_K)_{\bar{i},\bar{k}}$. Then

$$\big\|\bar{A}^j_K - A_j\big\|_{[L^\infty(K)]^{m\times m}} = \big|(A_j)_{\bar{i},\bar{k}}(\bar{\mathbf{x}}) - (A_j)_{\bar{i},\bar{k}}(\mathbf{y})\big|$$

For all $f : K \to \mathbb{R}$ such that $f \in C^1(K)$: For all $\mathbf{x}, \mathbf{y} \in K$

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq h_K \max_{l=1,..,d} \|\partial_l f\|_{L^\infty(K)}$$

such that we get

$$\begin{aligned} \left\|\bar{A}_K^j - A_j\right\|_{[L^\infty(K)]^{m\times m}} &\leq h_K \max_{l=1,..,d} \|\partial_l (A_j)_{\bar{i},\bar{k}}\|_{L^\infty(K)} \\ &\leq h_K \max_{l=1,..,d} \|\partial_l A_j\|_{[L^\infty(K)]^{m\times m\times d}} \\ &\leq h_K \, \|A_j\|_{1,\infty} \, . \end{aligned} \tag{2.8.5}$$

Then using (2.8.4) and (2.8.5) yields

$$\begin{aligned} &\Big|\sum_{j=1}^d (\eta, A_j \partial_j \xi)_{\tau_h}\Big| \\ &= \Big|\sum_{j=1}^d \sum_{K\in\tau_h} (\eta, A_j\partial_j\xi)_K\Big| \leq \sum_{j=1}^d \sum_{K\in\tau_h} \int_K \left|\eta^T (\bar{A}_K^j - A_j)\partial_j\xi\right| \\ &\leq \sum_{K\in\tau_h} \sum_{j=1}^d \|\bar{A}_K^j - A_j\|_{[L^\infty(K)]^{m\times m}} \sum_{i,k=1}^m \int_K |\eta_i| \, |\partial_j \xi_k| \\ &\leq \sum_{j=1}^d \|A_j\|_{1,\infty} \sum_{K\in\tau_h} h_K \sum_{i,k=1}^m \int_K |\eta_i| \, |\partial_j\xi_k| \\ &\leq \max_l \|A_l\|_{1,\infty} \sum_{K\in\tau_h} h_K \sum_{j=1}^d \sum_{i,k=1}^m \int_K |\eta_i| \, |\partial_j\xi_k| \, \frac{\theta_K}{\theta_K} \\ &\leq \max_l \|A_l\|_{1,\infty} \Big(\sum_{K\in\tau_h} \frac{h_K^2}{\theta_K^2} d^2 \sum_{i=1}^m \|\eta_i\|_{L^2(K)}^2\Big)^{\frac{1}{2}} \Big(\sum_{K\in\tau_h} \theta_K^2 \sum_{j=1}^d \sum_{k=1}^m \|\partial_j\xi_k\|_{L^2(K)}^2\Big)^{\frac{1}{2}} \\ &\leq d \max_l \|A_l\|_{1,\infty} \Big(\sum_{K\in\tau_h} \frac{h_K^2}{\theta_K^2} \|\eta\|_K^2\Big)^{\frac{1}{2}} \Big(\sum_{K\in\tau_h} \theta_K^2 \sum_{k=1}^m \|\nabla\xi_k\|_{[L^2(K)]^d}^2\Big)^{\frac{1}{2}} . \end{aligned}$$

Taking the following inverse inequality for algebraic polynomials, see the book of Quarteroni and Valli [15],

$$\|\nabla\xi_k\|_{[L^2(K)]^d} \leq \frac{N_K^2}{h_K} \|\xi_k\|_{L^2(K)}$$

and choosing

$$\theta_K = \frac{h_K}{N_K^2}$$

leads to

$$\Big|\sum_{j=1}^d (\eta, A_j\partial_j\xi)_{\tau_h}\Big| \leq \frac{d \max_j \|A_j\|_{1,\infty}}{\sigma_0^{1/2}} \Big(\sum_{K\in\tau_h} N_K^4 \|\eta\|_K^2\Big)^{\frac{1}{2}} \|\sigma_0^{\frac{1}{2}}\xi\|_{\tau_h} .$$

Defining

$$C_A = \frac{\max_j \|A_j\|_{1,\infty}}{\sigma_0^{1/2}}$$

leads to

$$\Big|\sum_{j=1}^{d}(\eta, A_j\partial_j\xi)_{\tau_h}\Big| \quad\leq\quad C_A d\Big(\sum_{K\in\tau_h} N_K^4\|\eta\|_K^2\Big)^{\frac{1}{2}}|||\xi|||\,.$$

□ **Lemma** (2.11)

Applying this lemma for $\mathcal{I}_2 = -\sum_{j=1}^{d}(\eta, A_j\partial_j\xi)_{\tau_h}$ yields

$$\mathcal{I}_2 \leq |\sum_{j=1}^{d}(\eta, A_j\partial_j\xi)_{\tau_h}| \leq C_A d\Big(\sum_{K\in\tau_h} N_K^4\|\eta\|_K^2\Big)^{\frac{1}{2}}|||\xi|||$$

- Finally let us develop the last term $\mathcal{I}_3$:

$$\begin{aligned}
\mathcal{I}_3 \quad &= \quad \sum_{K\in\tau_h}\Big\{(D\eta,\xi)_{\partial K} + \big(M^-\,[\eta],\xi\big)_{\Gamma_i^K} + (M^-\,\eta,\xi)_{\Gamma_e^K}\Big\} \\
&= \quad \sum_{K\in\tau_h}\Big\{(M^+\eta,\xi)_{\partial K} - (M^-\eta,\xi)_{\partial K} \\
&\qquad\qquad + \big(M^-\,[\eta],\xi\big)_{\Gamma_i^K} + (M^-\,\eta,\xi)_{\Gamma_e^K}\Big\} \\
&= \quad \sum_{K\in\tau_h}\Big\{(M^+\eta,\xi)_{\Gamma_i^K} + (M^+\eta,\xi)_{\Gamma_e^K} - (M^-\eta,\xi)_{\Gamma_i^K} \\
&\qquad\qquad - (M^-\eta,\xi)_{\Gamma_e^K} + \big(M^-\,[\eta],\xi\big)_{\Gamma_i^K} + (M^-\,\eta,\xi)_{\Gamma_e^K}\Big\} \\
&= \quad \sum_{K\in\tau_h}\Big\{(M^+\eta,\xi)_{\Gamma_i^K} + (M^+\eta,\xi)_{\Gamma_e^K} - (M^-\eta_{ext},\xi)_{\Gamma_i^K}\Big\}
\end{aligned}$$

Observe that

$$\sum_{K'\in\tau_h}(M^+\,\eta,\xi)_{\Gamma_i^{K'}} = \sum_{K\in\tau_h}(M^-\,\eta_{ext},\xi_{ext})_{\Gamma_i^K}\,.$$

Hence

$$\mathcal{I}_3 = \sum_{K\in\tau_h}\Big\{(M^+\,\eta,\xi)_{\Gamma_e^K} + (M^-\,\eta_{ext},\xi_{ext}-\xi)_{\Gamma_i^K}\Big\}$$

Firstly, M^+ being semi-positive definite, one can apply the Cauchy-Schwarz inequality

$$\sum_{K\in\tau_h}\big(M^+\,\eta,\xi\big)_{\Gamma_e^K} \leq \Big(\sum_{K\in\tau_h}\big(M^+\,\eta,\eta\big)_{\Gamma_e^K}\Big)^{\frac{1}{2}}\Big(\sum_{K\in\tau_h}\big(M^+\,\xi,\xi\big)_{\Gamma_e^K}\Big)^{\frac{1}{2}}$$

and observe that

$$\Big(\sum_{K\in\tau_h}\big(M^+\,\eta,\eta\big)_{\Gamma_e^K}\Big)^{\frac{1}{2}} \leq \Big(\sum_{K\in\tau_h}\big(M\,\eta,\eta\big)_{\Gamma_e^K}\Big)^{\frac{1}{2}} \leq \Big(\sum_{K\in\tau_h}\big(M\,\eta,\eta\big)_{\partial K}\Big)^{\frac{1}{2}} = |]\eta[|$$

and

$$\frac{1}{2}\Big(\sum_{K\in\tau_h}\big(M^+\,\xi,\xi\big)_{\Gamma_e^K}\Big)^{\frac{1}{2}} \leq |||\xi|||$$

so that

$$\sum_{K\in\tau_h} \left(M^+ \eta, \xi\right)_{\Gamma_e^K} \le 2\, |]\eta[|\ |||\xi|||\,.$$

Secondly, using that M^- is semi-positive definite

$$\begin{aligned}
\sum_{K\in\tau_h} (M^- \eta_{ext}, \xi_{ext} - \xi)_{\Gamma_i^K} &\le \Big(\sum_{K\in\tau_h} (M^- \eta_{ext}, \eta_{ext})_{\Gamma_i^K}\Big)^{\frac{1}{2}} \Big(\sum_{K\in\tau_h} (M^- [\xi], [\xi])_{\Gamma_i^K}\Big)^{\frac{1}{2}} \\
&\le 2\Big(\sum_{K\in\tau_h} (M^+ \eta, \eta)_{\Gamma_i^K}\Big)^{\frac{1}{2}} |||\xi||| \\
&\le 2\Big(\sum_{K\in\tau_h} (M\, \eta, \eta)_{\Gamma_i^K}\Big)^{\frac{1}{2}} |||\xi||| \\
&\le 2\, |]\eta[|\ |||\xi|||
\end{aligned}$$

Thereby

$$\mathcal{I}_3 \le 4\, |]\eta[|\ |||\xi|||$$

Considering the bounds of all three terms $\mathcal{I}_1, \mathcal{I}_2$ and $\mathcal{I}_3$ leads to the result.

□ **Lemma** (2.9)

The next Lemma is also an intermediate result which estimates the projection error in the seminorm $|]\cdot[|$.

Lemma 2.12 *Suppose that $\mathbf{v}_{|K} \in [H^{l_K}(K)]^m$, $\forall K \in \tau_h$ and for some integers $l_K \ge 1$. Then, for any integer $s_K, 1 \le s_K \le \min(N_K+1, l_K)$ with $N_K \ge 1$ for all $K \in \tau_h$:*

$$|]\mathbf{v} - \mathbf{P}_\delta \mathbf{v}[|^2 \le \mu_0 C^2(d, \mathbf{s}) \sum_{K\in\tau_h} \frac{h_K^{2s_K-1}}{(N_K+1)^{2s_K-1}} |\mathbf{v}|^2_{K,s_K}$$

where $\mu_0 = \|M\|_{0,\infty}$ is a positive constant and $\mathbf{s} = (s_{K_1}, s_{K_2}, ..)$.

Proof. Observe that $\mathbf{v}^T M \mathbf{v} \le \|M\|_{0,\infty} \mathbf{v}^T\mathbf{v} = \mu_0 \mathbf{v}^T\mathbf{v}$ since $D \in [L^\infty(\Omega)]^m$ and let be $\rho = \mathbf{v} - \mathbf{P}_\delta\mathbf{v}$. Thus

$$|]\rho[|^2 = \sum_{K\in\tau_h} \left(M\rho, \rho\right)_{\partial K} \le \mu_0 \sum_{K\in\tau_h} \|\rho\|^2_{[L^2(\partial K)]^m}\,.$$

Now, applying Lemma 2.5 leads to

$$\begin{aligned}
\|\rho\|^2 &\le \mu_0 \sum_{K\in\tau_h} C_K^2(d, s_K) \frac{h_K^{2s_K-1}}{(N_K+1)^{2s_K-1}} |\mathbf{v}|^2_{K,s_K} \\
&\le \mu_0 C^2(d, \mathbf{s}) \sum_{K\in\tau_h} \frac{h_K^{2s_K-1}}{(N_K+1)^{2s_K-1}} |\mathbf{v}|^2_{K,s_K}\,.
\end{aligned}$$

□ **Lemma** (2.12)

Lemma 2.13 *Suppose that* $\mathbf{v}_{|_K} \in [H^{l_K}(K)]^m$, $\forall K \in \tau_h$ *and for some integers* $l_K \geq 1$. *Then, for any integer* $s_K, 1 \leq s_K \leq \min(N_K+1, l_K)$ *with* $N_K \geq 1$ *for all* $K \in \tau_h$*:*

$$|||\mathbf{v} - \mathbf{P}_\delta \mathbf{v}|||^2 \leq \sigma_0 \tilde{C}^2(d) \sum_{K\in\tau_h} \frac{h_K^{2s_K}}{N_K^{2s_K}} |\mathbf{v}|^2_{K,s_K} + \frac{3}{2}\mu_0 C^2(d,\mathbf{s}) \sum_{K\in\tau_h} \frac{h_K^{2s_K-1}}{N_K^{2s_K-1}} |\mathbf{v}|^2_{K,s_K}$$

where $\mathbf{s} = (s_{K_1}, s_{K_2}, ..)$.

Proof. Let $\rho = \mathbf{v} - \mathbf{P}_\delta \mathbf{v}$ and by the definition of $|||\cdot|||$

$$|||\rho|||^2 = \|\sigma_0^{\frac{1}{2}}\rho\|^2_{\tau_h} + \frac{1}{2}\sum_{K\in\tau_h}\left\{\left(M^-[\rho],[\rho]\right)_{\Gamma_i^K} + \left(M\rho,\rho\right)_{\Gamma_e^K}\right\}.$$

Observe that

- By Lemma 2.4

$$\begin{aligned} \|\sigma_0^{\frac{1}{2}}\rho\|^2_{\tau_h} &= \sigma_0 \sum_{K\in\tau_h}\|\rho\|^2_K \leq \sigma_0 \sum_{K\in\tau_h}\left(\tilde{C}_K(d)\frac{h_K^{s_K}}{N_K^{s_K}}|\mathbf{v}|_{K,s_K}\right)^2 \\ &\leq \sigma_0\, \tilde{C}^2(d)\sum_{K\in\tau_h}\frac{h_K^{2s_K}}{N_K^{2s_K}}|\mathbf{v}|^2_{K,s_K} \end{aligned}$$

- Using the following inequality for a semi-positive definite matrix A

$$(\mathbf{a}-\mathbf{b})^T A(\mathbf{a}-\mathbf{b}) \leq 2\left(\mathbf{a}^T A\mathbf{a} + \mathbf{b}^T A\mathbf{b}\right)$$

and applying this with respect to the semi-definite matrix M^- yields

$$\begin{aligned} \sum_{K\in\tau_h}\frac{1}{2}\left(M^-[\rho],[\rho]\right)_{\Gamma_i^K} &\leq \sum_{K\in\tau_h}\left[\left(M^-\rho,\rho\right)_{\Gamma_i^K} + \left(M^-\rho_{ext},\rho_{ext}\right)_{\Gamma_i^K}\right] \\ &= \sum_{K\in\tau_h}\left[\left(M^-\rho,\rho\right)_{\Gamma_i^K} + \left(M^+\rho,\rho\right)_{\Gamma_i^K}\right] \\ &\leq \sum_{K\in\tau_h}\left[\left(M^-\rho,\rho\right)_{\partial K} + \left(M^+\rho,\rho\right)_{\partial K}\right] \\ &= \sum_{K\in\tau_h}\left(M\rho,\rho\right)_{\partial K} \leq \mu_0\sum_{K\in\tau_h}\|\rho\|^2_{\partial K}. \end{aligned}$$

Then by Lemma 2.5

$$\begin{aligned} \sum_{K\in\tau_h}\frac{1}{2}\left(M^-[\rho],[\rho]\right)_{\Gamma_i^K} &\leq \mu_0\sum_{K\in\tau_h}\left(C(d,s_K)\frac{h_K^{2s_K-1}}{(N_K+1)^{2s_K-1}}|\mathbf{v}|_{K,s_K}\right)^2 \\ &= \mu_0 C^2(d,\mathbf{s})\sum_{K\in\tau_h}\frac{h_K^{2s_K-1}}{(N_K+1)^{2s_K-1}}|\mathbf{v}|^2_{K,s_K} \end{aligned}$$

- Also by Lemma 2.5

$$\begin{aligned} \frac{1}{2}\sum_{K\in\tau_h}\left(M\rho,\rho\right)_{\Gamma_e^K} &\leq \sum_{K\in\tau_h}\frac{1}{2}\left(M\rho,\rho\right)_{\partial K} \leq \frac{\mu_0}{2}\sum_{K\in\tau_h}\|\rho\|^2_{\partial K} \\ &\leq \frac{\mu_0}{2}C^2(d,\mathbf{s})\sum_{K\in\tau_h}\frac{h_K^{2s_K-1}}{(N_K+1)^{2s_K-1}}|\mathbf{v}|^2_{K,s_K} \end{aligned}$$

Respecting all three bounds leads to the result.

□ **Lemma** (2.13)

Theorem 2.14 (Global Convergence) *Suppose that* $\mathbf{u} \in [H^1(\Omega)]^m$, $\mathbf{u}_{|_K} \in [H^{l_K}(K)]^m$, $\forall K \in \tau_h$ *and for some integers* $l_K \geq 1$, *and that* $A_j \in [W^{1,\infty}(\tau_h)]^{m\times m}$ *for all* $j = 1,..,d$. *Then, for any integer* $s_K, 1 \leq s_K \leq \min(N_K+1, l_K)$ *with* $N_K \geq 1$ *for all* $K \in \tau_h$*:*

$$|||\mathbf{u} - \mathbf{u}_{DG}||| \leq C\Big(\sum_{K\in\tau_h} \frac{h_K^{2s_K-1}}{N_K^{2s_K-1}} |\mathbf{u}|^2_{K,s_K}\Big)^{\frac{1}{2}} \tag{2.8.6}$$

where $\mathbf{u}$ *is the exact solution of (2.7.3),* $\mathbf{u}_{DG}$ *the solution of (2.7.3),*

$$C = \Big[\tilde{C}(d)h^{\frac{1}{2}}\Big(C_0\frac{1}{N_m^{1/2}} + C_A d N^{\frac{3}{2}}\Big) + 4\,\mu_0^{\frac{1}{2}}C(d,\mathbf{s}) + \Big(\sigma_0\tilde{C}^2(d)\frac{h}{N_m} + \frac{3}{2}\mu_0 C^2(d,\mathbf{s})\Big)^{\frac{1}{2}}\Big]$$

and $N = \max_{K\in\tau_h} N_K$, $N_m = \min_{K\in\tau_h} N_K$ *and* $h = \max_{K\in\tau_h} h_K$.

Remark 2.15 *Assume that a uniform mesh of mesh size* h *combined with a constant polynomial order* N *is used. Observe the presence of the term* $N^{\frac{3}{2}}$. *Due to this term, the result can also be presented as*

$$|||\mathbf{u} - \mathbf{u}_{DG}||| \leq C\frac{h^{s-\frac{1}{2}}}{N^{s-2}}|\mathbf{u}|_{\tau_h,\mathbf{s}}$$

where the constant C *denotes a generic constant and is not equal to the one in (2.8.6). The error estimate is a power of* $3/2$ *suboptimal in* N *with respect to the* $L^2(\partial K)$*-projection error.*

Remark 2.16 *If a uniform mesh of mesh size* h *combined with a constant polynomial order* N *is used and if* $h^{\frac{1}{2}} \leq (C_A d N^{\frac{3}{2}})^{-1}$, *then the result can be presented as:*

$$|||\mathbf{u} - \mathbf{u}_{DG}||| \leq C\frac{h^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}}|\mathbf{u}|_{\tau_h,\mathbf{s}}$$

where $\mathbf{s}$ *denotes the uniform vector* $s\mathbf{1}$.

Remark 2.17 *If a uniform mesh combined with a constant polynomial order is used and if the matrices* $A_j \in [\mathbb{P}_0(\tau_h)]^{m\times m}$, *then the bound of Lemma 2.11 is zero. Since the matrices* A_j *are all constant on each element,* $\sum_{j=1}^d A_j\partial_j\xi \in V_\delta$. *By the orthogonality of the* L^2*-projector*

$$\big(\eta, \sum_{j=1}^{d} A_j\partial_j\xi\big)_{\tau_h} = 0$$

As consequence, the term of $N^{\frac{3}{2}}$ *in the constant* C *vanishes. The result can be formulated as:*

$$|||\mathbf{u} - \mathbf{u}_{DG}||| \leq C\frac{h^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}}|\mathbf{u}|_{\tau_h,\mathbf{s}}$$

Thus, the error estimate is optimal with respect to the L^2*-projector on the boundary and suboptimal of a power of* $1/2$ *in both* h *and* N *with respect to the* L^2*-projector on each* K.

Proof. By the triangle inequality, we have that

$$|||\mathbf{u}-\mathbf{u}_{DG}||| \leq |||\eta||| + |||\xi||| \,.$$

Firstly by Lemma 2.13

$$\begin{aligned} |||\eta||| &\leq \Big(\sigma_0\tilde{C}^2(d)\sum_{K\in\tau_h}\frac{h_K^{2s_K}}{N_K^{2s_K}}|\mathbf{u}|^2_{K,s_K} + \frac{3}{2}\mu_0 C^2(d,\mathbf{s})\sum_{K\in\tau_h}\frac{h_K^{2s_K-1}}{N_K^{2s_K-1}}|\mathbf{u}|^2_{K,s_K}\Big)^{\frac{1}{2}} \\ &\leq \Big(\sigma_0\tilde{C}^2(d)\sum_{K\in\tau_h}\frac{h_K}{N_K}\frac{h_K^{2s_K-1}}{N_K^{2s_K-1}}|\mathbf{u}|^2_{K,s_K} + \frac{3}{2}\mu_0 C^2(d,\mathbf{s})\sum_{K\in\tau_h}\frac{h_K^{2s_K-1}}{N_K^{2s_K-1}}|\mathbf{u}|^2_{K,s_K}\Big)^{\frac{1}{2}} \,. \end{aligned}$$

Let us denote $N = \max_{K\in\tau_h} N_K$, $N_m = \min_{K\in\tau_h} N_K$ and $h = \max_{K\in\tau_h} h_K$. Then

$$|||\eta||| \leq \Big(\sigma_0\tilde{C}^2(d)\frac{h}{N_m} + \frac{3}{2}\mu_0 C^2(d,\mathbf{s})\Big)^{\frac{1}{2}}\Big(\sum_{K\in\tau_h}\frac{h_K^{2s_K-1}}{N_K^{2s_K-1}}|\mathbf{u}|^2_{K,s_K}\Big)^{\frac{1}{2}} \,.$$

Secondly by coercivity (Lemma 2.7), the Galerkin orthogonality (Lemma 2.8) and Lemma 2.9

$$\begin{aligned} |||\xi|||^2 &= a(\xi,\xi) = a(\xi,\xi) + a(\mathbf{u}_{DG}-\mathbf{u},\xi) = a(\eta,\xi) \\ &\leq \Big[C_0\|\eta\|_{\tau_h} + C_A d\Big(\sum_{K\in\tau_h} N_K^4\|\eta\|_K^2\Big)^{\frac{1}{2}} + 4\,|]\eta[|\Big]\; |||\xi||| \end{aligned}$$

Therefore

$$|||\xi||| \leq C_0\|\eta\|_{\tau_h} + C_A d\Big(\sum_{K\in\tau_h} N_K^4\|\eta\|_K^2\Big)^{\frac{1}{2}} + 4\,|]\eta[|$$

Using Lemma 2.12 and Lemma 2.4 yields

$$\begin{aligned} |||\xi||| &\leq C_0\Big(\sum_{K\in\tau_h}\frac{h_K}{N_K}\Big\|\Big(\frac{N_K}{h_K}\Big)^{\frac{1}{2}}\eta\Big\|_K^2\Big)^{\frac{1}{2}} \\ &\quad + C_A d\Big(\sum_{K\in\tau_h} h_K N_K^3\Big\|\Big(\frac{N_K}{h_K}\Big)^{\frac{1}{2}}\eta\Big\|_K^2\Big)^{\frac{1}{2}} + 4\,|]\eta[| \\ &\leq \Big(C_0\Big(\frac{h}{N_m}\Big)^{\frac{1}{2}} + C_A d h^{\frac{1}{2}} N^{\frac{3}{2}}\Big)\Big(\sum_{K\in\tau_h}\Big\|\Big(\frac{N_K}{h_K}\Big)^{\frac{1}{2}}\eta\Big\|_K^2\Big)^{\frac{1}{2}} + 4\,|]\eta[| \\ &\leq \Big(C_0\Big(\frac{h}{N_m}\Big)^{\frac{1}{2}} + C_A d h^{\frac{1}{2}} N^{\frac{3}{2}}\Big)\Big(\sum_{K\in\tau_h}\tilde{C}^2(d)\frac{h_K^{2s_K-1}}{N_K^{2s_K-1}}|\mathbf{u}|^2_{K,s_K}\Big)^{\frac{1}{2}} \\ &\quad + 4\,\mu_0^{\frac{1}{2}} C(d,\mathbf{s})\Big(\sum_{K\in\tau_h}\frac{h_K^{2s_K-1}}{N_K^{2s_K-1}}\,|\mathbf{u}|^2_{K,s_K}\Big)^{\frac{1}{2}} \\ &\leq \Big[\tilde{C}(d)h^{\frac{1}{2}}\Big(C_0\frac{1}{N_m^{1/2}} + C_A d N^{\frac{3}{2}}\Big) + 4\,\mu_0^{\frac{1}{2}} C(d,\mathbf{s})\Big] \\ &\quad \Big(\sum_{K\in\tau_h}\frac{h_K^{2s_K-1}}{N_K^{2s_K-1}}\,|\mathbf{u}|^2_{K,s_K}\Big)^{\frac{1}{2}} \,. \end{aligned}$$

Finally, taking the sum of $|||\eta|||$ and $|||\xi|||$ leads to the result.

□ **Theorem** (2.14)

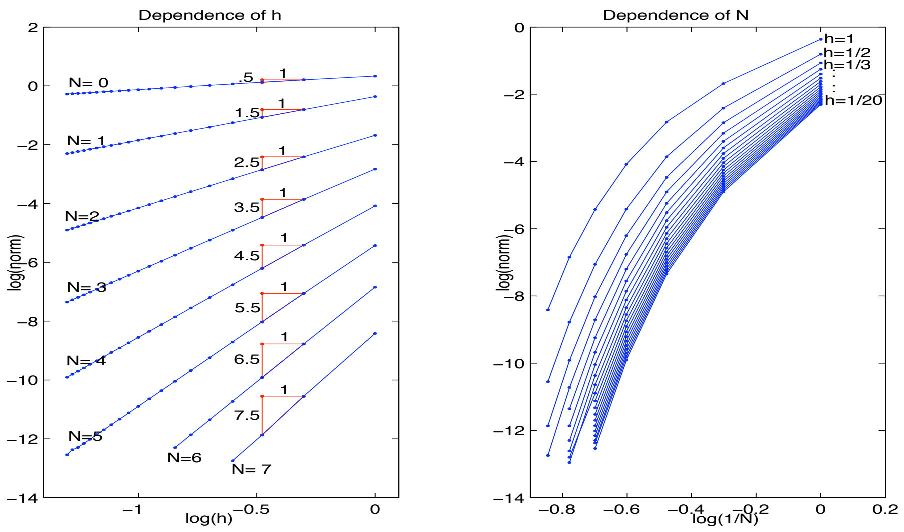

Figure 2.1: *h- resp. N-refinement in the case of example 1.*

2.9 Numerical Results

The convergence accuracy of the spectral discontinuous Galerkin (SDG) method for linear hyperbolic systems is analysed. The aim is to confirm numerically the theoretical result of Theorem 2.14.

As domain Ω, the interval $I = (0,1) \subset \mathbb{R}$ is chosen. The dimension of the domain does not influence the accuracy. Let $x_i = ih$ and $I_i = (x_{i-1}, x_i)$ for a uniform mesh size h. The set of all I_i for i varying from 1 to $N_x = \frac{1}{h}$ builds a partition of Ω. The intervals I_i are the elements. Then the space V_δ is defined by

$$V_\delta = \left\{ \mathbf{v} \in [L^2(\Omega)]^m \;\middle|\; \mathbf{v}_{|I_i} \in \mathcal{Q}_N(I_i) \quad \forall i = 1..N_x \right\} .$$

All problems are of the following form.

$$A\partial_x \mathbf{u} + B\mathbf{u} = \mathbf{f} \qquad \text{in } I \; .$$

In the first problem, a regular solution is considered. The coefficient matrices A and B are then constant. The second example proves the convergence rates of an irregular solution with constant coefficients. Whereas in the third example the matrix A is not constant anymore. A is then piecewise constant. Consequently it depends on the mesh $\mathcal{T}_h$ if $A \in [W^{1,\infty}(\tau_h)]^m$ or not. These two cases are implemented.

2.9.1 The code

A Matlab code is developed to solve the linear hyperbolic system with the SDG-method. For computing the matrix A_{LS} and the right hand side $\mathbf{f}_{LS}$ of the linear system, symbolic calculation is used by employing Maple commands in Matlab. To solve the linear system $A_{LS}\mathbf{u} = \mathbf{f}_{LS}$ the GMRes algorithm is used with restarting after 20 inner iterations.

2.9.2 Example 1

In this section, the convergence behavior of the SDG-method is analysed for a regular solution and for constant coefficients of the hyperbolic system. As coefficients and right-hand side, the following functions are chosen

$$B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ and } \quad \mathbf{f}(x) = \begin{pmatrix} 2\sinh(x) + x^2 - 1 \\ 2\cosh(x) + x \end{pmatrix}.$$

Then the solution reads

$$\mathbf{u}(x) = \begin{pmatrix} e^x + x^2 \\ e^{-x} - x \end{pmatrix}.$$

It is obvious that $\mathbf{u} \in [C^\infty(\Omega)]^2$ and therefore $\mathbf{u} \in [H^1(\Omega)]^2 \cap [H^k(\tau_h)]^2$, for all $k \geq 0$. Due to this regularity of the solution $\mathbf{u}$, the integer s of Theorem 2.14 gets $s = \min(N+1, k) = N+1$ and the convergence accuracy becomes

$$|||\mathbf{u} - \mathbf{u}_{DG}||| \leq C \frac{h^{N+\frac{1}{2}}}{N^{N+\frac{1}{2}}} |\mathbf{u}|_{\tau_h, N+1}.$$

h**-refinement**: This means that we get an algebraic convergence rate of $N + 1/2$ for a fixed N and h-refinement. So that we expect a straight line of slope $N+1/2$ in the in the log-log diagram, which we can observe in the numerical results, see Figure 2.1.

N**-refinement**: For N-refinement, due to the regularity of the solution, an exponential convergence rate is obtained, which can be quantitatively observed, see Figure 2.1.

2.9.3 Example 2

In this example, the convergence rate for an irregular solution with constant coefficients is analysed. The coefficients are defined by

$$B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and the right-hand side is defined by

$$\mathbf{f} = \begin{pmatrix} 2\sinh(x) + \mathbf{1}_{[x>0.5]}\, \alpha\, (x-0.5)^{\alpha-1} + (x-0.5)^\alpha \\ 2\cosh(x) + \mathbf{1}_{[x>0.5]}\, \alpha\, (x-0.5)^{\alpha-1} + (x-0.5)^\alpha \end{pmatrix}$$

so that the solution $\mathbf{u} : I \to \mathbb{R}^2$ is

$$\mathbf{u}_1 = \begin{cases} e^x & \text{if } x \leq 0.5 \\ e^x + (x-0.5)^\alpha & \text{otherwise} \end{cases}$$

and

$$\mathbf{u}_2 = \begin{cases} e^{-x} & \text{if } x \leq 0.5 \\ e^{-x} + (x-0.5)^\alpha & \text{otherwise} \end{cases}.$$

The parameter α is chosen among $\frac{1}{2}, \frac{3}{2}, \ldots$. One can show that $\mathbf{u} \in [H^{\alpha+\frac{1}{2}-\varepsilon}(\tau_h)]^2$, for all $\varepsilon > 0$. The biggest integer $\tilde{\alpha}$ such that $\mathbf{u} \in [H^{\tilde{\alpha}}(\tau_h)]^2$ is $\alpha - \frac{1}{2}$. Then the integer s of Theorem 2.14 is given by $s = \min(N+1, \alpha - \frac{1}{2})$ and the convergence result becomes

$$|||\mathbf{u} - \mathbf{u}_{DG}||| \leq C \frac{h^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}} |\mathbf{u}|_{\tau_h, s}.$$

Figure 2.2 shows the accuracy for different values of the parameter α, where α takes the values of $5/2$, $7/2$ and $9/2$.

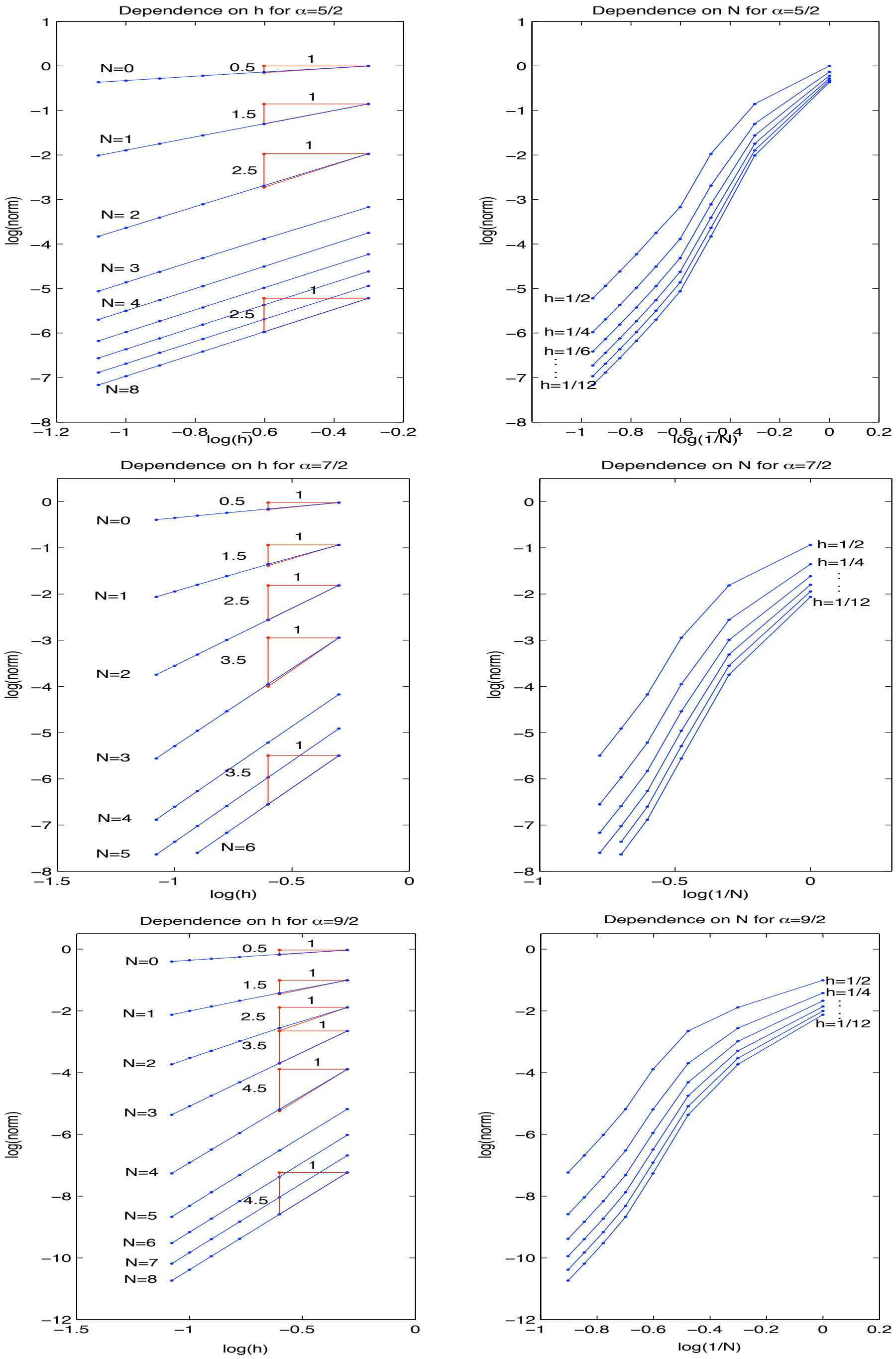

Figure 2.2: *h- resp. N-refinement in the case of example 2 with α=$\frac{5}{2}$,$\frac{7}{2}$ and $\frac{9}{2}$.*

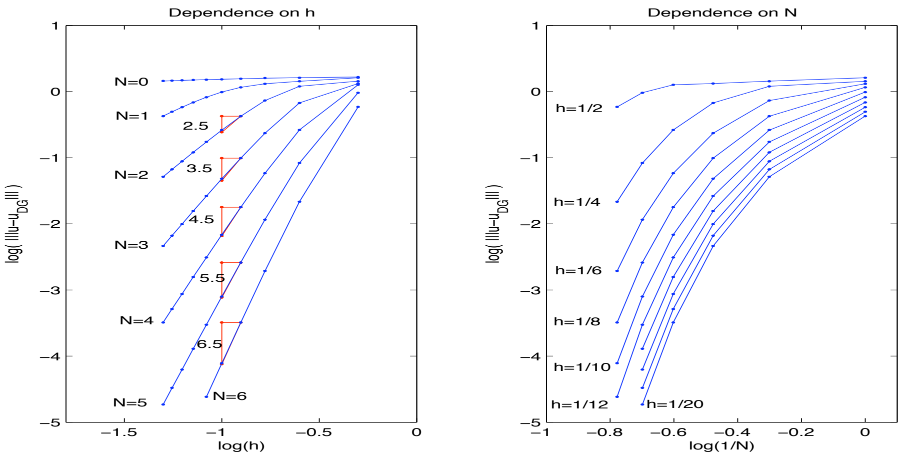

Figure 2.3: *Convergence rates for example 3 with elementwise constant coefficients.*

Knowing where the qualitative behavior of the solution changes, the mesh is adapted to this. A uniform mesh is used where the mesh size is $h = 1/(2n)$ for integers $n \geq 1$. Then $x = 0.5$ is never contained interior an element. But note that even if $x = 0.5$ is never contained in an element, the solution is still only an element of $\mathbf{u} \in [H^{\alpha+\frac{1}{2}-\varepsilon}(\tau_h)]^2$, since it satisfies only $\mathbf{u}_{|_K} \in [H^{\alpha+\frac{1}{2}-\varepsilon}(K)]^2$ for the particular element $K = (0.5, 0.5+h) \in \tau_h$.

h**-refinement**: The convergence rate for h-refinement should be $s = \min(N+1, \alpha - 1/2)$ and we should observe straight lines with slope s. This means that while increasing the polynomial order N, first the convergence rate increases, then it stays fixed due to the low regularity of the solution. But one can observe that the numerical results shows convergence rates as $\mathbf{u} \in [H^{\alpha+\frac{1}{2}}(\tau_h)]^2$. This means that we observe straight lines with slope $\min(N+1, \alpha+1/2)$.

N**-refinement**: Fixing h and varying N in the convergence result implies that first the error should increase exponentially and for $N = \alpha - 1/2$ the convergence becomes algebraic with rate $\alpha + 1/2$. This can be observed in Figure 2.2.

2.9.4 Example 3

In this example, we consider the following problem on $I = (0,1)$:

find $(u,v) : I \times (0,T) \to \mathbb{C}^2$ such that

$$u_t + cv_x = 0$$
$$v_t + cu_x = 0$$

where

$$c(x) = \begin{cases} \frac{\omega}{k_1} & \text{if } x < 0.5 \\ \frac{\omega}{k_2} & \text{otherwise} \end{cases} .$$

Remark 2.18 *If the coefficient $c(x)$ were constant, then the above defined equation would be equivalent to the following wave equation:*

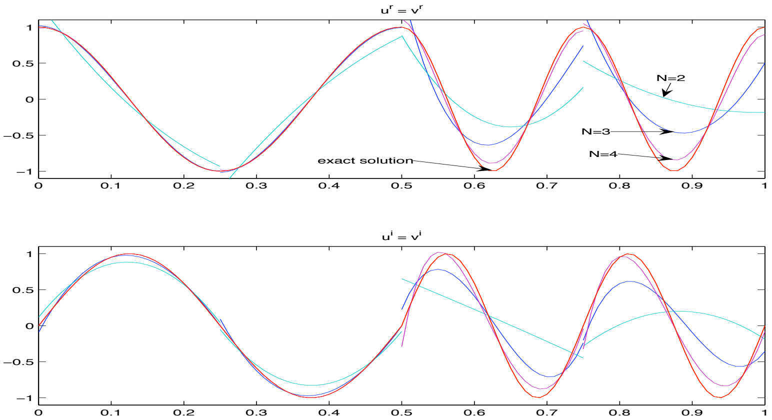

Figure 2.4: *The exact solution and its approximations for $h = 1/4$ and $N = 2, 3, 4$ for example 3.*

find $u : I \times (0, T) \to \mathbb{C}$ such that

$$\partial_{tt}u - c^2\, \partial_{xx}u \quad = \quad 0\,. \tag{2.9.1}$$

Let us assume that u and v are of the following form

$$u(x,t) = e^{-i\omega t}u^*(x) \text{ and } v(x,t) = e^{-i\omega t}v^*(x)\,.$$

Then, the problem is equivalent to

find $(u^*, v^*) : I \to \mathbb{C}^2$ such that

$$\begin{aligned} -i\omega u^* + cv_x^* &= 0 \\ -i\omega v^* + cu_x^* &= 0\,. \end{aligned}$$

Next, we split the complex functions into the real and imaginary part

$$u^*(x) = u^r(x) + iu^i(x) \quad \text{ and } \quad v^*(x) = v^r(x) + iv^i(x)$$

so that finally the problem becomes

find $\bar{\mathbf{u}} = (u^r, u^i, v^r, v^i) : I \to \mathbb{R}^4$ such that

$$\omega B\bar{\mathbf{u}} + cA\partial_x\bar{\mathbf{u}} = 0 \tag{2.9.2}$$

with Dirichlet boundary conditions imposed on the incoming characteristics and where

$$B = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}\,.$$

Note that the solution of (2.9.1) with the corresponding boundary conditions is

$$u(x,t) = \begin{cases} e^{i(k_1 x - \omega t)} & \text{if } x < 0.5 \\ e^{i(k_2 x - \omega t)} & \text{otherwise} \end{cases}$$

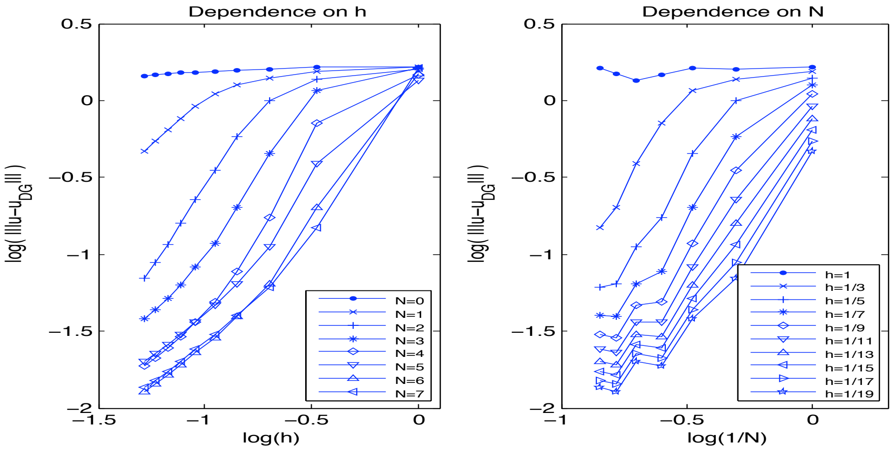

Figure 2.5: *Convergence rates for example 3 where the discontinuity lies in the interior of an element.*

so that the solution of (2.9.2) becomes

$$u^r(x) = v^i(x) = \cos\left(\frac{\omega}{c}x\right) \quad \text{and} \quad u^i(x) = v^r(x) = \sin\left(\frac{\omega}{c}x\right) .$$

As cA is not constant, it depends on the mesh $\mathcal{T}_h$ whether cA lies in $[W^{1,\infty}(\mathcal{T}_h)]^{m\times m}$ or not. For the numerical tests, two different types of meshes are used. The first is as the one already used in example 2 where a uniform mesh is used such that the discontinuity at $x = 0.5$ is never contained interior an element. The second mesh-type is one where the discontinuity lies in the interior of an element.

The first mesh type

A uniform mesh is used where the mesh size is $h = 1/(2n)$ for integers $n \geq 1$. The coefficient are then elementwise constant, thereby $\bar{\mathbf{u}} \in [H^k(\mathcal{T}_h)]^4, \forall k \geq 0$ and $cA \in [W^{1,\infty}(\mathcal{T}_h)]^{m\times m}$. Note that we still satisfy the conditions of Theorem 2.14. Thus the convergence estimation is the same as in example 1, that is

$$|||\mathbf{u} - \mathbf{u}_{DG}||| \leq C\frac{h^{N+\frac{1}{2}}}{N^{N+\frac{1}{2}}}|\mathbf{u}|_{\mathcal{T}_h,N+1} .$$

Figure 2.3 shows the convergence result for h-refinement and N-refinement of this example. Figure 2.4 shows the approximations of u^r, u^i, v^r and v^i for $h = 1/4$ and $N = 2, 3, 4$.

h-refinement: One can observe the estimated convergence rates for small h and for $N \geq 1$. For $N = 0$, the fact that the estimated accuracy is not obtained for big h can be explained that the period of the solution in $(0.5, 1)$ is $1/4$. Mainly for small N there is not enough liberty to catch the frequency effects.

N-refinement: An exponential convergence can be observed what corresponds to the theoretical result of Theorem 2.14.

The second mesh type

The second mesh type is one where the discontinuity at $x = 0.5$ lies in the interior of an element. A uniform mesh is used where the mesh size is $h = 1/(2n+1)$ for $n \in \mathbb{N}$. The coefficient cA is no longer elementwise constant. Thereby we do not anymore satisfy the conditions of Theorem 2.14, since $cA \notin [W^{1,\infty}(\tau_h)]^{m\times m}$. Figure 2.5 shows the convergence result for h-refinement and N-refinement of this example.

h**-refinement**: One can observe that the method does not converge for $N = 0$. For increased polynomial orders N and sufficient small mesh size h, one observes the same accuracy for N and $N+1$. Comparing to the case of the first mesh type, the convergence rate is low.

N**-refinement**: The numerical test shows that for $h = 1$ the method does not converge for the polynomial orders we have tried, this means up to $N = 7$.

2.9.5 Concluding Remark

Theorem 2.14 estimates the convergence rate of $|||\mathbf{u} - \mathbf{u}_{DG}|||$ depending on the mesh size h and the polynomial order N. Theoretically, the convergence rate of
$||\mathbf{u} - \mathbf{u}_{DG}||_{[L^2(\Omega)]^m}$ is not developed in the context of this work. But numerical results show the following behavior

$$||\mathbf{u} - \mathbf{u}_{DG}||_{[L^2(\Omega)]^m} \approx Ch^s$$

for h-refinement and sufficient smooth matrices A_j. Compared to the triple norm, a factor of $\frac{1}{2}$ is gained. The lost of this factor $\frac{1}{2}$ for the triple norm is due to the L^2-estimation on the boundary of each element.

2.10 Extension to Time-Dependent Linear Hyperbolic Systems

The convergence analysis of the time-dependent scalar transport equation (section 1.4.2) could be extended to the case of hyperbolic systems using the same notations and definitions as in section 2.8.

Let us consider the following time-dependent linear hyperbolic system:

find $\mathbf{u} : \Omega \times (0,T) \to \mathbb{R}^m$ such that

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^{d} A_j \partial_j \mathbf{u} + B\mathbf{u} = \mathbf{f} \qquad \text{in} \quad \Omega \tag{2.10.1}$$

with boundary conditions according to section 2.3. Then its semi-discrete SDG-scheme reads:

$\forall t > 0$, find $\mathbf{u}_{DG}(t) : \Omega \to \mathbb{R}^m$ such that

$$\big(\partial \mathbf{u}_{DG}(t), \mathbf{v}_\delta\big)_{\tau_h} + a\big(\mathbf{u}_{DG}(t), \mathbf{v}_\delta\big) = F(\mathbf{v}_\delta) \qquad \forall \mathbf{v}_\delta \in V_\delta$$

where $a(\cdot,\cdot)$ and $F(\cdot)$ are defined by 2.7.4. Choosing a basis for the finite element space V_δ, this formulation leads to as ordinary differential equation (ODE) with respect to the time variable and can be solved by a Runge-Kutta method.

Using a same approach as in section 1.4.2 combined with the approximation results of section 2.8

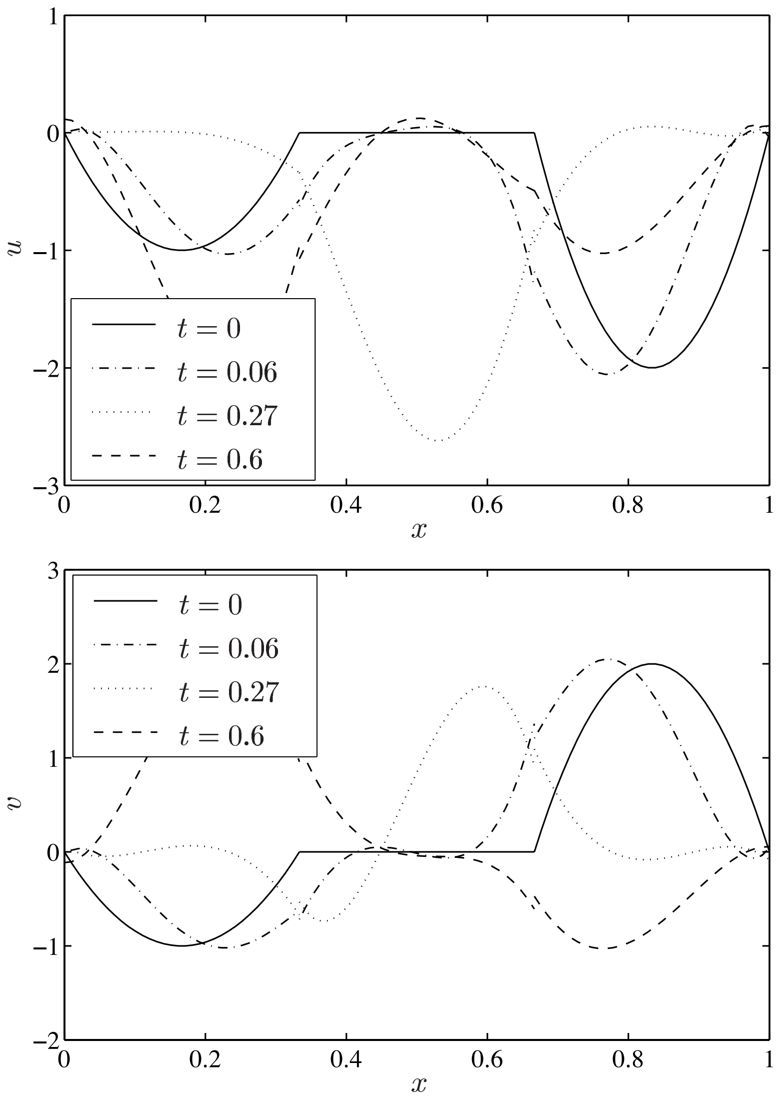

Figure 2.6: *Solution of (2.10.1) computed with $N = 4$ and $h = 1/3$ at different time levels.*

leads to an a priori error estimation analogous to Theorem 1.8 resp. Corollary 1.9 and so convergence of the space discretization would be guaranteed. Let us consider the following example:

find $\mathbf{u} : (0,1) \times (0,T) \to \mathbb{R}^2$ such that

$$\frac{\partial \mathbf{u}}{\partial t} + A\frac{\partial \mathbf{u}}{\partial x} = 0$$

where

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} .$$

Homogeneous Dirichlet boundary conditions are imposed on the incoming characteristics. The initial conditions are drawn in Figure 2.6.

Figure 2.6-2.8 shows the computed solutions of the above defined problem using the Matlab function ”ode45” (this is a Runge-Kutta method) for solving the ordinary differential equation.

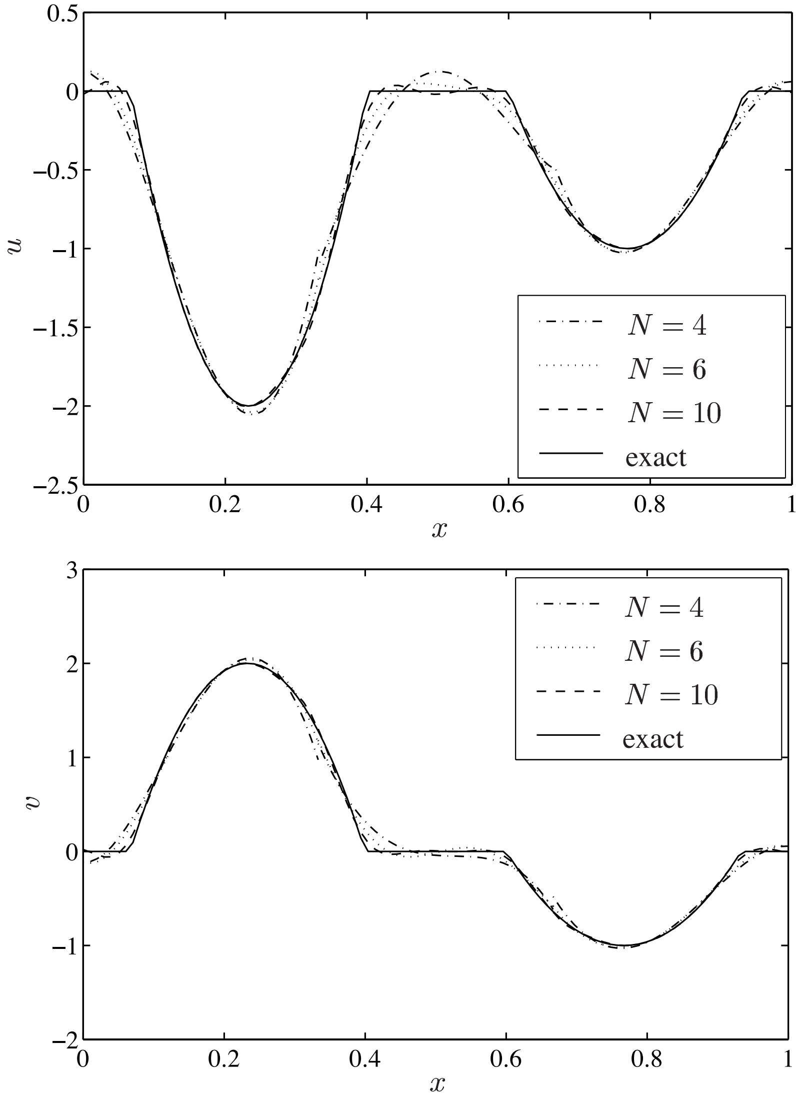

Figure 2.7: *Solutions u (above) and v (below) of (2.10.1) at $t = 0.6$ for $h = \frac{1}{3}$ and different values of N.*

Figure 2.6 shows the computed solution for $N = 4$ and $h = \frac{1}{3}$ at different time levels. Figure 2.7 shows the dependence on N of the accuracy of the computed solutions at the time level $t = 0.6$ while in Figure 2.8 we plot the computed solutions using a constant number of degrees of freedom $2(N + 1)/h = 24$ for different values of h and N at $t = 0.6$. One can observe that for big mesh sizes h and increased polynomial orders N, the method is more precise that for small mesh size h and low polynomial order N.

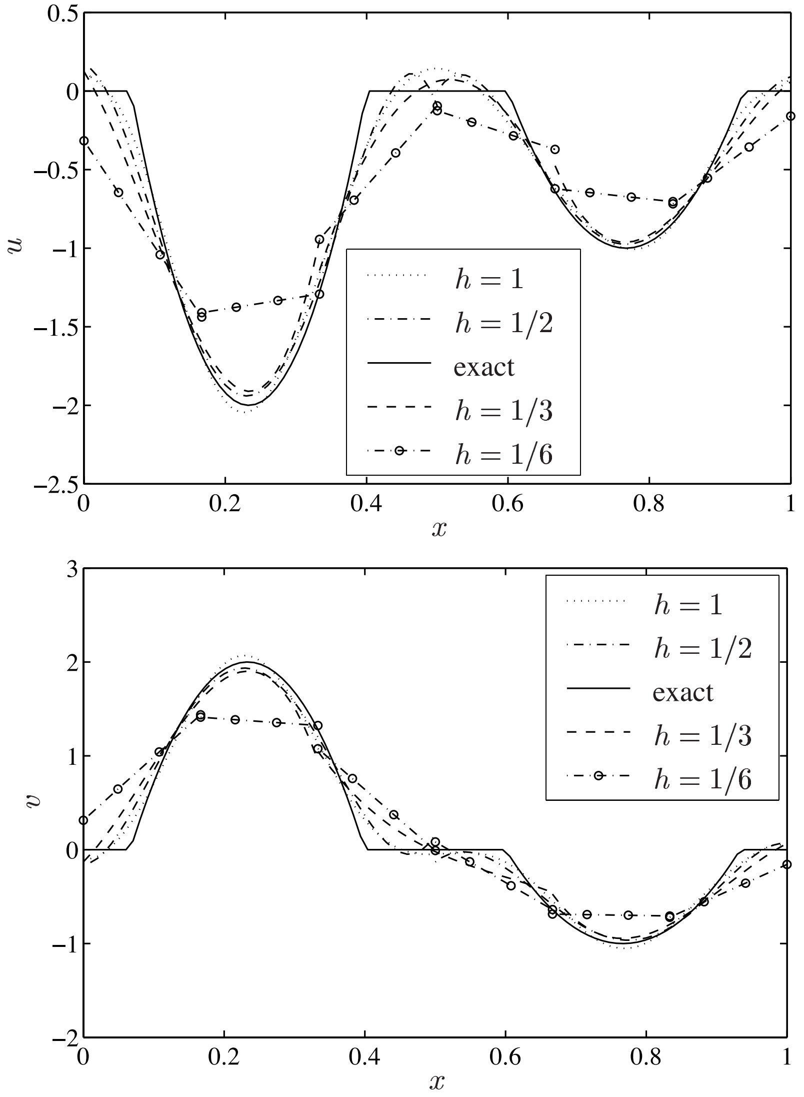

Figure 2.8: *Solutions* u *(above) and* v *(below) of (2.10.1) at* $t = 0.6$ *computed using a constant number of degrees of freedom* $2(N+1)h = 24$.

2.11 Conclusion

As model problem, a linear hyperbolic system was considered and the spectral discontinuous Galerkin (SDG) method was developed for this kind of equations. Due to the jumps across every face of the grid, the derivation and formulation of the SDG-method becomes more technical than that of the continuous finite element method. The formulation leads to an algebraic linear system. This linear system can be solved with standard methods like GMRes and others. We show that under some stability conditions, generalized boundary conditions can be imposed. This means that not only boundary conditions on the incoming characteristics can be imposed but also on certain linear combinations of the physical variables.

A convergence result was developed. The error estimate is suboptimal by a power of$\frac{1}{2}$ in the two refinement parameters h (mesh size) and N (polynomial order) for sufficient smooth entries compared to the L^2-projection on each element. Three numerical test cases confirm quantitatively the predicted convergence rates.

Then the SDG-method for the space discretization of a time-dependent linear hyperbolic system is formulated briefly. This leads to an ordinary differential equation with respect to the time time variable which can be solved by a Runge-Kutta method. A numerical example illustrates the computed solutions. We show qualitatively that N-refinement is much more precise than h-refinement. A possible application could be the non-stationary equations of gas dynamics for example. The extension to nonlinear hyperbolic systems would yield new important applications such as the Burgers' equation, the shallow water equations or the 1D model for the Navier-Stokes equation for blood flow.

Chapter 3

A Posteriori Estimations for Linear Hyperbolic Problems

3.1 Introduction

In practise, it is often interesting to minimize the quantity $|F_D(\mathbf{u} - \mathbf{u}_{DG})|$ where $\mathbf{u}$ is the exact solution of a linear hyperbolic system, $\mathbf{u}_{DG}$ its SDG-approximation and $F_D(\cdot)$ a linear functional. In the context of this work, we chose $F_D(\cdot)$ as the integral of the outgoing characteristics on the boundary. Note that for the quantity $|||\mathbf{u}-\mathbf{u}_{DG}|||$ a convergence result is already derived in section 2.8. The main goal of this section is to develop a convergence result for $F_D(\mathbf{u} - \mathbf{u}_{DG})$ and show using a duality argument that for weaker control of the error than that of $||| \cdot |||$ we can expect higher order convergence.

We present in section 3.2 the dual problem and develop the spectral discontinuous Galerkin (SDG) method for this problem. The relation between the primal and the dual SDG-formulation is shown in Lemma 3.1. The convergence rate depending on the local mesh size h_K and the local polynomial order N_K is the same as for the primal problem and presented in Theorem 3.5.

In section 3.3 we propose an a posteriori error estimation for $F_D(\mathbf{u}-\mathbf{u}_{DG})$. For each element, we can quantify its contribution to the error such that the global error is given by the sum of all local errors. Inspired by the article of Houston and Süli [11] we develop Theorem 3.9 that describes the convergence of $F_D(\mathbf{u} - \mathbf{u}_{DG})$ depending on the local mesh size h_K and the local polynomial order N_K. In [11], the result is developed for a scalar transport equation whereas here we extend it to hyperbolic systems. In general the exact dual solution is not known. That is why its solution has to be approximated by a SDG-method. For this, we keep the same mesh and increase the polynomial order up to k. Using this SDG-approximation of the dual solution we estimate the error by $F_D^k(\mathbf{u} - \mathbf{u}_{DG})$ which should be close to $F_D(\mathbf{u} - \mathbf{u}_{DG})$. Theorem 3.13 shows how the difference of this two functionals depend on k, N_K and h_K.

Finally we implement three test cases with the purpose of assessing quantitatively the predicted theoretical properties.

We use the same notations and technical results as in chapter 2. We refer to section 2.6 for the required definitions, notations and technical results.

3.2 The Dual Problem

Here, we define the dual problem that will be used in the next section for the a posteriori estimation. Let us recall the equations of a linear hyperbolic system with boundary conditions imposed on the incoming characteristics:

find $\mathbf{u}:\Omega\to\mathbb{R}^m$ such that

$$\begin{array}{rcll} B\mathbf{u}+\sum_{j=1}^{d}A_j\partial_j\mathbf{u} & = & \mathbf{f} & \text{in } \Omega \\ \mathbf{z}^- & = & \mathbf{g} & \text{on } \Gamma=\partial\Omega\,. \end{array} \tag{3.2.1}$$

The dual equation is then defined by

find $\varphi:\Omega\to\mathbb{R}^m$ such that

$$\begin{array}{rcll} B^T\varphi-\sum_{j=1}^{d}\partial_j(A_j\varphi) & = & \mathbf{0} & \text{in } \Omega \\ \psi^+ & = & \mathbf{1} & \text{on } \Gamma=\partial\Omega \end{array} \tag{3.2.2}$$

where $\mathbf{z}^-$ and ψ^+ are the negative resp. positive parts of the characteristic variables associated to $\mathbf{u}$ resp. φ, see section 2.3 for more details.
The dual solution can be approximated by a SDG-method. Let us define the local problem on one element K:

find $\varphi:\Omega\to\mathbb{R}^m$ such that

$$\begin{array}{rcll} B^T\varphi-\sum_{j=1}^{d}\partial_j(A_j\varphi) & = & \mathbf{0} & \text{in } \Omega \\ \psi^+ & = & \psi_{ext} & \text{on } \Gamma_i^K \\ \psi^+ & = & \mathbf{1} & \text{on } \Gamma_e^K\,. \end{array}$$

We can follow the same approach as in 2.7.2 to find the spectral discontinuous Galerkin scheme. Let $\tilde{A}_j=-A_j$, $\tilde{D}=\sum_{j=1}^{d}n_j\tilde{A}_j$ with $\tilde{D}=S\tilde{\Lambda}S^T$, $\tilde{M}^-=S\tilde{\Lambda}^-S^T$ and $\tilde{M}^+=S\tilde{\Lambda}^+S^T$. Then the SDG-scheme reads:

find $\varphi_{\mathbf{DG}}\in V_\delta$ such that

$$\begin{aligned}\Big(B^T\varphi_{DG}+\sum_{j=1}^{d}\partial_j(\tilde{A}_j\varphi_{DG}),\zeta_\delta\Big)_{\tau_h}+\sum_{K\in\tau_h}\Big\{\big(\tilde{M}^-(\varphi_{DG}-\varphi_{ext}),\zeta_\delta\big)_{\Gamma_i^K}+\big(\tilde{M}^-\varphi_{DG},\zeta_\delta\big)_{\Gamma_e^K}\Big\}\\ =\sum_{K\in\tau_h}\big(\tilde{M}^-S\mathbf{g}_D,\zeta_\delta\big)_{\Gamma_e^K}\qquad\forall\zeta_\delta\in V_\delta\end{aligned}$$

where $\mathbf{g}_D$ is composed by zero's and one's depending on whether the associated characteristic variable is negative or positive.
Observe that $\tilde{D}=-D$ where $D=\sum_{j=1}^{d}n_jA_j$ and $D=S\Lambda S^T$. As consequence $\tilde{\Lambda}^-=\Lambda^+$ and $\tilde{M}^-=M^+$ and we may write the problem.

find $\varphi_{\mathbf{DG}}\in V_\delta$ such that

$$\begin{aligned}\Big(B^T\varphi_{DG}-\sum_{j=1}^{d}\partial_j(A_j\varphi_{DG}),\zeta_\delta\Big)_{\tau_h}+\sum_{K\in\tau_h}\Big\{\big(M^+(\varphi_{DG}-\varphi_{ext}),\zeta_\delta\big)_{\Gamma_i^K}+\big(M^+\varphi_{DG},\zeta_\delta\big)_{\Gamma_e^K}\Big\}\\ =\sum_{K\in\tau_h}\big(M^+S\mathbf{g}_D,\zeta_\delta\big)_{\Gamma_e^K}\qquad\forall\zeta_\delta\in V_\delta\,.\end{aligned}$$

We now define

$$\begin{aligned} a_D(\mathbf{v},\mathbf{w}) &= \big(B^T\mathbf{v} - \sum_{j=1}^{d} \partial_j(A_j\mathbf{v}), \mathbf{w}\big)_{\tau_h} + \sum_{K\in\tau_h} \Big\{(M^+(\mathbf{v}-\mathbf{v}_{ext}),\mathbf{w})_{\Gamma_i^K} + (M^+\mathbf{v},\mathbf{w})_{\Gamma_e^K}\Big\} \\ F_D(\mathbf{v}) &= \sum_{K\in\tau_h} (M^+S\mathbf{g}_D,\mathbf{v})_{\Gamma_e^K} \end{aligned}$$

So that the problem takes the condensed form:

find $\varphi_{DG} \in V_\delta$ such that

$$a_D(\varphi_{DG},\zeta_\delta) = F_D(\zeta_\delta) \qquad \forall\zeta_\delta \in V_\delta\,. \tag{3.2.3}$$

The following Lemma shows the relation between the bilinear form of the primal problem $a(\cdot,\cdot)$ and $a_D(\cdot,\cdot)$.

Lemma 3.1 *The bilinear form $a_D(\cdot,\cdot)$ of the SDG-scheme for the dual problem satisfies*

$$a_D(\mathbf{v},\mathbf{w}) = a(\mathbf{w},\mathbf{v}) \quad \forall\mathbf{v},\mathbf{w}\in W_\Omega$$

where $a(\cdot,\cdot)$ is the bilinear form associated to the primal problem.

Proof. Integrating by parts yields

$$\begin{aligned} a_D(\mathbf{v},\mathbf{w}) &= \big(\mathbf{v}, B\mathbf{w} + \sum_{j=1}^{d} A_j\partial_j\mathbf{w}\big)_{\tau_h} + \sum_{K\in\tau_h} \Big\{\big(M^+(\mathbf{v}-\mathbf{v}_{ext}),\mathbf{w}\big)_{\Gamma_i^K} + \big(M^+\mathbf{v},\mathbf{w}\big)_{\Gamma_e^K} \\ &\qquad -\big(M^+\mathbf{v},\mathbf{w}\big)_{\Gamma_i^K} - \big(M^+\mathbf{v},\mathbf{w}\big)_{\Gamma_e^K} + \big(M^-\mathbf{v},\mathbf{w}\big)_{\Gamma_i^K} + \big(M^-\mathbf{v},\mathbf{w}\big)_{\Gamma_e^K}\Big\} \\ &= \big(\mathbf{v}, B\mathbf{w} + \sum_{j=1}^{d} A_j\partial_j\mathbf{w}\big)_{\tau_h} \\ &\qquad - \sum_{K\in\tau_h} \Big\{\big(M^+\mathbf{v}_{ext},\mathbf{w}\big)_{\Gamma_i^K} + \big(M^-\mathbf{v},\mathbf{w}\big)_{\Gamma_i^K} + \big(M^-\mathbf{v},\mathbf{w}\big)_{\Gamma_e^K}\Big\}\,. \end{aligned}$$

Using an analogous argument as for (2.8.1) one can write

$$\sum_{K\in\tau_h} (M^+\mathbf{v}_{ext},\mathbf{w})_{\Gamma_i^K} = \sum_{K\in\tau_h} (M^-\mathbf{v},\mathbf{w}_{ext})_{\Gamma_i^K}\,.$$

Then

$$\begin{aligned} a_D(\mathbf{v},\mathbf{w}) &= \big(\mathbf{v}, B\mathbf{w} + \sum_{j=1}^{d} A_j\partial_j\mathbf{w}\big)_{\tau_h} + \sum_{K\in\tau_h} \Big\{\big(\mathbf{v}, M^-(\mathbf{w}-\mathbf{w}_{ext})\big)_{\Gamma_i^K} + \big(\mathbf{v}, M^-\mathbf{w}\big)_{\Gamma_e^K}\Big\} \\ &= a(\mathbf{w},\mathbf{v}) \end{aligned}$$

□ **Lemma** (3.1)

Using the previous Lemma, the discrete dual problem becomes:

find $\varphi_{DG} \in V_\delta$ such that

$$a(\zeta_\delta,\varphi_{DG}) = F_D(\zeta_\delta) \qquad \forall\zeta_\delta \in V_\delta\,.$$

The Galerkin orthogonality holds also for the dual problem.

Lemma 3.2 (Galerkin orthogonality for the dual problem) *Let φ be the exact solution of problem (3.2.2) and φ_{DG} the solution of (3.2.3). If $\varphi \in [H^1(\Omega)]^m$, then*

$$a_D(\varphi_{DG} - \varphi, \mathbf{v}_\delta) = 0 \qquad \forall\, \mathbf{v}_\delta \in V_\delta \,.$$

Remark 3.3 *Applying Lemma 3.1, the Galerkin orthogonality can be formulated as*

$$a(\mathbf{v}_\delta, \varphi_{DG} - \varphi) = 0 \qquad \forall\, \mathbf{v}_\delta \in V_\delta \,.$$

Proof. φ_{DG} satisfies (3.2.3), this means

$$a_D(\varphi_{DG}, \mathbf{v}_\delta) = F_D(\mathbf{v}_\delta) \qquad \forall \mathbf{v}_\delta \in V_\delta \,. \tag{3.2.4}$$

Additionally, the fact that φ is the exact solution of (3.2.2) implies that

$$\Big(B^T\varphi - \sum_{j=1}^{d} \partial_j(A_j\varphi), \mathbf{v}_\delta\Big)_{\tau_h} = 0 \qquad \forall \mathbf{v}_\delta \in V_\delta \,.$$

Since $\varphi \in [H^1(\Omega)]^m$, $\varphi \in [H^1(K)]^d$ for all $K \in \tau_h$. Then the trace on every element K is well defined. Since M^+ is positive semi-definite

$$0 = \sum_{K\in\tau_h} \big([\varphi],[\varphi]\big)_{\Gamma_i^K} \leq \sum_{K\in\tau_h} \big(M^+\,[\varphi],[\varphi]\big)_{\Gamma_i^K} \leq \|M^+\|_{0,\infty} \sum_{K\in\tau_h} \big([\varphi],[\varphi]\big)_{\Gamma_i^K} = 0$$

and therefore

$$\sum_{K\in\tau_h} \big(M^+\,[\varphi],[\varphi]\big)_{\Gamma_i^K} = 0 \,.$$

The boundary condition $\psi^+ = \mathbf{1}$ implies

$$\begin{aligned}
\big(M^-\varphi, \mathbf{v}_\delta\big)_{\Gamma_e^K} &= \big(|\Lambda^-|\psi, \mathbf{y}_\delta\big)_{\Gamma_e^K} = \big(|\lambda^-|\psi^-, \mathbf{y}_\delta^-\big)_{\Gamma_e^K} \\
&= \big(|\lambda^-|\mathbf{1}, \mathbf{y}_\delta^-\big)_{\Gamma_e^K} = \big(|\Lambda^-|\mathbf{g}_D, \mathbf{y}_\delta\big)_{\Gamma_e^K} \\
&= \big(S|\Lambda^-|\mathbf{g}_D, \mathbf{v}_\delta\big)_{\Gamma_e^K} \qquad \forall \mathbf{v}_\delta \in V_\delta
\end{aligned}$$

where ψ and $\mathbf{y}_\delta$ are the characteristic variables of φ and $\mathbf{v}_\delta$. The vector $\mathbf{g}_D$ is composed of ones. Then we conclude that

$$a_D(\varphi, \mathbf{v}_\delta) = F_D(\mathbf{v}_\delta) \qquad \forall \mathbf{v}_\delta \in V_\delta \tag{3.2.5}$$

and taking the difference between (3.2.4) and (3.2.5) leads to the result.

□ **Lemma** (3.2)

The next lemma is an intermediary result which will serve in the proof for the convergence result for the dual SDG-problem.

Lemma 3.4 *If $A_j \in [W^{1,\infty}(\tau_h)]^{m\times m}$ for all $j = 1, .., d$, then*

$$a(\xi, \eta) \leq \Big[C_0\|\eta\|_{\tau_h} + C_A d\Big(\sum_{K\in\tau_h} N_K^4\|\eta\|_K^2\Big)^{\frac{1}{2}} + 4\,|]\eta[|\Big]\; |||\xi||| \,.$$

Proof. By definition of the bilinear form $a(\cdot,\cdot)$, see (2.7.4):

$$a(\xi,\eta) \;=\; \big(B\xi,\eta\big)_{\tau_h} + \sum_{j=1}^{d}\big(A_j\partial_j\xi,\eta\big)_{\tau_h} + \sum_{K\in\tau_h}\Big\{\big(M^-\,[\xi],\eta\big)_{\Gamma_i^K} + \big(M^-\,\xi,\eta\big)_{\Gamma_e^K}\Big\}\,.$$

Firstly, using Lemma 2.10, we conclude that

$$(B\xi,\eta)_{\tau_h} \leq C_0\|\eta\|_{\tau_h}\;|||\xi|||$$

and secondly applying Lemma 2.11 leads to

$$\sum_{j=1}^{d}(A_j\partial_j\xi,\eta)_{\tau_h} \leq \Big|\sum_{j=1}^{d}(A_j\partial_j\xi,\eta)_{\tau_h}\Big| \leq C_A d\Big(\sum_{K\in\tau_h} N_K^4\|\eta\|_K^2\Big)^{\frac{1}{2}}|||\xi|||\,.$$

For the third term the Cauchy-Schwarz inequality is used

$$\begin{aligned}
\sum_{K\in\tau_h}(M^-\,[\xi],\eta)_{\Gamma_i^K} \;&\leq\; \Big(\sum_{K\in\tau_h}(M^-\,[\xi],[\xi])_{\Gamma_i^K}\Big)^{\frac{1}{2}}\Big(\sum_{K\in\tau_h}(M^-\,\eta,\eta)_{\Gamma_i^K}\Big)^{\frac{1}{2}}\\
&\leq\; 2\,|||\xi|||\,\Big(\sum_{K\in\tau_h}(M\;\eta,\eta)_{\Gamma_i^K}\Big)^{\frac{1}{2}}\\
&\leq\; 2\,|||\xi|||\,\Big(\sum_{K\in\tau_h}(M\;\eta,\eta)_{\partial K}\Big)^{\frac{1}{2}}\\
&\leq\; 2\,|||\xi|||\;|]\eta[|\,.
\end{aligned}$$

Finally the fourth term is developed:

$$\begin{aligned}
\sum_{K\in\tau_h}(M^-\,\xi,\eta)_{\Gamma_e^K} \;&\leq\; \Big(\sum_{K\in\tau_h}(M^-\,\xi,\xi)_{\Gamma_e^K}\Big)^{\frac{1}{2}}\Big(\sum_{K\in\tau_h}(M^-\,\eta,\eta)_{\Gamma_e^K}\Big)^{\frac{1}{2}}\\
&\leq\; 2\,|||\xi|||\,\Big(\sum_{K\in\tau_h}(M\;\eta,\eta)_{\partial K}\Big)^{\frac{1}{2}}\\
&\leq\; 2\,|||\xi|||\;|]\eta[|\,.
\end{aligned}$$

Respecting the bounds of all four terms leads to the result.

□ **Lemma** (3.4)

Then as in the case of the primal problem, a convergence result can be developed.

Theorem 3.5 (Global Convergence for the Dual Problem) *Suppose that $\varphi \in [H^1(\Omega)]^m$, $\varphi \in \mathbf{H}^{\mathbf{l}}(\tau_h)$ for some integers $\mathbf{l} \geq \mathbf{1}$, and $A_j \in [W^{1,\infty}(\tau_h)]^{m\times m}$ $\forall j = 1,..,d$. Then, for any integer $s_K, 1 \leq s_K \leq \min(N_K+1, l_K)$ with $N_K \geq 1$ for all $K \in \tau_h$:*

$$|||\varphi - \varphi_{DG}||| \leq C\Big(\sum_{K\in\tau_h}\frac{h_K^{2s_K-1}}{N_K^{2s_K-1}}|\varphi|_{K,s_K}^2\Big)^{\frac{1}{2}} \tag{3.2.6}$$

where φ is the exact solution of (3.2.2), φ_{DG} the solution of (3.2.3),

$$C = \Big[\tilde{C}(d)h^{\frac{1}{2}}\Big(C_0\frac{1}{N_m^{1/2}} + C_A d N^{\frac{3}{2}}\Big) + 4\,\mu_0^{\frac{1}{2}}C(d,\mathbf{s}) + \Big(\sigma_0\tilde{C}^2(d)\frac{h}{N_m} + \frac{3}{2}\mu_0 C^2(d,\mathbf{s})\Big)^{\frac{1}{2}}\Big]$$

and $N = \max_{K\in\tau_h} N_K$, $N_m = \min_{K\in\tau_h} N_K$ and $h = \max_{K\in\tau_h} h_K$.

Remark 3.6 *If a uniform mesh of mesh size h combined with a constant polynomial order N is used and if the matrices $A_j \in [\mathbb{P}_0(K)]^{m\times m}$ for each element $K \in \tau_h$, then*

$$|||\varphi - \varphi_{DG}||| \le C \frac{h^{s-\frac{1}{2}}}{N^{s-\frac{1}{2}}} |\mathbf{u}|_{\tau_h,\mathbf{s}}$$

where $\mathbf{s}$ is the constant vector $\mathbf{s} = s\mathbf{1}$. For more details, see Remark 2.17.

Proof. Let us denote $\eta = \varphi - \mathbf{P}_\delta\varphi$ and $\xi = \varphi - \mathbf{P}_\delta\varphi$. Then

$$|||\varphi - \varphi_{DG}||| \le |||\varphi - \mathbf{P}_\delta\varphi||| + |||\mathbf{P}_\delta\varphi - \varphi_{DG}||| = |||\eta||| + |||\xi||| \, .$$

By Lemma 2.13 we get a bound for $|||\eta|||$:

$$\begin{aligned} |||\eta||| &\le \Big(\sigma_0 \tilde{C}^2(d) \sum_{K\in\tau_h} \frac{h_K^{2s_K}}{N_K^{2s_K}} |\varphi|^2_{K,s_K} + \frac{3}{2}\mu_0 C^2(d,\mathbf{s}) \sum_{K\in\tau_h} \frac{h_K^{2s_K-1}}{N_K^{2s_K-1}} |\varphi|^2_{K,s_K}\Big)^{\frac{1}{2}} \\ &\le \Big(\sigma_0 \tilde{C}^2(d) \sum_{K\in\tau_h} \frac{h_K}{N_K}\frac{h_K^{2s_K-1}}{N_K^{2s_K-1}} |\varphi|^2_{K,s_K} + \frac{3}{2}\mu_0 C^2(d,\mathbf{s}) \sum_{K\in\tau_h} \frac{h_K^{2s_K-1}}{N_K^{2s_K-1}} |\varphi|^2_{K,s_K}\Big)^{\frac{1}{2}} \, . \end{aligned}$$

Let us denote $N = \max_{K\in\tau_h} N_K$, $N_m = \min_{K\in\tau_h} N_K$ and $h = \max_{K\in\tau_h} h_K$. Then

$$|||\eta||| \le \Big(\sigma_0 \tilde{C}^2(d)\frac{h}{N_m} + \frac{3}{2}\mu_0 C^2(d,\mathbf{s})\Big)^{\frac{1}{2}} \Big(\sum_{K\in\tau_h} \frac{h_K^{2s_K-1}}{N_K^{2s_K-1}} |\varphi|^2_{K,s_K}\Big)^{\frac{1}{2}} \, .$$

For the second term $|||\xi|||$, coercivity (Lemma 2.7) and the Galerkin orthogonality (Lemma 3.2) for the dual problem is used:

$$|||\xi|||^2 \le a(\xi,\xi) + a(\xi,\varphi - \varphi_{DG}) = a(\xi,\eta)$$

since $\xi \in V_\delta$. Then applying Lemma 3.4 leads to

$$|||\xi|||^2 \quad \le \quad a(\xi,\eta) \le \Big[C_0 \|\eta\|_{\tau_h} + C_A d \Big(\sum_{K\in\tau_h} N_K^4 \|\eta\|_K^2\Big)^{\frac{1}{2}} + 4\,|]\eta[|\Big] \, |||\xi|||$$

and consequently

$$|||\xi||| \le C_0 \Big(\sum_{K\in\tau_h} \frac{h_K}{N_K} \Big\|\Big(\frac{N_K}{h_K}\Big)^{\frac{1}{2}}\eta\Big\|_K^2\Big)^{\frac{1}{2}} + C_A d\Big(\sum_{K\in\tau_h} N_K^3 h_K \Big\|\Big(\frac{N_K}{h_K}\Big)^{\frac{1}{2}}\eta\Big\|_K^2\Big)^{\frac{1}{2}} + 4\,|]\eta[| \, .$$

Applying Lemma 2.4 yields

$$\Big\|\Big(\frac{N_K}{h_K}\Big)^{\frac{1}{2}}\eta\Big\|_K^2 \quad \le \quad C_K(d)\Big(\frac{h_K}{N_K}\Big)^{2s_K-1} |\varphi|^2_{K,s_K} \, .$$

Then

$$|||\xi||| \quad \le \quad \tilde{C}(d)\,\Big(C_0 \frac{h^{1/2}}{N_m^{1/2}} + C_A d N^{\frac{3}{2}} h^{\frac{1}{2}}\Big)\,\Big(\sum_{K\in\tau_h}\Big(\frac{h_K}{N_K}\Big)^{2s_K-1} |\varphi|^2_{K,s_K}\Big)^{\frac{1}{2}} + 4\,|]\eta[|$$

Finally using Lemma 2.12 leads to

$$|||\xi||| \le \Big(\tilde{C}(d) h^{1/2}\,\big(C_0 \frac{1}{N_m^{1/2}} + C_A d N^{\frac{3}{2}}\big) + 4\,\mu_0 C^2(d,\mathbf{s})\Big)\,\Big(\sum_{K\in\tau_h}\Big(\frac{h_K}{N_K}\Big)^{2s_K-1} |\varphi|^2_{K,s_K}\Big)^{\frac{1}{2}} \, .$$

Respecting the bounds for $|||\eta|||$ and $|||\xi|||$ leads to the final result.

□ **Theorem** (3.5)

3.3 An A Posteriori Estimation

In this section, we derive an a posteriori error estimation using the dual problem. The aim is to describe the error $|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})|$. This global error is first decomposed in a sum of local errors where each term corresponds to an element $K \in \mathcal{T}_h$ and describes its contribution to the global error. Then a convergence result is proved which describes the convergence behavior of $|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})|$ depending on the local mesh size h_K and the local polynomial order N_K. Finally we give a result which describes the convergence behavior between the error estimated using the exact dual solution and the estimation using the SDG-approximation of the dual solution.

Let φ be the exact solution of the dual problem (3.2.2). Then φ satisfies

$$a_D(\varphi, \mathbf{v}) = a(\mathbf{v}, \varphi) = F_D(\mathbf{v}) \qquad \forall\, \mathbf{v} \in W_\Omega$$

using the same argumentation as for the Galerkin orthogonality (Lemma 3.2). Let $\mathbf{e} = \mathbf{u} - \mathbf{u}_{DG}$ where $\mathbf{u}$ is the exact solution of the primal problem (3.2.1) and $\mathbf{u}_{DG}$ its SDG-approximation. Then by the Galerkin orthogonality for the primal problem and since $\mathbf{P}_\delta\varphi \in V_\delta$

$$\begin{aligned}
F_D(\mathbf{e}) &= a(\mathbf{e}, \varphi) = a(\mathbf{e}, \varphi - \mathbf{P}_\delta\varphi) = F(\varphi - \mathbf{P}_\delta\varphi) - a(\mathbf{u}_{DG}, \varphi - \mathbf{P}_\delta\varphi) \\
&= \Big(\mathbf{f} - B\mathbf{u}_{DG} + \sum_{j=1}^{m} A_j \partial_j \mathbf{u}_{DG}, \varphi - \mathbf{P}_\delta\varphi\Big)_{\tau_h} \\
&\qquad + \sum_{K \in \tau_h} \Big\{ \big(S|\Lambda^-|\bar{\mathbf{g}} - M^-\mathbf{u}_{DG}, \varphi - \mathbf{P}_\delta\varphi\big)_{\Gamma_e^K} - \big(M^-[\mathbf{u}_{DG}], \varphi - \mathbf{P}_\delta\varphi\big)_{\Gamma_i^K} \Big\} \\
&= \big(\mathbf{R}_\delta, \varphi - \mathbf{P}_\delta\varphi\big)_{\tau_h} + \sum_{K \in \tau_h} \Big\{ \big(M^-\mathbf{r}_\delta, \varphi - \mathbf{P}_\delta\varphi\big)_{\Gamma_e^K} - \big(M^-[\mathbf{u}_{DG}], \varphi - \mathbf{P}_\delta\varphi\big)_{\Gamma_i^K} \Big\}
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{R}_\delta &= \mathbf{f} - B\mathbf{u}_{DG} - \textstyle\sum_{j=1}^{d} A_j \partial_j \mathbf{u}_{DG} && \text{in } \Omega \\
\mathbf{r}_\delta &= S\bar{\mathbf{g}} - \mathbf{u}_{DG} && \text{on } \Gamma\ .
\end{aligned}$$

Define the local a posteriori error estimation η_K for each $K \in \tau_h$ by

$$\eta_K = \big(\mathbf{R}_\delta, \varphi - \mathbf{P}_\delta\varphi\big)_K + \big(M^-\mathbf{r}_\delta, \varphi - \mathbf{P}_\delta\varphi\big)_{\Gamma_e^K} - \big(M^-[\mathbf{u}_{DG}], \varphi - \mathbf{P}_\delta\varphi\big)_{\Gamma_i^K}$$

and the global a posteriori error estimation $\eta = \sum_{K \in \tau_h} \eta_K$. Consequently

$$F_D(e) = \sum_{K \in \tau_h} \eta_K\ . \tag{3.3.1}$$

In general, the exact solution of the dual problem φ is not known. One possibility is to approximate the dual solution also by a spectral discontinuous Galerkin Method in the space V_{δ^k} where δ^k represents the couple $(\mathbf{h}, \mathbf{N} + \mathbf{k})$ with $\mathbf{k} = k\mathbf{1}$ and $k \in \mathbb{N}$. Note that the primal solution is sought in V_δ. Let $\varphi_{DG} \in V_{\delta^k}$ be the approximated dual solution. Then

$$\begin{aligned}
F_D(\mathbf{e}) \approx F_D^k(\mathbf{e}) &= \sum_{K \in \tau_h} \Big\{ \big(\mathbf{R}_\delta, \varphi_{DG} - \mathbf{P}_\delta\varphi_{DG}\big)_K + \big(M^-\mathbf{r}_\delta, \varphi_{DG} - \mathbf{P}_\delta\varphi_{DG}\big)_{\Gamma_e^K} \\
&\qquad\qquad - \big(M^-[\mathbf{u}_{DG}], \varphi_{DG} - \mathbf{P}_\delta\varphi_{DG}\big)_{\Gamma_i^K} \Big\} \\
&= \sum_{K \in \tau_h} \eta_K^k\ .
\end{aligned}$$

In the following we use two inequalities which are quoted without proof. First, the following trace inequality is used:

$$\|v\|^2_{L^2(\partial K)} \leq C_0^2 \Big(\|\nabla v\|_{L^2(K)} \, \|v\|_{L^2(K)} + \frac{1}{h_K} \|v\|^2_{L^2(K)} \Big)$$

for $v \in H^1(K)$. Additionally we use the algebraic inverse inequality, see [15]. Let v be a polynomial on K, then there exists a positive constant C, dependent only on d and the shape-regularity of τ_h, such that

$$\|\nabla v\|_{L^2(K)} \leq C \frac{N_K^2}{h_K} \|v\|_{L^2(K)} \, .$$

These two inequalities leads to the following Lemma.

Lemma 3.7 *Let $K \in \tau_h$ be an arbitrary element and $v \in V_\delta^K$. Then*

$$\|\mathbf{v}\|^2_{\partial K} \leq C^2 \frac{N_K^2 + 1}{h_K} \|\mathbf{v}\|^2_K \, .$$

Proof. Using the trace inequality yields

$$\|\mathbf{v}\|^2_{\partial K} \;=\; \sum_{i=1}^m \|v_i\|^2_{L^2(\partial K)} \leq C_0^2 \sum_{i=1}^m \Big(\|\nabla v_i\|_{L^2(K)} \, \|v_i\|_{L^2(K)} + \frac{1}{h_K} \|v_i\|^2_{L^2(K)} \Big) \, .$$

Applying the Inverse inequality

$$\|\nabla v_i\|_{L^2(K)} \leq \tilde{C} \frac{N_K^2}{h_K} \|v_i\|_{L^2(K)}$$

leads to

$$\|\mathbf{v}\|^2_{\partial K} \leq C_0^2 \sum_{i=1}^m \Big[\Big(\tilde{C} \frac{N_K^2}{h_K} + \frac{1}{h_K} \Big) \|v_i\|^2_{L^2(K)} \Big] \leq C^2 \frac{N_K^2 + 1}{h_K} \|\mathbf{v}\|^2_K \, .$$

□ **Lemma** (3.7)

Let us denote

$$\eta(\mathbf{u}, \mathbf{u}_{DG}, \mathbf{v}) = \sum_{K \in \tau_h} \Big\{ \big(\mathbf{R}_\delta, \mathbf{v} - \mathbf{P}_\delta \mathbf{v}\big)_K + \big(M^- \mathbf{r}_\delta, \mathbf{v} - \mathbf{P}_\delta \mathbf{v}\big)_{\Gamma_e^K} - \big(M^- [\mathbf{u}_{DG}], \mathbf{v} - \mathbf{P}_\delta \mathbf{v}\big)_{\Gamma_i^K} \Big\}$$

Lemma 3.8 *Let $\mathbf{u}$ be the exact solution of (3.2.1), $\mathbf{u}_{DG}$ its SDG-approximation and $\mathbf{v}$ an arbitrary function such that $\mathbf{v} \in [L^2(\tau_h)]^m$. If $\mathbf{f} \in V_\delta$, $B\mathbf{v}_\delta + \sum_{j=1}^d A_j \partial_j \mathbf{v}_\delta \in V_\delta$ for all $\mathbf{v}_\delta \in V_\delta$, then*

$$\big|\eta(\mathbf{u}, \mathbf{u}_{DG}, \mathbf{v})\big| \leq C \, |||\mathbf{u} - \mathbf{u}_{DG}||| \, \Big(\sum_{K \in \tau_h} \|\mathbf{v} - \mathbf{P}_\delta \mathbf{v}\|^2_{\partial K} \Big)^{\frac{1}{2}} \, .$$

Proof. Observe that by hypothesis $\mathbf{R}_\delta$ lies in the finite element space V_δ and by the definition of the projector $\mathbf{P}_\delta$:

$$\sum_{K \in \tau_h} \big(\mathbf{R}_\delta, \mathbf{v} - \mathbf{P}_\delta \mathbf{v}\big)_K = 0 \, .$$

Then applying the Cauchy-Schwarz inequality yields

$$
\begin{aligned}
\left|\eta(\mathbf{u},\mathbf{u}_{DG},\mathbf{v})\right| \;\le\; & \Big(\sum_{K\in\tau_h} \big(M^-\mathbf{r}_\delta,\mathbf{r}_\delta\big)_{\Gamma_e^K}\Big)^{\frac{1}{2}} \Big(\sum_{K\in\tau_h} \big(M^-(\mathbf{v}-\mathbf{P}_\delta\mathbf{v}),\mathbf{v}-\mathbf{P}_\delta\mathbf{v}\big)_{\Gamma_e^K}\Big)^{\frac{1}{2}} \\
& +\Big(\sum_{K\in\tau_h} \big(M^-[\mathbf{u}_{DG}],[\mathbf{u}_{DG}]\big)_{\Gamma_i^K}\Big)^{\frac{1}{2}} \Big(\sum_{K\in\tau_h} \big(M^-(\mathbf{v}-\mathbf{P}_\delta\mathbf{v}),\mathbf{v}-\mathbf{P}_\delta\mathbf{v}\big)_{\Gamma_i^K}\Big)^{\frac{1}{2}} \\
\le\; & \Big[\Big(\sum_{K\in\tau_h} \big(M\mathbf{r}_\delta,\mathbf{r}_\delta\big)_{\Gamma_e^K}\Big)^{\frac{1}{2}} + \Big(\sum_{K\in\tau_h} \big(M^-[\mathbf{u}_{DG}],[\mathbf{u}_{DG}]\big)_{\Gamma_i^K}\Big)^{\frac{1}{2}}\Big] \\
& \Big(\sum_{K\in\tau_h} \big(M^-(\mathbf{v}-\mathbf{P}_\delta\mathbf{v}),\mathbf{v}-\mathbf{P}_\delta\mathbf{v}\big)_{\partial K}\Big)^{\frac{1}{2}} \\
\le\; & \|M^-\|_{0,\infty}^{\frac{1}{2}}\Big[\Big(\sum_{K\in\tau_h} \big(M\mathbf{r}_\delta,\mathbf{r}_\delta\big)_{\Gamma_e^K}\Big)^{\frac{1}{2}} + \Big(\sum_{K\in\tau_h} \big(M^-[\mathbf{u}_{DG}],[\mathbf{u}_{DG}]\big)_{\Gamma_i^K}\Big)^{\frac{1}{2}}\Big] \\
& \Big(\sum_{K\in\tau_h} \|\mathbf{v}-\mathbf{P}_\delta\mathbf{v}\|_{\partial K}^2\Big)^{\frac{1}{2}} .
\end{aligned}
$$

By definition of the triple norm $|||\cdot|||$,

$$
\left|\eta(\mathbf{u},\mathbf{u}_{DG},\mathbf{v})\right| \;\le\; C\,|||\mathbf{u}-\mathbf{u}_{DG}|||\,\Big(\sum_{K\in\tau_h} \|\mathbf{v}-\mathbf{P}_\delta\mathbf{v}\|_{\partial K}^2\Big)^{\frac{1}{2}} .
$$

□ **Lemma** (3.8)

In the next theorem, the convergence rate of the quantity $|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})|$ depending on the local mesh size h_K and on the local polynomial order N_K is described.

Theorem 3.9 *Let $\mathbf{u}$ be the exact solution of (3.2.1), $\mathbf{u}_{DG}$ its DG-approximation and φ the exact dual solution of (3.2.2) such that $\varphi \in \mathbf{H}^{\mathbf{k}}(\tau_h)$ for some integers $\mathbf{k} \ge \mathbf{1}$. If $\mathbf{f} \in V_\delta$, $B\mathbf{v}_\delta + \sum_{j=1}^{d} A_j\partial_j\mathbf{v}_\delta \in V_\delta$ for all $\mathbf{v}_\delta \in V_\delta$, $\mathbf{u} \in [H^1(\Omega)]^m$ and $\mathbf{u} \in \mathbf{H}^{\mathbf{l}}(\tau_h)$ for some integers $\mathbf{l} \ge \mathbf{1}$. Then, for any integer $s_K, 1 \le s_K \le \min(N_K+1, l_K)$, $t_K, 1 \le t_K \le \min(N_K+1, k_K)$ with $N_K \ge 1$ for all $K \in \tau_h$:*

$$
\left|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})\right| \le C\,\Big(\sum_{K\in\tau_h}\Big(\frac{h_K}{N_K}\Big)^{2s_K-1}|\mathbf{u}|_{K,s_K}^2\Big)^{\frac{1}{2}}\,\Big(\sum_{K\in\tau_h}\Big(\frac{h_K}{N_K}\Big)^{2t_K-1}|\varphi|_{K,t_K}^2\Big)^{\frac{1}{2}} .
$$

Remark 3.10 *If an uniform mesh of mesh size h and an uniform polynomial order N is used and if $s_K = s$, $t_K = t$, $l_K = l$ and $k_K = k$; then the result becomes*

$$
\left|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})\right| \le C\,\Big(\frac{h}{N}\Big)^{s+t-1}|\mathbf{u}|_{\tau_h,\mathbf{s}}\,|\varphi|_{\tau_h,\mathbf{t}} .
$$

Remark 3.11 *For N-refinement, this result means that it suffices that either the primal or the dual solution is regular in order to obtain an exponential convergence rate. If the primal solution is irregular and the dual solution regular, in contrast to the quantity $|||\mathbf{u} - \mathbf{u}_{DG}|||$ one can still expect an exponential convergence rate for the quantity $\left|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})\right|$.*

Remark 3.12 *If one wishes to develop an fully adaptive algorithm for minimizing $\big|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})\big|$, the regularities of both solutions, the primal and dual, has to be estimated. Then an N-refinement is chosen if at least one of the solutions is regular, otherwise h-refinement is favored.*

Proof. Observe that

$$\big|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})\big| = \Big|\sum_{K\in\tau_h} \eta_K\Big| = \big|\eta(\mathbf{u},\mathbf{u}_{DG},\varphi)\big|$$

and using Lemma 3.8 yields

$$\big|\eta(\mathbf{u},\mathbf{u}_{DG},\varphi)\big| \le C\, |||\mathbf{u}-\mathbf{u}_{DG}|||\, \Big(\sum_{K\in\tau_h} \|\varphi - \mathbf{P}_\delta\varphi\|^2_{\partial K}\Big)^{\frac{1}{2}}\,.$$

Applying Lemma 2.5 leads to

$$\big|\eta(\mathbf{u},\mathbf{u}_{DG},\varphi)\big| \quad\le\quad C\, |||\mathbf{u}-\mathbf{u}_{DG}|||\, \Big(\sum_{K\in\tau_h} \Big(\frac{h_K}{N_K}\Big)^{2t_K-1} |\varphi|^2_{K,t_K}\Big)^{\frac{1}{2}}\,.$$

Finally, Theorem 2.14 under the hypothesis of this theorem becomes

$$|||\mathbf{u}-\mathbf{u}_{DG}||| \le C\Big(\sum_{K\in\tau_h} \Big(\frac{h_K}{N_K}\Big)^{2s_K-1} |\mathbf{u}|^2_{K,s_K}\Big)^{\frac{1}{2}}$$

and we get the final result

$$\big|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})\big| \le C\,\Big(\sum_{K\in\tau_h} \Big(\frac{h_K}{N_K}\Big)^{2s_K-1} |\mathbf{u}|^2_{K,s_K}\Big)^{\frac{1}{2}}\,\Big(\sum_{K\in\tau_h} \Big(\frac{h_K}{N_K}\Big)^{2t_K-1} |\varphi|^2_{K,t_K}\Big)^{\frac{1}{2}}\,.$$

□ **Theorem** (3.9)

Theorem 3.13 *Assume an uniform mesh of mesh size h and an uniform polynomial order N. Let $\mathbf{u}$ be the exact solution of (3.2.1), $\mathbf{u}_{DG}$ its SDG-approximation of uniform polynomial order N, φ the exact dual solution of (3.2.2) such that $\varphi \in [H^1(\Omega)]^m$ and $\varphi \in \mathbf{H}^{\mathbf{r}}(\tau_h)$, for some integer vector $\mathbf{r}$ with $\mathbf{r} = r\mathbf{1}$ and $r \ge 1$. Let φ_{DG} be its SDG-approximation of uniform polynomial order $N+k$. If $\mathbf{f} \in V_\delta$, $B\mathbf{v}_\delta + \sum_{j=1}^d A_j\partial_j\mathbf{v}_\delta \in V_\delta$ for all $\mathbf{v}_\delta \in V_\delta$, $\mathbf{u} \in [H^1(\Omega)]^m$ and $\mathbf{u} \in \mathbf{H}^{\mathbf{l}}(\tau_h)$ for some integer $l \ge 1$ and $\mathbf{l} = l\mathbf{1}$. Then, for any integer $s, 1 \le s \le \min(N+1, l)$, $t, 1 \le t \le \min(N+k+1, r)$ with $N \ge 1$:*

$$\big|F_D(\mathbf{e}) - F_D^k(\mathbf{e})\big| \le C\,(N+k)\Big(\frac{h}{N}\Big)^{s-\frac{1}{2}}\,\Big(\frac{h}{N+k}\Big)^{t-1}\, |\mathbf{u}|_{\tau_h,\mathbf{s}}\, |\varphi|_{\tau_h,\mathbf{t}}$$

where $\mathbf{e} = \mathbf{u} - \mathbf{u}_{DG}$, $\mathbf{s} = s\mathbf{1}$ and $\mathbf{t} = t\mathbf{1}$.

Remark 3.14 *If the mesh size h and the polynomial order N of the SDG-method for the primal problem are fixed, then the theorem estimates the behavior of the error between the real error $F_D(\mathbf{e})$ and the estimated error $F_D^k(\mathbf{e})$ using a SDG-method of mesh size h and polynomial order $N+k$ for the dual problem. Note also that the behavior of the convergence rate of the quantity $|F_D(\mathbf{e}) - F_D^k(\mathbf{e})|$ is independent of the regularity of the primal solution.*

Proof. First, observe that for $\mathbf{v} = \varphi - \varphi_{DG}$:

$$\begin{aligned}
\left|F_D(\mathbf{e}) - F_D^k(\mathbf{e})\right| &= \Big| \sum_{K\in\tau_h} (\eta_K - \eta_K^k)\Big| \\
&= \Big| \sum_{K\in\tau_h} \Big\{ \big(\mathbf{R}_\delta, \mathbf{v} - \mathbf{P}_\delta \mathbf{v}\big)_K + \big(M^- \mathbf{r}_\delta, \mathbf{v} - \mathbf{P}_\delta \mathbf{v}\big)_{\Gamma_e^K} - \big(M^-[\mathbf{u}_{DG}], \mathbf{v} - \mathbf{P}_\delta \mathbf{v}\big)_{\Gamma_i^K} \Big\}\Big| \\
&= \big|\eta(\mathbf{u}, \mathbf{u}_{DG}, \mathbf{v})\big| = \big|\eta(\mathbf{u}, \mathbf{u}_{DG}, \varphi - \varphi_{DG})\big| \,.
\end{aligned}$$

Applying Lemma 3.8 yields

$$\big|\eta(\mathbf{u}, \mathbf{u}_{DG}, \varphi - \varphi_{DG})\big| \leq C\, |||\mathbf{u} - \mathbf{u}_{DG}|||\, \Big(\sum_{K\in\tau_h} \big\|\varphi - \varphi_{DG} - \mathbf{P}_\delta(\varphi - \varphi_{DG})\big\|_{\partial K}^2 \Big)^{\frac{1}{2}} .$$

Let us introduce an intermediary result for estimating

$$\big\|\varphi - \varphi_{DG} - \mathbf{P}_\delta(\varphi - \varphi_{DG})\big\|_{\partial K}^2$$

on an element $K \in \tau_h$.

Lemma 3.15 *Assume a uniform mesh of mesh size h and a uniform polynomial order N. Let φ be the exact dual solution of problem (3.2.2) and φ_{DG} its DG- approximation of polynomial order $N + k$. If $\varphi \in \mathbf{H}^\mathbf{r}(\tau_h)$, for a fixed $K \in \tau_h$ and an integer $r \geq 1$, then for any integer $t, 1 \leq t \leq \min(N + k + 1, r)$ with $N \geq 1$:*

$$\begin{aligned}
&\big\|\varphi - \varphi_{DG} - \mathbf{P}_\delta(\varphi - \varphi_{DG})\big\|_{\partial K}^2 \\
&\leq \; C^2\Big[\big((N+k)^{\frac{1}{2}} + 1\big)^2 \Big(\frac{h}{N+k}\Big)^{2t-1} |\varphi|_{K,t}^2 + \frac{(N+k)^2}{h} \|\varphi - \varphi_{DG}\|_K^2\Big] \,.
\end{aligned}$$

Proof. Using the triangle inequality leads to

$$\big\|\varphi - \varphi_{DG} - \mathbf{P}_\delta(\varphi - \varphi_{DG})\big\|_{\partial K} \leq \|\varphi - \varphi_{DG}\|_{\partial K} + \|\mathbf{P}_\delta(\varphi - \varphi_{DG})\|_{\partial K}$$

and let us denote

$$\begin{aligned}
I_1 &= \|\varphi - \varphi_{DG}\|_{\partial K} \\
I_2 &= \|\mathbf{P}_\delta(\varphi - \varphi_{DG})\|_{\partial K} \,.
\end{aligned}$$

Applying a second time the triangle inequality for the first term I_1 yields

$$I_1 \leq \|\varphi - \mathbf{P}_{\delta^k}\varphi\|_{\partial K} + \|\mathbf{P}_{\delta^k}\varphi - \varphi_{DG}\|_{\partial K} = I_3 + I_4$$

where δ^k represents the couple $(h, N + k)$. Observe that by Lemma 2.5 the term I_3 is bounded by

$$I_3 \leq C\Big(\frac{h}{N+k}\Big)^{t-\frac{1}{2}} |\varphi|_{K,t} \,.$$

For I_4, let us apply Lemma 3.7, use an other time the triangle inequality and apply Lemma 2.4:

$$\begin{aligned}
I_4 &\leq C\frac{N+k}{h^{1/2}} \|\mathbf{P}_{\delta^k}\varphi - \varphi_{DG}\|_K \\
&\leq C\frac{N+k}{h^{1/2}} \Big(\|\varphi - \mathbf{P}_{\delta^k}\varphi\|_K + \|\varphi - \varphi_{DG}\|_K\Big) \\
&\leq C\frac{N+k}{h^{1/2}} \Big[\Big(\frac{h}{N+k}\Big)^t |\varphi|_{K,t} + \|\varphi - \varphi_{DG}\|_K\Big]
\end{aligned}$$

such that considering the bounds for I_3 and I_4 we get:

$$I_1 \leq C\Big[\big((N+k)^{\frac{1}{2}}+1\big)\Big(\frac{h}{N+k}\Big)^{t-\frac{1}{2}}|\varphi|_{K,t} + \frac{N+k}{h^{1/2}}\|\varphi-\varphi_{DG}\|_K\Big] .$$

Finally for the term I_2, Lemma 3.7 is used:

$$I_2 \leq C\frac{N}{h^{1/2}}\,\|\mathbf{P}_\delta(\varphi-\varphi_{DG})\|_K .$$

Observe that

$$\begin{aligned}\|\mathbf{P}_\delta\mathbf{v}\|_K^2 &= \sum_{j=1}^{m}\|(\mathbf{P}_\delta\mathbf{v})_j\|_{L^2(K)}^2 = \sum_{j=1}^{m}\sum_{i=0}^{N_K}c_i\mathbf{v}_{j,i}^2 \leq \sum_{j=1}^{m}\sum_{i=0}^{\infty}c_i\mathbf{v}_{j,i}^2 = \sum_{j=1}^{m}\|\mathbf{v}_j\|_{L^2(K)}^2 \\ &\leq \|\mathbf{v}\|_K^2\end{aligned}$$

where $\mathbf{v}_{j,i}^2$ is the i-th Legendre coefficient of the j-th component of $\mathbf{v}$ and c_i some coefficients. Combining these two arguments leads to:

$$I_2 \;\leq\; C\frac{N}{h^{1/2}}\,\|\mathbf{P}_\delta(\varphi-\varphi_{DG})\|_K \leq C\frac{N}{h^{1/2}}\,\|\varphi-\varphi_{DG}\|_K .$$

Respecting the bounds of I_1 and I_2 yields

$$\begin{aligned}&\big\|\varphi-\varphi_{DG}-\mathbf{P}_\delta(\varphi-\varphi_{DG})\big\|_{\partial K}^2 \\ &\leq\; C^2\Big[\big((N+k)^{\frac{1}{2}}+1\big)\Big(\frac{h}{N+k}\Big)^{t-\frac{1}{2}}|\varphi|_{K,t} + \frac{2N+k}{h^{1/2}}\|\varphi-\varphi_{DG}\|_K\Big]^2 .\end{aligned}$$

Observe that $2N+k \leq 2(N+k)$. Using the following inequality

$$(a+b)^2 \leq 2a^2+2b^2 = C(a^2+b^2)$$

yields

$$\begin{aligned}&\big\|\varphi-\varphi_{DG}-\mathbf{P}_\delta(\varphi-\varphi_{DG})\big\|_{\partial K}^2 \\ &\leq\; C^2\Big[\big((N+k)^{\frac{1}{2}}+1\big)^2\Big(\frac{h}{N+k}\Big)^{2t-1}|\varphi|_{K,t}^2 + \frac{(N+k)^2}{h}\|\varphi-\varphi_{DG}\|_K^2\Big] .\end{aligned}$$

□ **Lemma** (3.15)

Applying the previous Lemma 3.15 and the definition of the triple norm $|||\cdot|||$ leads to

$$\begin{aligned}&\big|\eta(\mathbf{u},\mathbf{u}_{DG},\varphi-\varphi_{DG})\big| \\ &\leq\; C\,|||\mathbf{u}-\mathbf{u}_{DG}|||\cdot \\ &\quad\Big(\sum_{K\in\tau_h}\Big\{\big((N+k)^{\frac{1}{2}}+1\big)^2\Big(\frac{h}{N+k}\Big)^{2t-1}|\varphi|_{K,t}^2 + \frac{(N+k)^2}{h}\|\varphi-\varphi_{DG}\|_K^2\Big\}\Big)^{\frac{1}{2}} \\ &\leq\; C\,|||\mathbf{u}-\mathbf{u}_{DG}|||\cdot \\ &\quad\Big(\big((N+k)^{\frac{1}{2}}+1\big)^2\Big(\frac{h}{N+k}\Big)^{2t-1}\sum_{K\in\tau_h}|\varphi|_{K,t}^2 + \frac{(N+k)^2}{h}\sum_{K\in\tau_h}\|\varphi-\varphi_{DG}\|_K^2\Big)^{\frac{1}{2}} \\ &\leq\; C\,|||\mathbf{u}-\mathbf{u}_{DG}|||\cdot\Big(\big((N+k)^{\frac{1}{2}}+1\big)^2\Big(\frac{h}{N+k}\Big)^{2t-1}|\varphi|_{\tau_h,\mathbf{t}}^2 + \frac{(N+k)^2}{h}|||\varphi-\varphi_{DG}|||^2\Big)^{\frac{1}{2}}\end{aligned}$$

Under the assumptions of this Theorem, Theorem 2.14 becomes

$$|||\mathbf{u}-\mathbf{u}_{DG}||| \leq C\Big(\frac{h}{N}\Big)^{s-\frac{1}{2}}|\mathbf{u}|_{\mathcal{T}_h,\mathbf{s}}$$

and the convergence result for the dual problem, Theorem 3.5 becomes

$$|||\varphi-\varphi_{DG}||| \leq C\Big(\frac{h}{N+k}\Big)^{t-\frac{1}{2}}|\varphi|_{\mathcal{T}_h,\mathbf{t}}\ .$$

Finally we get

$$\begin{aligned}
&\big|\eta(\mathbf{u},\mathbf{u}_{DG},\varphi-\varphi_{DG})\big| \\
&\leq\ C\ \Big(\frac{h}{N}\Big)^{s-\frac{1}{2}}|\mathbf{u}|_{\mathcal{T}_h,\mathbf{s}}\Big(\big((N+k)^{\frac{1}{2}}+1\big)^2\Big(\frac{h}{N+k}\Big)^{2t-1}+\frac{(N+k)^2}{h}\Big(\frac{h}{N+k}\Big)^{2t-1}\Big)^{\frac{1}{2}}\ |\varphi|_{\mathcal{T}_h,\mathbf{t}} \\
&\leq\ C\ \Big(\frac{h}{N}\Big)^{s-\frac{1}{2}}|\mathbf{u}|_{\mathcal{T}_h,\mathbf{s}}\Big((N+k)\Big(\frac{h}{N+k}\Big)^{2t-1}+(N+k)\Big(\frac{h}{N+k}\Big)^{2t-2}\Big)^{\frac{1}{2}}\ |\varphi|_{\mathcal{T}_h,\mathbf{t}} \\
&\leq\ C\,(N+k)\Big(\frac{h}{N}\Big)^{s-\frac{1}{2}}\Big(\frac{h}{N+k}\Big)^{t-1}\ |\mathbf{u}|_{\mathcal{T}_h,\mathbf{s}}\ |\varphi|_{\mathcal{T}_h,\mathbf{t}}
\end{aligned}$$

and consequently the final result.

□ **Theorem** (3.13)

3.4 Numerical Results

We want to confirm numerically the theoretical results of Theorem (3.9) and (3.13). We present three test cases and illustrate the convergence rates in the context of Theorem 3.9 and 3.13.
As domain Ω, the interval $I=(0,1)\subset\mathbb{R}$ is chosen. Let be $x_i = ih$ and $I_i=(x_{i-1},x_i)$ for a uniform mesh size h. The set of all I_i for i from 1 to $N_x=\frac{1}{h}$ build a partition of Ω. The intervals I_i are the elements. Then the spaces V_δ and V_{δ^k} are defined by

$$V_\delta=\{\mathbf{v}\in[L^2(\Omega)]^m \mid \mathbf{v}_{|_{I_i}}\in\mathcal{Q}_N(I_i)\quad\forall i=1..N_x\}$$

and

$$V_{\delta^k}=\{\mathbf{v}\in[L^2(\Omega)]^m \mid \mathbf{v}_{|_{I_i}}\in\mathcal{Q}_{N+k}(I_i)\quad\forall i=1..N_x\}\ .$$

All problems are of the following form:

find $\mathbf{u}: I\to\mathbb{R}^m$ such that

$$\begin{aligned}
B\mathbf{u}+A\partial_x\mathbf{u} &= \mathbf{f} \quad \text{in } \Omega \\
\mathbf{z}^- &= \mathbf{g} \quad \text{on } \Gamma=\partial\Omega\ .
\end{aligned}$$

Then the dual equation is defined by:

find $\varphi: I\to\mathbb{R}^m$ such that

$$\begin{aligned}
B^T\varphi-\partial_x(A\varphi) &= \mathbf{0} \quad \text{in } \Omega \\
\psi^+ &= \mathbf{1} \quad \text{on } \Gamma=\partial\Omega
\end{aligned}$$

where $\mathbf{z}^-$ and ψ^+ are the incoming resp. outgoing parts of the characteristic variables associated to $\mathbf{u}$ resp. φ.
First, a problem where the primal and dual solutions are regular is considered. The second example proves the convergence rates for an irregular primal solution and a regular dual solution. Whereas the primal and dual solutions of the third problem are both irregular.

3.4.1 The code

A Matlab code is developed to solve the linear hyperbolic system using the spectral discontinuous Galerkin method for the primal and dual problems. For computing the matrix A_{LS} and the right hand side $\mathbf{f}_{LS}$ of the linear system, symbolic calculation is used by employing Maple commands in Matlab. To solve the linear system $A_{LS}\mathbf{u} = \mathbf{f}_{LS}$ the GMRes algorithm is used with restarting after 20 inner iterations.

3.4.2 Example 1

First we will study a case where both the primal and dual solutions are regular. Consider the following scalar hyperbolic problem on $\Omega = (0, 1)$:

find $u : (0, 1) \to \mathbb{R}$ such that

$$\begin{array}{rcl} u + \partial_x u & = & x + 1 \quad \text{on } (0,1) \\ u(0) & = & 1 \end{array}$$

with corresponding solution $u(x) = e^{-x} + x$. Then, the associated dual problem reads:

find $\varphi : (0, 1) \to \mathbb{R}$ such that

$$\begin{array}{rcl} \varphi - \partial_x \varphi & = & 0 \quad \text{on } (0,1) \\ \varphi(1) & = & 1 \end{array} .$$

The solution is $\varphi(x) = \frac{e^x}{e}$. Observe that both solutions are contained in $C^\infty(\Omega)$.

Numerical verification of Theorem 3.9
As already mentioned, the primal and dual solutions are both regular. All hypothesis of Theorem 3.9 are satisfied. The error estimation reads

$$\left|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})\right| \le C \left(\frac{h}{N}\right)^{2N+1} |\mathbf{u}|_{\mathcal{T}_h, N+1} |\varphi|_{\mathcal{T}_h, N+1} .$$

For h-refinement we expect an algebraic convergence rate $2N + 1$, whereas for N-refinement an exponential accuracy is estimated. Figure 3.1 shows the h and N-refinement for this example. The numerical results confirm the expected convergence behavior.

Numerical verification of Theorem 3.13
Observe that all conditions of Theorem 3.13 are satisfied. Let us fix the mesh size h and the polynomial order N for the primal problem. The dual solution is sought in V_{δ^k} and let us vary k. Since the dual solution is regular Theorem 3.13 reads

$$\left|F_D(\mathbf{e}) - F_D^k(\mathbf{e})\right| \le C h^{N+k} \left(\frac{1}{N+k}\right)^{N+k-1} |\varphi|_{\mathcal{T}_h, N+k+1}$$

where $\mathbf{e} = \mathbf{u} - \mathbf{u}_{DG}$. Figure 3.2 shows the behavior of $|F_D(\mathbf{e}) - F_D^k(\mathbf{e})|$ while increasing k for fixed $h = \frac{1}{2}$ and $N = 0$. As expected, an exponential convergence can be observed.

3.4.3 Example 2

In this example, we consider a linear hyperbolic system where the primal solution is irregular and the dual solution regular. The primal problem reads:

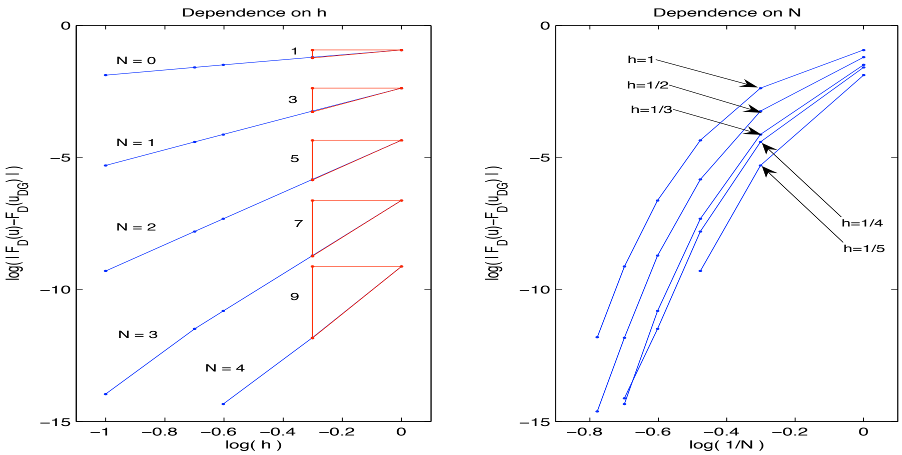

Figure 3.1: *Convergence rates of the quantity* $|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})|$ *depending on the mesh size* h *and the polynomial order* N *for example 1. The slope of the triangles indicates the convergence rates and are chosen to* $2N+1$.

find $\mathbf{u} : (0,1) \to \mathbb{R}$ such that

$$B\mathbf{u} + A\partial_x \mathbf{u} \quad = \quad \mathbf{f} \quad \text{on } (0,1)$$

with Dirichlet boundary conditions on the incoming characteristics according to the exact solution. The coefficient matrices and the right-hand side are defined by

$$B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad , \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and

$$\mathbf{f} = \begin{pmatrix} 2\sinh(x) + \mathbf{1}_{[x>0.5]}\ \alpha\,(x-0.5)^{\alpha-1} + (x-0.5)^{\alpha} \\ 2\cosh(x) + \mathbf{1}_{[x>0.5]}\ \alpha\,(x-0.5)^{\alpha-1} + (x-0.5)^{\alpha} \end{pmatrix}$$

where α is chosen among $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$. Then the solution is

$$\mathbf{u}(x) = \begin{pmatrix} e^{x} + \mathbf{1}_{[x>0.5]}\ (x-0.5)^{\alpha} \\ e^{-x} + \mathbf{1}_{[x>0.5]}\ (x-0.5)^{\alpha} \end{pmatrix}$$

and satisfies $\mathbf{u} \in [H^{\alpha+\frac{1}{2}-\varepsilon}(\tau_h)]^2$ for all $\varepsilon > 0$. The biggest integer $\tilde{\alpha}$ such that $\mathbf{u} \in [H^{\tilde{\alpha}}(\tau_h)]^2$ is given by $\alpha - \frac{1}{2}$.
The dual problem reads

find $\varphi : (0,1) \to \mathbb{R}$ such that

$$\begin{aligned} B\varphi - A\partial_x\varphi \quad &= \quad 0 \quad \text{on } (0,1) \\ \psi^{+} \quad &= \quad \mathbf{1} \quad \text{at } x=0 \text{ and } x=1 \end{aligned}$$

and its solution is

$$\varphi = \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{e^x}{e} - e^{-x} \\ \frac{e^x}{e} + e^{-x} \end{pmatrix} .$$

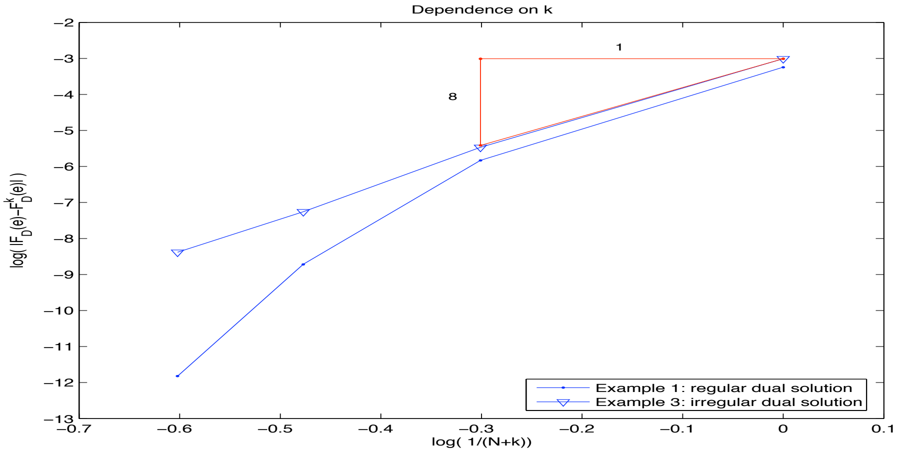

Figure 3.2: *Convergence rates of the quantity* $|F_D(\mathbf{e}) - F_D^k(\mathbf{e})|$ *depending on k for example 1 and 3 where* $\mathbf{e} = \mathbf{u} - \mathbf{u}_{DG}$. *The primal solution is of polynomial order* $N = 0$ *and mesh size* $h = \frac{1}{2}$. *This quantity describes the error between the exact error* $F_D(\mathbf{e})$ *and the a posteriori estimation* $F_D^k(\mathbf{e})$ *if the dual solution is approximated by a SDG-method of order* $N + k$.

Observe that the dual solution satisfies $\varphi \in C^\infty(\Omega)$.

Numerical verification of Theorem 3.9

Observe that the conditions of Theorem 3.9 are not satisfied since $\mathbf{f} \notin V_\delta$. But assume that the result holds nevertheless. The regularities of the primal and dual solution would lead to the following error estimation by Theorem 3.9. Let $s = \min(N+1, \alpha - \frac{1}{2})$, then

$$\left|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})\right| \leq C \left(\frac{h}{N}\right)^{N+s} |\mathbf{u}|_{\tau_h,\mathbf{s}} |\varphi|_{\tau_h,N+1} .$$

The numerical results shows that the convergence rates for h-refinement behave as if $\mathbf{u} \in [H^{\alpha+\frac{1}{2}}(\tau_h)]^2$ and consequently $s = \min(N+1, \alpha + \frac{1}{2})$, see Figure 3.3. Even under this assumption the estimation is still suboptimal for increased polynomial orders N.

For N-refinement, we expect an exponential convergence rate although the primal solution is not enough regular, according to remark 3.11. Numerical results reflect this behavior for small polynomial orders N, see Figure 3.3. For increased polynomial orders an algebraic convergence might be observed which can not be explained by Theorem 3.9. But remember that theorem 3.9 does not hold in this example.

3.4.4 Example 3

We consider a case where both the primal and dual solutions are irregular. Consider the following scalar hyperbolic problem:

find $u : (0,1) \to \mathbb{R}$ such that

$$\begin{aligned} u + \tfrac{x}{\alpha+1}\, \partial_x u &= \left(1 + \tfrac{\alpha}{\alpha+1}\right) x^\alpha \quad \text{on } (0,1) \\ u(0) &= 0 \end{aligned}$$

Its solution reads $u(x) = x^\alpha$ and it satisfies $u \in H^{\alpha+\frac{1}{2}-\varepsilon}$ for all $\varepsilon > 0$ if $\alpha = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$. The

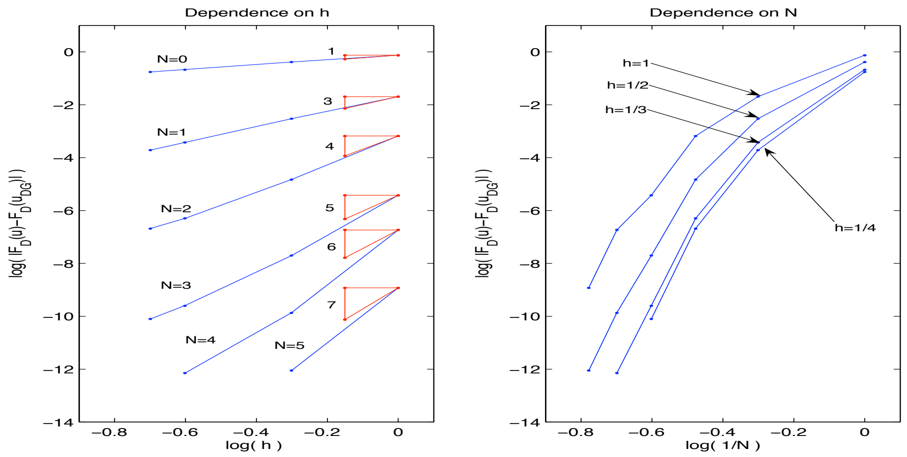

Figure 3.3: *Convergence rates of the quantity* $|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})|$ *depending on the mesh size* h *and the polynomial order* N *for example 2. The slope of the triangles indicates the convergence rates and are chosen to* $N + s$ *where* $s = \min(N+1, \alpha + \frac{1}{2})$.

biggest integer $\tilde{\alpha}$ such that $u \in H^{\tilde{\alpha}}$ is given by $\tilde{\alpha} = \alpha - \frac{1}{2}$. The dual problem becomes:

find $\varphi : (0,1) \to \mathbb{R}$ such that

$$\begin{array}{rcll} \varphi - \partial_x\big(\frac{x}{\alpha+1}\varphi\big) & = & 0 & \text{on } (0,1) \\ \varphi(1) & = & 1 & \end{array}$$

and its solution is given by $\varphi(x) = x^{\alpha}$. As the primal solution, it also satisfies $\varphi \in H^{\alpha+\frac{1}{2}-\varepsilon}$ for all $\varepsilon > 0$ and $\alpha = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$.

Numerical verification of Theorem 3.9

The conditions for Theorem 3.9 are not satisfied since $f \notin V_\delta$ for our choice of $\alpha = \frac{5}{2}$. Assume that the estimation holds also in this case, then due to the irregularity of the primal and dual solution we would expect the following convergence rate. Let $s = \min(N+1, \alpha - \frac{1}{2})$ then

$$\big|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})\big| \leq C \left(\frac{h}{N}\right)^{2s-1} |\mathbf{u}|_{[H^s(\mathcal{T}_h)]^m} |\varphi|_{[H^s(\mathcal{T}_h)]^m} .$$

This means that for h-refinement we can expect a convergence rate of $2s - 1$. In the numerical example the convergence rate of $2s - 1$ can be observed but for $s = \min(N+1, \alpha + \frac{1}{2})$, see Figure 3.4. This means that the convergence behavior is like $\mathbf{u}, \varphi \in H^{\alpha+\frac{1}{2}}$.
For the N-refinement and small polynomial orders N, the estimate predicts an exponential convergence rate. Whereas for increased polynomial orders the accuracy is estimated to be algebraic. This behavior is confirmed by the numerical example, see Figure 3.4.

Numerical verification of Theorem 3.13

As for Theorem 3.9, the conditions for Theorem 3.13 are not satisfied since $f \notin V_\delta$. Let us fix the mesh size h and the polynomial order N for the primal problem while varying k. Let $s = t = \min(N+k+1, \alpha - \frac{1}{2})$, then Theorem 3.13 becomes

$$\big|F_D(\mathbf{e}) - F_D^k(\mathbf{e})\big| \leq C h^{t-1} \left(\frac{1}{N+k}\right)^{t-2} |\mathbf{u}|_{\mathcal{T}_h,\mathbf{s}}\, |\varphi|_{\mathcal{T}_h,\mathbf{t}}$$

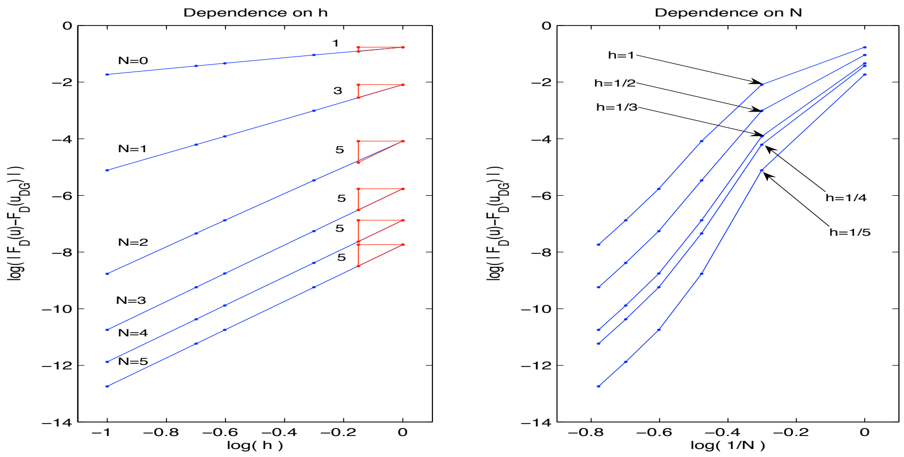

Figure 3.4: *Convergence rates of the quantity* $|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})|$ *depending on the mesh size* h *and the polynomial order* N *for example 3.*

where $\mathbf{e} = \mathbf{u} - \mathbf{u}_{DG}$. Figure 3.2 shows the behavior of $|F_D(\mathbf{e}) - F_D^k(\mathbf{e})|$ while increasing k for fixed $h = \frac{1}{2}$ and $N = 0$. Note that $s = t = \alpha - \frac{1}{2}$ since $N = 0$ and $k > 0$. Then the estimation becomes

$$\left|F_D(\mathbf{e}) - F_D^k(\mathbf{e})\right| \leq C \left(\frac{1}{N+k}\right)^{\alpha - \frac{5}{2}} |\mathbf{u}|_{\mathcal{T}_h, \alpha - \frac{1}{2}} \, |\varphi|_{\mathcal{T}_h, \alpha - \frac{1}{2}}$$

such that an algebraic convergence of $\alpha - \frac{5}{2}$ can be expected. But we chose α to be $\frac{5}{2}$ such that no convergence can be guaranteed by the theorem. However, the numerical example shows an algebraic convergence behavior of rate 8. This numerical result can only be explained by the fact that Theorem 3.13 is not sharp. This is not surprising since Theorem 3.13 is not expected to be sharp. This shows the effect or the sub-optimality of the Cauchy-Schwarz inequality or the inverse inequality.

3.5 Conclusion

We have presented the dual problem associated to an linear hyperbolic system. The spectral discontinuous Galerkin (SDG) method for its approximation was formulated. As for the primal problem, a theorem (Theorem 2.14) describes the convergence rates for $|||\varphi - \varphi_{DG}|||$ depending of the local mesh size h_K and the local polynomial order N_K.

In practise, the quantity of interest is often some linear functional $F_D(\mathbf{u})$ and we would therefor like to control the associated error $|F_D(\mathbf{u} - \mathbf{u}_{DG})|$. $F_D(\cdot)$ is here chosen to be the integral of the outgoing characteristics on the boundary $\partial\Omega$. An a posteriori estimation was developed which estimates the global error $|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})|$ as a sum of local errors associated to the elements. Inspired by the article of Houston and Süli [11], we derived a theorem (Theorem 3.9) which describes the convergence rates of $|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})|$ depending on the local mesh size h_K and local polynomial order N_K. It shows, under the restrictive condition that the residual lies in the finite element space, that an exponential convergence rate can be expected if the primal or the dual solution is regular enough.

In general, the dual solution is not known and it has to be approximated. For this, the same mesh is used with an increased polynomial order $N+k$, supposing a uniform polynomial order N for the primal problem. Using this approximation of the dual solution we get a new a posteriori estimation $F_D^k(\mathbf{e})$ where $\mathbf{e} = \mathbf{u} - \mathbf{u}_{DG}$. In Theorem 3.13, we show the convergence rate of $|F_D(\mathbf{e}) - F_D^k(\mathbf{e})|$ depending on h, N and k, under the condition that the residual lies in the finite element space.

In a further work, it would be interesting to consider the techniques of regularity estimation. This means to estimate the local Sobolev regularity, i.e. determine l_K, k_K such that $\mathbf{u} \in [H^{l_K}(K)]^m$ and $\varphi \in [H^{k_K}(K)]^m$. Using this together with the a posteriori estimate one has all ingredients to construct a fully adaptive algorithm to reduce $|F_D(\mathbf{u}) - F_D(\mathbf{u}_{DG})|$. In a first step, one would select the elements where the associated error has to be reduced using the a posteriori estimation. Then, for each selected element, one would increase the local polynomial order if either the primal or the dual solution is enough regular, otherwise h-refinement would be effected.

Summary

Problem Statement

In this project we address the numerical approximation of hyperbolic equations and systems using the discontinuous Galerkin (DG) method in combination with higher order polynomial degrees. In short, this is called the Spectral Discontinuous Galerkin (SDG) method. Our interest is to review the theoretical properties of the SDG method, particularly for what concerns stability, convergence, dissipation and dispersion. Special emphases is shed on the role of the two parameters, h (the grid-size) and N (the local polynomial degree).

Motivation, Aims and Goals

Standard continuous Galerkin-based finite element methods have poor stability properties when applied to transport-dominated flow problems, so numerical stabilization is needed. In contrast, the spectral discontinuous Galerkin method is known to have good stability properties when applied to first order hyperbolic problems.

The goal of this project is to study in detail the spectral discontinuous Galerkin method applied to hyperbolic systems and equations. Theoretical properties, such as stability, convergence, dissipation and dispersion. Starting from available theoretical results from literature, they should be investigated. The theoretical results should be validated numerically on some model cases.

Framework and State of the act

The spectral discontinuous Galerkin method applied to hyperbolic problems has been proposed in several articles. For second-order partial differential equations with nonnegative characteristic form, the paper of Houston, Schwab and Süli [10] presents a detailed error analysis. This class of equations includes second-order elliptic and parabolic equations, advection-reaction equations, as well as problems of mixed hyperbolic-elliptic-parabolic type.

In the article of Brezzi, Marini and Süli [4], the transport-reaction equation is treated with special focus on different forms of the numerical flux. For the same equation, the reader can find a dispersion and dissipation error analysis in the article of Ainsworth [2], also for a parametrized numerical flux.

For the second order wave equation the spectral discontinuous Galerkin method is presented in the paper of Ainsworth, Monk and Muniz [1]. Dispersive and dissipative properties are studied.

In the article of Monk and Richter [12] a symmetric hyperbolic time-dependent system is considered. A space-time spectral discontinuous Galerkin method is presented followed by a stability and convergence analysis.

Techniques and Methods

The spectral discontinuous Galerkin method is applied to a scalar time-dependent transport equation and symmetric linear hyperbolic system. In the time-dependent case, the time discretisation could be subject to a further project. For both cases, a Matlab code is developed. For the case of the symmetric linear hyperbolic system, the reader will find the code in Appendix A. Exact integration is used by employing Maple commands in Matlab. Thereby we avoid possible errors due to numerical integration.

Innovation and Original Elements

The most original and innovative parts of this work are the convergence analysis, section 1.4.2 and 2.8, and the a posteriori estimation in section 3.3. All three sections are based on two lemmas quoted from the literature [10]. The main idea of section 1.4.2 is inspired by [5], where the same idea is used for a different method. Section 2.8 is fully worked out in this project. The a posteriori estimation of the SDG-method for linear hyperbolic systems and its convergence result (Theorem 3.9) are inspired by the article of Houston and Süli [11]. Theorem 3.13 is completely worked out in the context of this project.

The Matlab codes of all different problems as well as all numerical results are developed in the context of this project.

Acknowledgements

I would like to take this opportunity to acknowledge and thank those who made this work possible and encouraged me during the studies. I thank Prof. Alfio Quarteroni who made me discover the fascinating branch of applied mathematics.

I wish to thank to Prof. Alfio Quarteroni and Erik Burman who gave me the opportunity to realize this master project. Special thanks to Erik for his endless encouragement, support and help.

My gratitude also goes to all members of Prof. Quarteroni's chair, the chair of Modelling and Scientific Computing (CMCS). You were there to help no matter time or day of the week. I enjoyed very much the high performance eatings!

A special thank to my family, my mother and father, Martha and Rudolf, and my brother Simon. They never failed to support me and were always patient and loving even when I decided to study in the far Lausanne. I will be grateful that they never stood in my way, even when they didn't agree with my choices.

To my girlfriend Fabienne, I wish to offer my deepest thanks. Words can not express how grateful I am to be in her life an how much this work was enhanced and made easier by her being in mine.

My friends, they made every day special. Thanks for the nice time I enjoyed outside EPFL and the huge number of coffee breaks we spent together.

Bibliography

[1] M. Ainsworth, P. Monk, and W. Muniz. Dispersive and dissipative properties of discontinuous galerkin finite element methods for the second order wave equation. *Technical report*, 2004.

[2] Mark Ainsworth. Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods. *J. Comput. Phys.*, 198(1):106–130, 2004.

[3] Douglas N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19(4):742–760, 1982.

[4] F. Brezzi, L.D. Marini, and E. Süli. Discontinuous galerkin methods for first-order hyperbolic problems. *Technical report*, 2004.

[5] Erik Burman and Niklas Ericsson. Edge stabilization for the time dependent convection-diffusion-reaction equation. *In preparation*, 2004.

[6] C. Canuto and A. Quarteroni. Approximation results for orthogonal polynomials in Sobolev spaces. *Math. Comp.*, 38(157):67–86, 1982.

[7] Alexandre Ern and Jean-Luc Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[8] Richard S. Falk and Gerard R. Richter. Explicit finite element methods for linear hyperbolic systems. In *Discontinuous Galerkin methods (Newport, RI, 1999)*, volume 11 of *Lect. Notes Comput. Sci. Eng.*, pages 209–219. Springer, Berlin, 2000.

[9] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products*. Fourth edition prepared by Ju. V. Geronimus and M. Ju. Ceĭtlin. Translated from the Russian by Scripta Technica, Inc. Translation edited by Alan Jeffrey. Academic Press, New York, 1965.

[10] Paul Houston, Christoph Schwab, and Endre Süli. Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 39(6):2133–2163 (electronic), 2002.

[11] Paul Houston and Endre Süli. *hp*-adaptive discontinuous Galerkin finite element methods for first-order hyperbolic problems. *SIAM J. Sci. Comput.*, 23(4):1226–1252 (electronic), 2001.

[12] Peter Monk and Gerard R. Richter. A discontinuous galerkin method for linear hyperbolic systems in inhomogeneous media. *Technical report*, 2003.

[13] H. Padé. Sur la répresentation approchée d'une fonction par des fractions rationnelles. *Ann. de l'Éc. Nor.*, 9, 1892.

[14] Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics*, volume 37 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2000.

[15] Alfio Quarteroni and Alberto Valli. *Numerical approximation of partial differential equations*, volume 23 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1994.

[16] Richard S. Varga. On higher order stable implicit methods for solving parabolic partial differential equations. *J. Math. and Phys.*, 40:220–231, 1961.

[17] Qiang Zhang and Chi-Wang Shu. Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin methods for scalar conservation laws. *SIAM J. Numer. Anal.*, 42(2):641–666 (electronic), 2004.

Appendix A

In the context of this work, many Matlab codes have been developed. It is not possible to specify all codes. Here we list the Matlab code of the spectral discontinuous Galerkin method to solve a linear hyperbolic system. For the computation of the matrix and the right hand side of the linear system, symbolic calculation is used by employing Maple commands in Matlab. Since the polynomial basis, the Legendre’s polynomials, are smooth enough. Thereby, we avoid possible errors due to numerical integration.

```
function Norm=DGHyperbolicSystems(handles)
%
% Solve a linear hyperboli system of the form
%
%       A u_x + Bu = f      on I
%
% handles is a handle that should contain:
% a,b:      I=(a,b)
% A,B:      the matrices of the above equation
% m:        the dimension of the hyperbolic system
% f:        the right-hand side of the above equation
% ga,gb:    the boundary conditions at a resp. b
% Nx:       number of subintervals on I
% N:        degree of polynomial approximation space for each subinterval
% points:   Nx+1 node points which define the subintervals
% pw:       is true if f and uexact are given in the form of a "piecewise"
%           Maple statement
% uexact    optional: the exact solution
% sigma     necessary if the exact solution is known: constant such that
%           B+B’-A_x >=sigma I > 0
%
% EXAMPLE:    a=0; b =1; nx=3; n=4; m =2;
%             B = cell(2,2);
%             B{1,1} = ’1’;
%             B{1,2} = ’0’;
%             B{2,1} = ’0’;
%             B{2,2} = ’1’;
%             A = cell(2,2);
%             A{1,1} = ’0’;
%             A{1,2} = ’1’;
%             A{2,1} = ’1’;
%             A{2,2} = ’0’;
%             f = cell(2,1);
%             f{1,1} = ’piecewise(x<.5,2*sinh(x),x>=.5,2*sinh(x)+4.5*(x-.5)^3.5+(x-.5)^4.5)’;
%             f{2,1} = ’piecewise(x<.5,2*cosh(x),x>=.5,2*cosh(x)+4.5*(x-.5)^3.5+(x-.5)^4.5)’;
%             ga = zeros(2,1);
%             ga(1,1) = 1;
%             ga(2,1) = 1;
%             gb = zeros(2,1);
%             gb(1,1) = exp(1)+.5^4.5;
%             gb(2,1) = exp(-1)+.5^4.5;
%             uexact = cell(2,1);
%             uexact{1,1} = [’piecewise(x<.5,exp(x),x>=.5,exp(x)+(x-.5)^4.5)’];
%             uexact{2,1} = [’piecewise(x<.5,exp(-x),x>=.5,exp(-x)+(x-.5)^4.5)’];
%             sigma = 1;
%             pw=1;
%
```

```
%           This program needs the following M-files: dgnorm1d, modstring
%
%           Beni Stamm, EPFL/SB/IACS/CMCS, Jan. 2005

%% Initialization
a = handles.a;
b = handles.b;
A = handles.A;
B = handles.B;
m = handles.m;
f = handles.f;
ga = handles.ga;
gb = handles.gb;
Nx = handles.nx;
N = handles.n;
pw = handles.pw;
uexact = handles.uexact;
points = handles.points;
nhelp = size(N);

% input dimension test
if (nhelp(1)~=m || nhelp(2)~=Nx)
    disp(’false dimensions of N’)
    return;
end
if (points(1)~=a || points(Nx+1)~=b)
    disp(’false values of points at extremeties’)
    return;
end

% define constants for visualization
NumberOfVisualizationPoints = 100;

% define constants
for i=1:Nx
    hx(i) = points(i+1)-points(i);
    PointsPerInterval(i) = round(NumberOfVisualizationPoints*hx(i));
    VisH = 1/NumberOfVisualizationPoints;
end
Nmax=max(max(N));
Nsize = 0;


% get MMinus at a and b
LambdaPlus = zeros(m,m);
LambdaMinus = zeros(m,m);
MMinusAtA = zeros(Nx+1,m,m);
MMinusAtB = zeros(Nx+1,m,m);
SAtA = zeros(m,m);
SAtB = zeros(m,m);
PlusAtA = ones(m,1);
PlusAtB = zeros(m,1);

for k=0:Nx
    x = points(k+1);
    for i=1:m
        for j=1:m
            if(pw)
                AatX(i,j)=str2num(maple(’eval’,A{i,j},[’x=’ num2str(x,15)]));
            else
                AatX(i,j)=eval(A{i,j});
            end
        end
    end
    [VectPr,ValPr]=eigs(AatX);
    for i=1:m
        if (ValPr(i,i) >= 0)
            LambdaPlus(i,i) = ValPr(i,i);
            if k==0
```

```
                PlusAtA(i) = 0;
            end
            if k==Nx
                PlusAtB(i) = 1;
            end
        else
            LambdaMinus(i,i) = ValPr(i,i);
        end
        % computations for Nsize
        if(k>0)
            Nsize = Nsize + N(i,k) + 1;
        end
    end
    APlus = VectPr*LambdaPlus*VectPr’;
    AMinus = VectPr*LambdaMinus*VectPr’;
    % negative part of -A as absolute value
    MMinusAtA(k+1,:,:) = APlus;
    % negative part of A as absulute value
    MMinusAtB(k+1,:,:) = -AMinus;
    if (k==0)
        % positive and negative part of -A as absolute value
        MatA = VectPr*(LambdaPlus - LambdaMinus)*VectPr’;
        MPlusA = -AMinus;
        SAtA = VectPr;
    end
    if (k==Nx)
        % positive and negative part of A as absolute value
        MatB = VectPr*(LambdaPlus - LambdaMinus)*VectPr’;
        MPlusB = APlus;
        SAtB = VectPr;
    end
end
Mhelp1 = zeros(m,m);
Mhelp2 = zeros(m,m);
Mhelp1(:,:) = MMinusAtA(1,:,:);
Mhelp2(:,:) = MMinusAtB(Nx+1,:,:);

% additional stuff
Ga = ga;
Gb = gb;
ga=-2*Mhelp1*ga;
gb=-2*Mhelp2*gb;

% determine if exact solution is known
exact = false;
SizeOfU = size(uexact);
if (SizeOfU  ~= [0 0])
    exact = true;
    Lnorm2 = 0;
    DGnorm2 = 0;
    DGnorm2All = 0;
    Lnorm2All = 0;
    sigma = handles.sigma;
    Lhandles.hx = hx;
    Lhandles.Nx = Nx;
    Lhandles.MMinusAtA = MMinusAtA;
    Lhandles.MMinusAtB = MMinusAtB;
    Lhandles.MatA = MatA;
    Lhandles.MatB = MatB;
    Lhandles.m = m;
    Lhandles.pw = handles.pw;
    Lhandles.a = a;
    Lhandles.uexact = uexact;
    Lhandles.points = points;
end

% load maple orthogonal polynomials
maple(’with’,’orthopoly’)

```

```
% Phi at -1 and 1
PhiAtA = zeros(Nmax+1,1);
PhiAtB = zeros(Nmax+1,1);
for i=0:Nmax
    PhiAtB(i+1)=sqrt((2*i+1)/2);
    PhiAtA(i+1)=sqrt((2*i+1)/2)*(-1)^i;
end

% define matrices
AA = zeros(Nsize,Nsize);
C = zeros(Nsize,Nsize);
D = zeros(Nsize,Nsize);
E = zeros(Nsize,Nsize);
F = zeros(Nsize,1);



% loop over intervals
Ind = 0;
Jnd = 0;
IndStart = 1;
JndOld = 1;
JndMinus = 1;
JndPlus = 0;%sum(N(:,1))+m;
for k=1:Nx
    invTransX = [’2/(’ num2str(hx(k),15) ’)*(x-’ num2str(points(k),15) ’)-1’];
    IntReg = [’x=’ num2str(points(k),15) ’..’ num2str(points(k+1),15)];

    JndMinusOld = JndOld;
    JndOld=Jnd+1;

    % loop over first system index
    for i=1:m
        % loop over first polynomial order
        for j=0:N(i,k)
            Ind = Ind + 1;
            Jnd = JndOld-1;
            JndMinus = JndMinusOld - 1;
            phi_j = [num2str(sqrt((2*j+1)/2),15) ’*(’ maple(’P’,j,invTransX) ’)’];
            % loop over second system index
            for kk=1:m
                % loop over second polynomial order
                for l=0:N(kk,k)
                    Jnd = Jnd + 1;
                    phi_l = [num2str(sqrt((2*l+1)/2),15) ’*(’ maple(’P’,l,invTransX) ’)’];
                    phi_l_prime = maple(’diff’,phi_l,’x’);

                    % (Bu,v)
                    % integration
                    str = [’(’ phi_l ’)*(’ B{i,kk} ’)*(’ phi_j ’)’];
                    maple([’func:=(x)->’ str ]);
                    intstr = maple(’int’,’func(x)’,IntReg);
                    C(Ind,Jnd)=str2num(intstr);

                    % (Au_x,v)
                    str = [’(’ phi_l_prime ’)*(’ A{i,kk} ’)*(’ phi_j ’)’];
                    maple([’func:=(x)->’ str]);
                    intstr = maple(’int’,’func(x)’,IntReg);
                    D(Ind,Jnd)=str2num(intstr);

                    % (M-[u],v)
                    E(Ind,Jnd)=PhiAtA(l+1)*MMinusAtA(k,kk,i)*PhiAtA(j+1)...
                        +PhiAtB(l+1)*MMinusAtB(k+1,kk,i)*PhiAtB(j+1);
                end
                % shifted term
                if (k>1)
                    for l=0:N(kk,k-1)
                        JndMinus = JndMinus+1;
                        E(Ind,JndMinus)=-PhiAtB(l+1)*MMinusAtA(k,kk,i)*PhiAtA(j+1);
                    end
                end
                if k<Nx
                    if i==1 && j==0 && kk==1
                        JndPlus = JndPlus + sum(N(:,k)) + m;
                    end
                    for l=0:N(kk,k+1)
                        JndPlus = JndPlus + 1;
                        E(Ind,JndPlus)=-PhiAtA(l+1)*MMinusAtB(k+1,kk,i)*PhiAtB(j+1);
                    end
                end
            end
            JndPlus = Jnd;

            % (f,v)
            s=maple('P',j,invTransX);
            str = ['(' f{i,1} ')*(' s ')'];
            ff = sqrt((2*j+1)/2)*str2num(maple('int',str,IntReg));
            F(Ind) = ff;
            if (k==1)
                F(Ind)=F(Ind)-.5*ga(i,1)*PhiAtA(j+1);
            end
            if (k==Nx)
                F(Ind)=F(Ind)-.5*gb(i,1)*PhiAtB(j+1);
            end
        end
    end
    % create grid
    for i=0:PointsPerInterval(k)
        X(k,i+1) = points(k) + i*VisH;
    end
end
AA = C+D+E;

% solve linear system
U = gmres(AA,F,20,10e-15,500);

Ind = 0;
IndD = 0;

% get solution in form of string
for k=1:Nx
    invTransX = ['2/(' num2str(hx(k),15) ')*(x-' num2str(points(k),15) ')-1'];
    x0str = maple('eval',points(k));
    x0strPlus = maple('eval',points(k+1));
    xeq = ['x=' x0str '..' x0strPlus];
    % loop over system index
    for i=1:m
        % loop over order of basis
        for j=0:N(i,k)
            Ind = Ind + 1;
            if (j==0)
                Ufunc(k,i) = {['(' num2str(U(Ind),15) '*' num2str(sqrt((2*j+1)/2),15) ...
                            '*(' maple('P',0,invTransX) '))']};
            else
                Ufunc(k,i) = {[char(Ufunc(k,i)) '+(' num2str(U(Ind),15) ...
                            '*' num2str(sqrt((2*j+1)/2),15) '*(' maple('P',j,invTransX) '))']};
            end
        end

        % evaluation on grid for visualization
        ffunc = ModString(char(Ufunc(k,i)));
        x = X(k,1:PointsPerInterval(k)+1);
        z=eval(ffunc);
        if (size(z)==[1 1])
            sx = size(x);
            v(k,i,1:PointsPerInterval(k)+1) = z*ones(sx(1),sx(2));
        else
            v(k,i,1:PointsPerInterval(k)+1) = z;
        end

        if(exact)
            % L2norm
            maple(['func:=(x)->' num2str(sigma,15) '^.5*((' char(uexact(i,1))...
                    ')-(' char(Ufunc(k,i)) '))^2']);
            nstr = maple('int','func(x)',xeq);
            Lnorm2(k,i) = str2num(nstr);
            Lnorm2All = Lnorm2All + Lnorm2(k,i);
        end
    end
    if (exact)
        % DG part of norm
        Lhandles.Ufunc = Ufunc;
        Lhandles.k = k;
        DGnorm2(k) = DGnorm1d(Lhandles);
        DGnorm2All = DGnorm2All + DGnorm2(k);
        if k<Nx
            Lnorm2(k+1,:) = zeros(1,m);
            DGnorm2(k+1) = 0;
        end
    end
end
figure(1)
clf;

%% visualization
vv = zeros(1,PointsPerInterval(k)+1);
for i=1:m
    subplot(m,1,i)
    for k=1:Nx
        vv = zeros(1,PointsPerInterval(k)+1);
        vv(1,1:PointsPerInterval(k)+1)=v(k,i,1:PointsPerInterval(k)+1);
        plot(X(k,1:PointsPerInterval(k)+1),vv)
        hold on;
    end
end

save Ufunctions Ufunc

if (exact)
    Norm(1) = sqrt(Lnorm2All);
    Norm(2) = sqrt(DGnorm2All);
    Norm(3) = sqrt(Lnorm2All+DGnorm2All);
    Norm
    norm=sqrt(DGnorm2+sum(Lnorm2'));
end
```