University of Basel

University of Basel
Department of Mathematics and Computer Science

Discontinuous Galerkin Finite Element Method for Wave Equations with Time-dependent Coefficients

Master Thesis

Mauro Morini

Supervisor: Prof. Dr. Marcus J. Grote
October 17, 2025

Contents

0.1 Introduction 1
0.2 Overview 1

1 DG for Elliptic Problem **2**
1.1 Problem 2
1.2 Discretization 3
1.3 Variational Formulation 3
1.4 Boundary Conditions 4
1.5 Matrix-Vector System 5
1.6 Basis of Finite Element Space 5
1.7 Stiffness Matrix Assembly 7
1.7.1 Assembly of A 8
1.7.2 Assembly of B consistency part 8
1.7.3 Assembly of B penalty part 9
1.7.4 System Vector Assembly 10
1.8 Existence of Discrete Solution 11
1.9 Numerical Results 16
1.9.1 Rate of Convergence 16
1.9.2 Influence of Quadrature Rule on the Convergence Rate 17

A Prerequisites **20**

Abstract

The symmetric interior penalty discontinuous Galerkin finite element method is presented for a second order elliptical problem in 1d which yields a symmetric, positive definite bilinear form for a sufficiently positive penalty parameter. This guarantees existence of and uniqueness of the discrete problem. Implementation and theoretical details are discussed. Numerical results confirm the optimal convergence rate of $\mathcal{O}(h^{r+1})$ in the L^2-norm.

0.1 Introduction

The classical continuous Finite Element Method (FEM) is a widely used and very powerful

0.2 Overview

Chapter 1

DG for Elliptic Problem

First we consider a time-independent elliptic problem. Not only is it useful for initiation to the subject to first consider a simpler elliptic problem, but it is also an essential preparational step in deriving the SIPG bilinear form for the elliptic part of the hyperbolic problem as well.
The goal of this chapter is to build all the necessary theoretical and practical tools to solve a given elliptic problem numerically and then experimentally test the method for different parameters. We will define the necessary notation and derive the SIPG variational formulation as well as in detail describe further implementation steps as for example what basis of the finite element space we chose and how to derive local matrices. Finally this chapter will also include some well established theoretical results in the context of discontinuous Galerkin methods. The derivation of the bilinear form is inspired by Chapter 1 in [6] as well as [2] and [3] for cross reference.

1.1 Problem

We consider the following elliptic model problem:

$$-(c(x)u'(x))' = f(x) \qquad \forall x \in \Omega \tag{1.1}$$

$$u(0) = g_0, u(1) = g_1 \tag{1.2}$$

where $\Omega = (0,1)$ is the domain, $g_0, g_1 \in \mathbb{R}$ are Dirichlet boundary conditions, $f \in L^2(\Omega)$ and $c : \Omega \to \mathbb{R}$ satisfies:

$$c_{\min} \leq c(x) \leq c_{\max} \qquad \forall x \in \Omega$$

for $0 < c_{\min} \leq c_{\max}$. Multiplying the solution by a test function and integrating by parts over Ω we get the standard weak formulation:
Find $u \in H^1(\Omega)$ such that:

$$a(u,v) = (f,v)_{L^2(\Omega)} \qquad \forall v \in C_c^\infty(\Omega) \tag{1.3}$$

Where

$$a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}, \qquad (u,v) \mapsto \int_\Omega c(x)u'(x)v'(x)\mathrm{d}x$$

defines the standard elliptic bilinear form on $H^1(\Omega)$ and

$$(u,v)_{L^2(\Omega)} = \int_\Omega uv \,\mathrm{d}x$$

denotes the L^2-inner product.
If we assume $f \in L^2(\Omega)$ and $c \in C^1(\Omega)$ we know by *Lax-Milgram* that there exists a unique (weak) solution of (1.3) and from the *elliptic regularity theory* we can conclude higher regularity of the weak solution, namely $u \in H^2(\Omega)$. For reference see chapter 6.2, 6.3 in [1].

1.2 Discretization

Let $0 = x_0 < \cdots < x_{N+1} = 1$ be the mesh faces, $I_n = (x_n, x_{n+1})$ for $n = 0, \dots, N$ be the elements and $\mathcal{T}_h = \{I_n\}_{n=0}^N$ a partition of Ω for some fixed $N \in \mathbb{N}$. We denote the element length by $h_n = x_{n+1} - x_n$ for $n = 0, \dots, N$ and the global meshsize by $h = \max_n h_n$. Next we define the discontinuous finite element space

$$V_h^r(\mathcal{T}_h) = \{v \in L^2(\Omega) | \, v|_{I_n} \in \mathcal{P}^r(I_n)\} \tag{1.4}$$

where $\mathcal{P}^r(I_n)$ denotes the space of polynomials $p : I_n \to \mathbb{R}$ of degree r for $r \in \mathbb{N}$. When the context allows it, we will denote the finite element space with just V_h for simplicity. V_h is our final approximation space in which the numerical solution lays. We observe that in contrast to a continuous finite element approximation space here the resulting solution is a priori discontinuous by construction. Furthermore we have here $V_h \not\subset H^1(\Omega)$. This is especially apparent in 1d due to the Sobolev embedding $H^1(\Omega) \subset C^0(\Omega)$. Any discontinuous element of V_h can therefore not be in $H^1(\Omega)$.
To proceed we will require the following trace operators:

Definition 1.1. *Let $v : \Omega \to \mathbb{R}$ be piecewise continuous and let $n \in \{1, \dots, N\}$*

(i) *We denote $v(x_n^+) := \lim_{x \searrow x_n} v(x), v(x_n^-) := \lim_{x \nearrow x_n} v(x)$ the limit from above/below.*

(ii) *We define the **jump** at x_n as*

$$[v(x_n)] := v(x_n^-) - v(x_n^+)$$

*and the **average** at x_n as*

$$\{v(x_n)\} := \frac{v(x_n^+) + v(x_n^-)}{2}$$

furthermore by convention we set:

$$[v(x_0)] := -v(x_0^+), \quad [v(x_{N+1})] := v(x_{N+1}^-), \quad \{v(x_0)\} := v(x_0^+), \quad \{v(x_{N+1})\} := v(x_{N+1}^-)$$

1.3 Variational Formulation

To derive the SIPG variational formulation, let $v \in V_h$ be a test function. As mentioned in section 1.1 we can assume that the coefficient $c \in C^1(\Omega)$ and the exact solution $u \in H^2(\Omega) \subset C^1(\Omega)$. Due to the discontinuity of the test function in contrast to continuous FEM we multiply u with v on each element I_n and integrate by parts locally

$$\int_{x_n}^{x_{n+1}} fv \, \mathrm{d}x = -\int_{x_n}^{x_{n+1}} (cu')'v \, \mathrm{d}x = \int_{x_n}^{x_{n+1}} cu'v' \, \mathrm{d}x - cu'v\Big|_{x_n}^{x_{n+1}} \qquad \forall n = 0, \dots, N$$

then sum over all elements

$$(f, v)_{L^2(\Omega)} = \sum_{n=0}^{N} \int_{I_n} cu'v' \, \mathrm{d}x - \sum_{n=0}^{N+1} [c(x_n)u'(x_n)v(x_n)] \tag{1.5}$$

where we have used that $\sum_{n=0}^N w\Big|_{x_n}^{x_{n+1}} = w(x_{N+1}^-) - w(x_N^+) + w(x_N^-) - \cdots - w(x_1^+) + w(x_1^-) - w(x_0^+) = \sum_{n=0}^{N+1} [w(x_n)]$ for any piece-wise continuous function w.
By our construction are c, u' continuous on Ω, this means

$$[c(x_n)u'(x_n)v(x_n)] = c(x_n)u'(x_n)[v(x_n)] = \{c(x_n)u'(x_n)\}[v(x_n)] \qquad \forall n = 0, \dots, N+1 \tag{1.6}$$

and

$$[u(x_n)] = 0 \qquad \forall n = 1, \dots, N \tag{1.7}$$

To derive the final variational form we will now have to add two additional terms to (1.5):
Step 1. Firstly we need to symmetrize our currently non-symmetrical right hand side which will correspond to the SIPG bilinear form. To do so we subtract $\sum_{n=0}^{N+1}\{c(x_n)v'(x_n)\}[u(x_n)]$ on both sides of (1.5) so we get

$$\begin{aligned}&(f,v)_{L^2(\Omega)} - g_1 c(x_{N+1}^-)v(x_{N+1}^-) + g_0 c(x_0^+)v(x_0^+)\\ =&\sum_{n=0}^{N}\int_{I_n} cu'v'\,\mathrm{d}x - \sum_{n=0}^{N+1}\{c(x_n)u'(x_n)\}[v(x_n)] + \{c(x_n)v'(x_n)\}[u(x_n)]\end{aligned}$$

where on the left hand side of the equation we have applied (1.7) for the interior node contributions of the sum (which therefore vanish), and the boundary condition (1.25) ensuring the left hand side to be soley dependent on v.
Step 2. The bilinear form we seek to create will (for now) be defined on $V_h \times V_h$ meaning it will intake discontinuous functions. In particular the numerical solution will be a discontinuous function wheras the exact solution is continuous. To counterweigh this discrepancy we need to integrate a penalization mechanism, seeking to minimize discontinuous behaviors. Technically speaking this penalization term will guarantee coercivity of the bilinear form (see section 1.8).
Let $\sigma > 0$ constant, we define:

$$\mathtt{c}_n := \begin{cases}\max(c(x_n^+), c(x_n^-)), & n = 1,\dots,N\\ c(x_n^+), & n = 0\\ c(x_n^-), & n = N+1\end{cases}, \qquad \mathtt{h}_n := \begin{cases}\min(h_n, h_{n-1}), & n = 1,\dots,N\\ h_n, & n \in \{0, N+1\}\end{cases}$$

with this we define our penalization parameter

$$\mathtt{a}_n := \frac{\sigma\,\mathtt{c}_n}{\mathtt{h}_n} > 0 \qquad \forall n = 0\dots, N+1 \tag{1.8}$$

Similarly to Step 1 we can now add the term $\sum_{n=0}^{N+1}\mathtt{a}_n[u(x_n)][v(x_n)]$ on both sides of (1.5) and get the final *discrete* SIPG variational formulation.

Find $u_h \in V_h$ such that:

$$b_h(u_h, v) = \ell(v), \qquad \forall v \in V_h \tag{1.9}$$

where

$$\begin{aligned}b_h(u,v) &= \sum_{n=0}^{N}\int_{I_n} cu'v'\,\mathrm{d}x - \sum_{n=0}^{N+1}\{c(x_n)u'(x_n)\}[v(x_n)] + \{c(x_n)v'(x_n)\}[u(x_n)] + \sum_{n=0}^{N+1}\mathtt{a}_n[u(x_n)][v(x_n)]\\ \ell(v) &= (f,v)_{L^2(\Omega)} - g_1 c(x_{N+1}^-)v(x_{N+1}^-) + g_0 c(x_0^+)v(x_0^+) + \mathtt{a}_{N+1}g_1 v(x_{N+1}^-) + \mathtt{a}_0 g_0 v(x_0^+)\end{aligned}$$

for $u, v \in V_h$.

1.4 Boundary Conditions

By adding the terms $-\sum_{n=0}^{N+1}\{c(x_n)v'(x_n)\}[u(x_n)], \sum_{n=0}^{N+1}\mathtt{a}_n[u(x_n)][v(x_n)]$ on both sides of (1.5) we *weakly* imposed the Dirichlet boundary conditions into the variational form. This stands in contrast to how boundary conditions are usually imposed in continuous FEM. Indeed one could also impose them strongly, meaning we could define

$$V_h^r(\mathcal{T}_h) = \{v \in L^2(\Omega) |\, v|_{I_n} \in \mathcal{P}^r(I_n), v(x_0) = g_0, v(x_{N+1}) = g_1\}$$

but this soley as a side note, we will continue to work with purely weakly imposed boundary conditions.

One could alternatively desire to implement *Neumann* boundary conditions, this slightly changes the variational formulation. We illustrate the idea on the following example boundary condition. A solution u should satisfy:

$$u(0) = g_0, u'(1) \cdot n_1 = g_1$$

where again $g_0, g_1 \in \mathbb{R}$ are the boundary values and n_1 denotes the outward normal of the domain at the upper boundary. In 1d we trivially have $n_1 = 1, n_0 = -1$, where n_0 denotes the outward normal at the lower boundary.

Now recall the initial incomplete formulation (1.5). First we take the Neumann boundary contribution $\{c(x_{N+1})u'(x_{N+1})\}[v(x_{N+1})]$ to the other side of the equation. We get

$$\sum_{n=0}^{N} \int_{I_n} cu'v' \,\mathrm{d}x - \sum_{n=0}^{N} \{c(x_n)u'(x_n)\}[v(x_n)] = (f,v)_{L^2(\Omega)} + g_1 c(x_{N+1}^-)v(x_{N+1}^-)$$

We have used $\{c(x_{N+1})u'(x_{N+1})\}[v(x_{N+1})] = c(x_{N+1}^-)u'(x_{N+1}^-)v(x_{N+1}^-) \cdot n_1 = g_1 c(x_{N+1}^-)v(x_{N+1}^-)$

From here on we proceed similarly as in the Dirichlet case. The main difference is that we always ommit the boundary face with the Neumann boundary condition.
We add the terms

$$-\sum_{n=0}^{N} \{c(x_n)v'(x_n)\}[u(x_n)] + \sum_{n=0}^{N} \mathsf{a}_n [u(x_n)][v(x_n)]$$

to both sides, using again that the real solution has zero jump on the interior faces and applying the boundary conditions we finally derive the variational form

$$\begin{aligned}
&\sum_{n=0}^{N} \int_{I_n} cu'v' \,\mathrm{d}x - \sum_{n=0}^{N} \{c(x_n)u'(x_n)\}[v(x_n)] + \{c(x_n)v'(x_n)\}[u(x_n)] + \sum_{n=0}^{N} \mathsf{a}_n [u(x_n)][v(x_n)] \\
&= (f,v)_{L^2(\Omega)} + g_0 c(x_0^+)v(x_0^+) + \mathsf{a}_0 g_0 v(x_0^+) + g_1 c(x_{N+1}^-)v(x_{N+1}^-)
\end{aligned}$$

1.5 Matrix-Vector System

We will now in derive the fully discrete Matrix-Vector system given by the variational form (1.9). To do so let $r \in \mathbb{N}$ denote the polynomial degree and consequently the element degree of freedom. Note that in this thesis we will only consider global polynomial degrees, meaning one set polynomial degree for all elements. Next let $\{\Phi_0, \ldots, \Phi_M\}$ be a basis of V_h, where $M = \dim(V_h)$. We can represent the sought Galerkin approximation as $u_h = \sum_{j=0}^{M} \alpha_j \Phi_j \in V_h$ for coefficients $\alpha_j \in \mathbb{R}$. Then (1.9) is equivalent to:

$$\sum_{j=0}^{M} \alpha_j b_h(\Phi_j, \Phi_i) = \ell(\Phi_i) \qquad \forall i = 0, \ldots, M$$

which corresponds to the system:

$$\mathbf{B}\mathbf{u} = \mathbf{l} \tag{1.10}$$

for $\mathbf{B} \in \mathbb{R}^{M \times M}, [\mathbf{B}]_{i,j} = b_h(\Phi_j, \Phi_i), \mathbf{u} \in \mathbb{R}^M, [\mathbf{u}]_j = \alpha_j, \mathbf{l} \in \mathbb{R}^M, [\mathbf{l}]_j = \ell(\Phi_j)$.

1.6 Basis of Finite Element Space

There are many ways of choosing basis functions for finite element spaces. In this thesis we soley focus on elementwise nodal Lagrangian basis functions, although there are alternatives which are just as valid, like the modal Legendre basis. When choosing basis nodes for a lagrangian basis we have a couple of choices to make.

1. *Should the quadrature nodes coincide with the basis nodes?*
Having the quadrature nodes coincide with the basis nodes simplifies the calculations, because by the property

of a Lagrangian basis the values of the basis functions at the quadrature nodes, i.e the basis nodes coincide with the Kroneker delta. So a value matrix composed of each value at the quadrature nodes for each basis function (per element) is just the identity matrix. Specifically we have:
Let $\xi_0, \dots, \xi_r$ be the basis- and quadrature nodes and let $\widehat{\phi}_0, \dots, \widehat{\phi}_r$ be the Lagrangian basis, then

$$\widehat{\phi}_i(\xi_j) = \delta_{i,j} := \begin{cases} 1 & , \text{ for } i = j \\ 0 & , \text{ for } i \neq j \end{cases}$$

therefore $\mathbf{I} = \mathbf{L}$ where $\mathbf{I}$ is the identity and $[\mathbf{L}]_{i,j} = [\widehat{\phi}_i(\xi_j)]$ the basis shape-function value-matrix. This reduces the computational cost, since there is no need to actually evaluate/interpolate the polynomials at the quadrature nodes. The clear downside is that we couple the exactness of the quadrature to the polynomial degree of the basis function. If for example we fix a polynomial degree $r \in \mathbb{N}$. The Gauss-Lobatto quadrature with $r+1$ nodes is exact for polynomials of degree $2r-1$. When computing the mass matrix we have to calculate an integral of the form $\int_I \Phi_j \Phi_i \mathrm{d}x$ where $I \in \mathcal{T}_h$ is a element and $\Phi_i|_I, \Phi_j|_I \in \mathcal{P}^r(I)$ are basis functions. This means $\Phi_j \cdot \Phi_i \in \mathcal{P}^{2r}$ and therefore we introduce an error.

2. *Should there be basis nodes on the element boundaries?*
In contrast to continuous FEM we do not need to impose continuity over the element boundaries, so we have the option of allowing the nodes to be exclusively in the interior of the elements. This allows us to use Gauss-Legendre quadrature nodes in 1d for example, which is exact for polynomials of degree $2r+1$ for $r+1$ quadrature nodes. This would remedy the issue of approximating the mass matrix inexactly as described before when choosing the same quadrature nodes as basis nodes. The downside here is that we sill require boundary values to incorporate boundary conditions and calculating the penalty, consistency and symmetry terms (see matrices $\mathbf{B}_{\text{cons}}, \mathbf{B}_{\text{penal}}$ in section 1.7), so the boundary values have to be interpolated.

In this thesis we will not try to justify the choice of basis function too much and use Gauss-Lobatto quadrature nodes as basis nodes. This is a commonly used nodal basis in continuous 1d-FEM and we will also use it for our DG method. In Appendix A.2.3 of [4] the *Fekete points*, which in 1d coincide with the Gauss-Lobatto points, are recommended with the argument that this nodal basis yields a mass matrix with an optimal condition number. For more detailed information on choosing basis functions and alternative approaches see for example Appendix A.2 in [4].

Let $n \in \{1, \dots, N\}$ and $I_n \in \mathcal{T}_h$ be an arbitrary element. We denote $\hat{I} = (-1, 1)$ the *reference element* and $F_n : \hat{I} \to I_n, \xi \mapsto \frac{x_n + x_{n+1}}{2} + \frac{h_n}{2}\xi$ the *element map*. This now allows us to define a basis on the reference element and extend it to all elements using the element map.
For a fixed polynomial degree $r \geq 2$ let $\xi_0, \dots, \xi_r \in [-1, 1]$ be the Gauss-Lobatto nodes.

$r = 2$	$\{-1, 1\}$
$r = 3$	$\{-1, 0, 1\}$
$r = 4$	$\{-1, -\frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}, 1\}$
$\vdots$	$\vdots$

The inner nodes are given by the roots of L'_{r-1}, the derivative of the $r-1$-th Legendre polynomial. We define the basis on the reference element as the Lagrangian nodal basis

$$\widehat{\phi}_i(\xi) := \prod_{\substack{j=0 \\ j \neq i}}^{r} \frac{\xi - \xi_j}{\xi_i - \xi_j}, \qquad \text{for } i = 0, \dots, r \tag{1.11}$$

and define the basis functions on the element I_n as

$$\phi_i^n : I_n \to \mathbb{R}, \quad \phi_i^n(x) := \widehat{\phi}_i(F_n^{-1}(x))$$

as a last step we extend the basis functions to the whole domain Ω by zero

$$\Phi_i^n : \Omega \to \mathbb{R}, \quad \Phi_i^n(x) := \begin{cases} \phi_i^n(x), & \text{for } x \in I_n \\ 0, & \text{else} \end{cases} \tag{1.12}$$

for $n = 0, \dots, N$ and $i = 0, \dots, r$. Clearly we have $\operatorname{span}(\widehat{\phi}_0, \dots, \widehat{\phi}_r) = \mathcal{P}^r(\hat{I})$ and by extension $\operatorname{span}(\Phi_0^0, \dots, \Phi_r^N) = V_h^r(\mathcal{T}_h)$. It is essential that our basis has only local support, meaning the basis functions are zero on most of the domain. This is the key property which allows the final matrices to be sparse. Choosing basis functions with global support, the computational cost would be unfeasible for small mesh sizes.
By having chosen a Lagrangian nodal basis the mesh nodes exactly coincide with the Gauss-Lobatto nodes on each element. To simplify the notation we introduce a *local-to-global* index map

$$T : \{0, \dots, N\} \times \{0, \dots, r\} \to \{1, \dots, M\} \tag{1.13}$$

where $M = (r+1)(N+1) = \dim(V_h)$. T takes an element index n and a local basis function index i as inputs and returns the globally assigned node index $T(n, i)$. T corresponds to the *connectivity matrix*. In the simplest case we have the global index ordered from left to right and get $T(n, i) = nr + i$.

1.7 Stiffness Matrix Assembly

With the basis functions in (1.12) defined we can now in detail investigate how to assemble the matrix $\mathbf{B}$ in (1.10). To do so we firstly separate the bilinear form b_h into different components

$$\begin{aligned} a_h(u, v) &:= \sum_{n=0}^{N} \int_{I_n} c u' v' \, \mathrm{d}x \\ b_h^{\text{cons}}(u, v) &:= \sum_{n=0}^{N+1} \{c(x_n) u'(x_n)\}[v(x_n)] + \{c(x_n) v'(x_n)\}[u(x_n)] \\ b_h^{\text{penal}}(u, v) &:= \sum_{n=0}^{N+1} \mathtt{a}_n [u(x_n)][v(x_n)] \end{aligned}$$

Let $u_h = \sum_{m=0}^{N} \sum_{j=0}^{r} \alpha_j^m \Phi_j^m \in V_h$ denote the Galerkin approximation, then as discussed in section 1.6 the discrete variational formulation (1.9) is equivalent to

$$\sum_{m=0}^{N} \sum_{j=0}^{r} \alpha_j^m \Big(a_h(\Phi_j^m, \Phi_i^n) - b_h^{\text{cons}}(\Phi_j^m, \Phi_i^n) + b_h^{\text{penal}}(\Phi_j^m, \Phi_i^n) \Big) = \ell(\Phi_i^n), \qquad \forall n = 0, \dots, N, i = 0, \dots, r \tag{1.14}$$

which corresponds to the matrix vector system (1.10) where we can write

$$\mathbf{B} = \mathbf{A} - \mathbf{B}_{\text{cons}} + \mathbf{B}_{\text{penal}}$$

we will assemble the three (symmetric) matrices separately.

$$\begin{aligned} [\mathbf{B}]_{T(n,i),T(m,j)} &= b_h(\Phi_j^m, \Phi_i^n), & [\mathbf{A}]_{T(n,i),T(m,j)} &= a_h(\Phi_j^m, \Phi_i^n) \\ [\mathbf{B}_{\text{cons}}]_{T(n,i),T(m,j)} &= b_h^{\text{cons}}(\Phi_j^m, \Phi_i^n) & [\mathbf{B}_{\text{penal}}]_{T(n,i),T(m,j)} &= b_h^{\text{penal}}(\Phi_j^m, \Phi_i^n) \end{aligned}$$

where T is given by (1.13)

1.7.1 Assembly of A

$\mathbf{A}$ is assembled similarly to the standard stiffness matrix in continuous finite element. The main difference is that there is no overlap in the elementwise contributions. Each set of local (element) basis functions only contributes to the integrals over said element. We can rewrite $\mathbf{A} = \sum_{s=0}^{N} \mathbf{A}^{(s)}$, where $[\mathbf{A}^{(s)}]_{T(n,i),T(m,j)} = \int_{I_s} c\,(\Phi_j^m)'(\Phi_i^n)'\,\mathrm{d}x$. Now since we have $\operatorname{supp}(\Phi_i^n) \subset I_n$ the only non-zero entries of $\mathbf{A}^{(s)}$ are the ones where both $n = m = s$. Pulling back the integral to the reference element using the chain rule and the substitution $F_s^{-1}(x) = \xi$ we find

$$\int_{I_s} c(x)\,(\Phi_j^s)'(x)(\Phi_i^s)'(x)\,\mathrm{d}x = \frac{2}{h_s}\int_{\hat{I}} c(F_s(\xi))\,\widehat{\phi}_j'(\xi)\widehat{\phi}_i'(\xi)\,\mathrm{d}\xi$$

This integral now only depends on the reference shape functions, the element length h_s and the values of the coefficient c. Using the Gauss-Lobatto quadrature rule we can approximate the integral only requiring the values of c at the nodes, whilst having the values of ϕ, ϕ' preloaded. The total assembly of $\mathbf{A}$ can therefore be achieved by calculating a local contribution matrix $\widehat{\mathbf{A}}^{(s)} \in \mathbb{R}^{(r+1)\times(r+1)}$ for each element I_s and adding it into $\mathbf{A}$.

Example 1.2. *For $c \equiv 1$ with $\mathcal{P}^1$-elements ($r = 1$) we have*

$$\widehat{\boldsymbol{A}}^{(s)} = \frac{1}{h_s}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

1.7.2 Assembly of B consistency part

As before we rewrite $\mathbf{B}_{\text{cons}} = \sum_{s=0}^{N+1} \mathbf{B}_{\text{cons}}^{(s)}$, where

$$[\mathbf{B}_{\text{cons}}^{(s)}]_{T(n,i),T(m,j)} = \{c(x_s)\Phi_j^{m\,\prime}(x_s)\}[\Phi_i^n(x_s)] + \{c(x_s)\Phi_i^{n\,\prime}(x_s)\}[\Phi_j^m(x_s)]$$

Interior Faces

First let $s \in \{1, \dots, N\}$ denote an interior face, we observe again, that the entries of $\mathbf{B}_{\text{cons}}^{(s)}$ can only be non-zero for an index tuple $(T(n,i), T(m,j))$ if $n, m \in \{s, s-1\}$ by the local support of the basis functions. Therefore assembling $\mathbf{B}_{\text{cons}}$ again comes down to calculating a local contribution matrix

$$\widehat{\mathbf{B}}_{\text{cons}}^{(s)} = \begin{bmatrix} \mathbf{C}_{\text{cons}}^{(s-1,s-1)} & \mathbf{C}_{\text{cons}}^{(s-1,s)} \\ \mathbf{C}_{\text{cons}}^{(s,s-1)} & \mathbf{C}_{\text{cons}}^{(s,s)} \end{bmatrix} \in \mathbb{R}^{2(r+1)\times 2(r+1)}$$

for each interior face x_s consisting of four blocks we will now lay out in more detail. We discuss the boundary case separately. Using again the local support of the basis functions we find

$$\begin{aligned}
[\mathbf{C}_{\text{cons}}^{(s-1,s-1)}]_{i,j} &= \frac{c(x_s^-)}{2}\Phi_j^{s-1\,\prime}(x_s^-)\Phi_i^{s-1}(x_s^-) + \frac{c(x_s^-)}{2}\Phi_i^{s-1\,\prime}(x_s^-)\Phi_j^{s-1}(x_s^-) \\
[\mathbf{C}_{\text{cons}}^{(s-1,s)}]_{i,j} &= \frac{c(x_s^+)}{2}\Phi_j^{s\,\prime}(x_s^+)\Phi_i^{s-1}(x_s^-) - \frac{c(x_s^-)}{2}\Phi_i^{s-1\,\prime}(x_s^-)\Phi_j^{s}(x_s^+) \\
[\mathbf{C}_{\text{cons}}^{(s,s)}]_{i,j} &= -\frac{c(x_s^+)}{2}\Phi_j^{s\,\prime}(x_s^+)\Phi_i^{s}(x_s^+) - \frac{c(x_s^+)}{2}\Phi_i^{s\,\prime}(x_s^+)\Phi_j^{s}(x_s^+)
\end{aligned}$$

where we have used the definitions of jump and average in (1.1). Note that $\mathbf{C}_{\text{cons}}^{(s-1,s)} = (\mathbf{C}_{\text{cons}}^{(s,s-1)})^T$ by the symmetry of the bilinear form b_h^{cons}. Next we represent the values of the basis functions Φ at the element boundary by the values of the reference shape functions $\widehat{\phi}$

$$\begin{aligned}
\Phi_i^s(x_s^+) &= \widehat{\phi}_i(-1), & \Phi_i^{s-1}(x_s^-) &= \widehat{\phi}_i(1) \\
\Phi_i^{s\,\prime}(x_s^+) &= \frac{2}{h_s}\widehat{\phi}_i'(-1), & \Phi_i^{s-1\,\prime}(x_s^-) &= \frac{2}{h_{s-1}}\widehat{\phi}_i'(1)
\end{aligned} \tag{1.15}$$

which finally yields

$$[\mathbf{C}_{\text{cons}}^{(s-1,s-1)}]_{i,j} = \frac{c(x_s^-)}{h_{s-1}}\widehat{\phi}_j'(1)\widehat{\phi}_i(1) + \frac{c(x_s^-)}{h_{s-1}}\widehat{\phi}_i'(1)\widehat{\phi}_j(1)$$
$$[\mathbf{C}_{\text{cons}}^{(s-1,s)}]_{i,j} = \frac{c(x_s^+)}{h_s}\widehat{\phi}_j'(-1)\widehat{\phi}_i(1) - \frac{c(x_s^-)}{h_{s-1}}\widehat{\phi}_i'(1)\widehat{\phi}_j(-1)$$
$$[\mathbf{C}_{\text{cons}}^{(s,s)}]_{i,j} = -\frac{c(x_s^+)}{h_s}\widehat{\phi}_j'(-1)\widehat{\phi}_i(-1) - \frac{c(x_s^+)}{h_s}\widehat{\phi}_i'(-1)\widehat{\phi}_j(-1)$$

Example 1.3. *Consider $c \equiv 1$ for $\mathcal{P}^1$-elements ($r = 1$) with an equidistant mesh with meshsize h we have*

$$\widehat{\boldsymbol{B}}_{cons}^{(s)} = \frac{1}{h}\begin{bmatrix} 0 & -1/2 & 1/2 & 0 \\ -1/2 & 1 & -1 & 1/2 \\ 1/2 & -1 & 1 & -1/2 \\ 0 & 1/2 & -1/2 & 0 \end{bmatrix}$$

Boundary Faces

For $s \in \{0, N+1\}$ and x_s a boundary face we now have a smaller local contribution matrix since the contribution can only come from the one element to which x_s belongs. Meaning we have $\widehat{\mathbf{B}}_{\text{cons}}^{(0)}, \widehat{\mathbf{B}}_{\text{cons}}^{(N+1)} \in \mathbb{R}^{(r+1)\times(r+1)}$ with

$$[\widehat{\mathbf{B}}_{\text{cons}}^{(0)}]_{i,j} = -\frac{2c(x_0^+)}{h_0}\widehat{\phi}_j'(-1)\widehat{\phi}_i(-1) - \frac{2c(x_0^+)}{h_0}\widehat{\phi}_i'(-1)\widehat{\phi}_j(-1)$$
$$[\widehat{\mathbf{B}}_{\text{cons}}^{(N+1)}]_{i,j} = \frac{2c(x_{N+1}^-)}{h_{N+1}}\widehat{\phi}_j'(1)\widehat{\phi}_i(1) + \frac{2c(x_{N+1}^-)}{h_{N+1}}\widehat{\phi}_i'(1)\widehat{\phi}_j(1)$$

Example 1.4. *For $c \equiv 1$ with $\mathcal{P}^1$-elements ($r = 1$) we have*

$$\widehat{\boldsymbol{B}}_{cons}^{(0)} = \frac{1}{h_0}\begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix}, \qquad \widehat{\boldsymbol{B}}_{cons}^{(N+1)} = \frac{1}{h_{N+1}}\begin{bmatrix} 0 & -1 \\ -1 & 2 \end{bmatrix}$$

1.7.3 Assembly of B penalty part

Again we rewrite $\mathbf{B}_{\text{penal}} = \sum_{s=0}^{N+1} \mathbf{B}_{\text{penal}}^{(s)}$, where

$$[\mathbf{B}_{\text{penal}}^{(s)}]_{T(n,i),T(m,j)} = \mathtt{a}_s[\Phi_j^m(x_s)][\Phi_i^n(x_s)]$$

We proceed analogously to 1.7.2.

Interior Faces

Let $s \in \{1, \dots, N\}$ and x_s denote an interior face. As before we have that the entries of $\mathbf{B}_{\text{penal}}^{(s)}$ can only be nonzero for an index tuple $(T(n,i), T(m,j))$ if $n, m \in \{s, s-1\}$, similar to 1.7.2 we find that the assembly boils down to adding up local contributions represented in a local contribution matrix

$$\widehat{\mathbf{B}}_{\text{penal}}^{(s)} = \begin{bmatrix} \mathbf{C}_{\text{penal}}^{(s-1,s-1)} & \mathbf{C}_{\text{penal}}^{(s-1,s)} \\ \mathbf{C}_{\text{penal}}^{(s,s-1)} & \mathbf{C}_{\text{penal}}^{(s,s)} \end{bmatrix} \in \mathbb{R}^{2(r+1)\times 2(r+1)}$$

and using (1.15), and the definition of the penalization parameter (1.8) we specifically find

$$[\mathbf{C}_{\text{penal}}^{(s-1,s-1)}]_{i,j} = \mathtt{a}_s\widehat{\phi}_j(1)\widehat{\phi}_i(1)$$
$$[\mathbf{C}_{\text{penal}}^{(s-1,s)}]_{i,j} = -\mathtt{a}_s\widehat{\phi}_j(-1)\widehat{\phi}_i(1)$$
$$[\mathbf{C}_{\text{penal}}^{(s,s)}]_{i,j} = \mathtt{a}_s\widehat{\phi}_j(-1)\widehat{\phi}_i(-1)$$

where again by symmetry of the penalty term we have $\mathbf{C}_{\text{penal}}^{(s-1,s)} = (\mathbf{C}_{\text{penal}}^{(s,s-1)})^T$ and $\mathtt{a}_s$ only depends on the two adjacent elements I_{s-1}, I_s.

Example 1.5. *Consider $c \equiv 1$ for $\mathcal{P}^1$-elements ($r = 1$) with an equidistant mesh with meshsize h we have*

$$\widehat{\boldsymbol{B}}_{penal}^{(s)} = \frac{\sigma}{h}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Boundary Faces

For $s \in \{0, N+1\}$ we again have only the respective boundary element contributing. So the local contribution matrices $\widehat{\mathbf{B}}_{\text{penal}}^{(s)} \in \mathbb{R}^{(r+1)\times(r+1)}$ satisfy

$$[\widehat{\mathbf{B}}_{\text{penal}}^{(0)}]_{i,j} = \mathtt{a}_0 \widehat{\phi}_j(-1)\widehat{\phi}_i(-1), \qquad [\widehat{\mathbf{B}}_{\text{penal}}^{(N+1)}]_{i,j} = \mathtt{a}_{N+1} \widehat{\phi}_j(1)\widehat{\phi}_i(1)$$

Example 1.6. *For $c \equiv 1$ with $\mathcal{P}^1$-elements ($r = 1$) we have*

$$\widehat{\boldsymbol{B}}_{penal}^{(0)} = \frac{\sigma}{h_0}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \widehat{\boldsymbol{B}}_{penal}^{(N+1)} = \frac{\sigma}{h_{N+1}}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

1.7.4 System Vector Assembly

We divide assembling the vector $\mathbf{l}$ in (1.10) into two parts.

$$\mathbf{l} = \mathbf{l}_{\text{load}} + \mathbf{l}_{\text{bc}}$$

First we recall the assembly of the load vector $\mathbf{l}_{\text{load}}$, i.e. the vector containing the contributions of the forcing term f and secondly we will describe how to add the Dirichlet boundary condition contributions ($\mathbf{l}_{\text{bc}}$).

Load Vector

The assembly of the load vector is completely analogous to the continuous finite element case. Using the local support of Φ_i^n we can rewrite

$$\int_\Omega f\Phi_i^n \mathrm{d}x = \sum_{s=0}^{N} \int_{I_s} f\Phi_i^n \mathrm{d}x = \int_{I_n} f(x)\Phi_i^n(x)\mathrm{d}x = \frac{h_n}{2}\int_{-1}^{1} f(F_n(\xi))\widehat{\phi}_i(\xi)\mathrm{d}\xi$$

meaning as before we can assemble $\mathbf{l}_{\text{load}} = \sum_{s=0}^{N} \mathbf{l}_{\text{load}}^{(s)}$ where

$$[\mathbf{l}_{\text{load}}^{(s)}]_{T(n,i)} = \int_{I_s} f\Phi_i^n \mathrm{d}x = \delta_{n,s}\frac{h_n}{2}\int_{-1}^{1} f(F_n(\xi))\widehat{\phi}_i(\xi)\mathrm{d}\xi$$

which can be characterized by the local contribution vector $\widehat{\mathbf{l}}_{\text{load}}^{(s)} \in \mathbb{R}^{r+1}$ defined as

$$[\widehat{\mathbf{l}}_{\text{load}}^{(s)}]_i = \frac{h_s}{2}\int_{-1}^{1} f(F_s(\xi))\widehat{\phi}_i(\xi)\mathrm{d}\xi$$

In practice we approximate the integral using the Gauss-Lobatto quadrature rule.

Example 1.7. *For f piecewise constant (i.e. $f|_{I_s} \equiv f_s \in \mathbb{R} \quad \forall s = 0, \ldots, N$) with $\mathcal{P}^1$-elements ($r = 1$) we have*

$$\widehat{\mathbf{l}}_{\text{load}}^{(s)} = \frac{f_s h_s}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Dirichlet Boundary Condition Vector

We have

$$[\mathbf{l}_{\mathrm{bc}}]_{T(n,i)} = -g_1 c(x_{N+1}^-)\Phi_i^n(x_{N+1}^-) + g_0 c(x_0^+)\Phi_i^n(x_0^+) + \mathtt{a}_{N+1} g_1 \Phi_i^n(x_{N+1}^-) + \mathtt{a}_0 g_0 \Phi_i^n(x_0^+)$$

where the entries can clearly only be non-zero at indices corresponding to boundary elements. We characterize the assembly using the local contribution vectors $\widehat{\mathbf{l}}_{\mathrm{bc}}^{(N+1)}, \widehat{\mathbf{l}}_{\mathrm{bc}}^{(0)} \in \mathbb{R}^{r+1}$ where

$$[\widehat{\mathbf{l}}_{\mathrm{bc}}^{(0)}]_i = g_0 c(x_0^-)\widehat{\phi}_i(-1) + \mathtt{a}_0 g_0 \widehat{\phi}_i(-1), \qquad [\widehat{\mathbf{l}}_{\mathrm{bc}}^{(N+1)}]_i = -g_1 c(x_{N+1}^-)\widehat{\phi}_i(1) + \mathtt{a}_{N+1} g_1 \widehat{\phi}_i(1)$$

1.8 Existence of Discrete Solution

Firstly we will recall some basic definitions:

Definition 1.8. *Let V be a normed vector space and $b : V \times V \to \mathbb{R}$ be a bilinear form.*

(i) *We say b is **continuous** if $\exists C_{cont} > 0$, such that*

$$|b(u,v)| \leq C_{cont} \|u\| \, \|v\| \qquad \forall u, v \in V$$

(ii) *We say b is **symmetric** if*

$$b(u,v) = b(v,u) \qquad \forall u, v \in V$$

(iii) *We say b is **coercive** if $\exists C_{coer} > 0$, such that*

$$b(u,u) \geq C_{coer} \|u\|^2 \qquad \forall u \in V$$

Since (1.9) corresponds to the finite dimensional system (1.10) uniqueness and existence of a solution are equivalent. The bilinear form b_h is *symmetric* by construction the goal of this section is to show that b_h is also *coercive*. From the coercivity of b_h it will follow that the matrix $\mathbf{B}$ in (1.10) is positive definite and hence invertible, which means there exists a (unique) solution of (1.9).

Lemma 1.9. *Let $V = span(\varphi_1, \ldots, \varphi_M)$ be a finite dimensional normed vector space with $\dim(V) = M \in \mathbb{N}$ and let $b : V \times V \to \mathbb{R}$ be a symmetric, coercive bilinear form, then the matrix $[\boldsymbol{B}]_{i,j} = [b(\varphi_j, \varphi_i)]_{i,j} \in \mathbb{R}^{N \times N}$ is symmetric positive definite.*

Proof. Clearly $\mathbf{B}$ is symmetric.
Let $\mathbf{v} = (v_1, \ldots, v_M) \in \mathbb{R}^M$ then $v = \sum_{i=1}^M v_i \varphi_i \in V$ and we have:

$$\mathbf{v}^T \mathbf{B} \mathbf{v} = \sum_{i,j=1}^M v_i v_j b(\varphi_j, \varphi_i) = b(v,v) \geq C_{\mathrm{coer}} \|w\|^2$$

where we have used the bilinearity and the coercivity of b. □

Next we will require a usefull tool often used in FEM proofs to bound a boundary integral with the integral over the interior domain. These kind of inequalities are in the literature often called *inverse (trace) inequalities* and are in essence trace inequalities on finite dimensional subspaces. We will here rely on a result and proof as presented in [7].

Lemma 1.10 (Inverse inequality)**.** *Let $r \in \mathbb{N}$ be the polynomial degree, $a, b \in \mathbb{R}$ with $a < b$ and let $\mathcal{P}^r([a,b])$ denote the space of polynomials of degree r defined on $[a,b]$.*

For any $v \in \mathcal{P}^r([a,b])$ we have:

1. $|v(a)|^2 \leq \dfrac{(r+1)^2}{|b-a|}\|v\|^2_{L^2([a,b])}$
2. $|v(b)|^2 \leq \dfrac{(r+1)^2}{|b-a|}\|v\|^2_{L^2([a,b])}$

Proof. We will prove the statements first for the reference element $\hat{I} = [-1,1]$ and then use a scaling argument to show the general case by applying a simple substitution.

Step 1 (*Setup*)**.**
We will make use of the Legendre orthonormal basis of $\mathcal{P}^r(\hat{I})$: Let $P_0, \dots, P_r$ denote the Legegndre polynomials on $\mathcal{P}^r(\hat{I})$. Recall the following well known facts (see for example [5]):

1. $\{P_0, \dots, P_r\}$ form an orthogonal basis of $\mathcal{P}^r(\hat{I})$ under the $L^2(\hat{I})$ inner product. Meaning:
$$\operatorname{span}(P_0, \dots, P_r) = \mathcal{P}^r(\hat{I}), \quad \int_{-1}^{1} P_i P_j \,\mathrm{d}\xi = \begin{cases} \frac{2}{2i+1}, & \text{for } i = j \\ 0, & \text{for } i \neq j \end{cases}$$
2. $P_i(1) = 1, P_i(-1) = (-1)^i, \qquad \forall i = 0, \dots, r$

Let $\psi_i = \sqrt{\frac{(2i+1)}{2}} P_i$ for $i = 0, \dots, r$ denote the normed basis function. Clearly we now have

$$\psi_i(-1) = (-1)^i \sqrt{\frac{2i+1}{2}}, \qquad \psi_i(1) = \sqrt{\frac{2i+1}{2}}, \qquad \int_{-1}^{1} \psi_i \psi_j \mathrm{d}\xi = \delta_{i,j}, \qquad \forall i = 0, \dots, r$$

where $\delta_{i,j} = \begin{cases} 1, & \text{for } i = j \\ 0, & \text{for } i \neq j \end{cases}$, and hence $\{\psi_0, \dots, \psi_r\}$ form an orthonomal basis.

Step 2 (*Proof on reference element*)**.**
For any $v \in \mathcal{P}^r(\hat{I})$ there exist coefficients $v_0, \dots, v_r \in \mathbb{R}$, such that $v = \sum_{i=0}^r v_i \psi_i$. By applying Cauchy-Schwarz we find

$$|v(-1)|^2 = \Big|\sum_{i=0}^r v_i \psi_i(-1)\Big|^2 \leq \Big(\sum_{i=0}^r v_i^2\Big)\Big(\sum_{i=0}^r \psi_i(-1)^2\Big) = \Big(\sum_{i=0}^r v_i^2\Big)\Big(\sum_{i=0}^r \frac{2i+1}{2}\Big) = \Big(\sum_{i=0}^r v_i^2\Big)\frac{(r+1)^2}{2}$$

and finally the orthonormality of the ψ_i yields

$$\frac{(r+1)^2}{2}\sum_{i=0}^r v_i^2 = \frac{(r+1)^2}{2}\sum_{i,j=0}^r v_i v_j \delta_{i,j} = \frac{(r+1)^2}{2}\|v\|^2_{L^2(\hat{I})}$$

This yields the first inequality for the reference element. The second inequality can be proven analogously.

Step 3 (*Scaling argument*)**.**
Now we assume that $v \in \mathcal{P}^r([a,b])$. Using the affine (element) map

$$F : [-1,1] \to [a,b], \xi \mapsto \frac{a+b}{2} + \frac{|b-a|}{2}\xi$$

we can pull v back to the reference element by defining $\widehat{v}(\xi) := v(F(\xi))$ for all $\xi \in \hat{I}$. Clearly $\widehat{v} \in \mathcal{P}^r(\hat{I})$ hence, by Step 2 we obtain

$$|v(a)|^2 = |\widehat{v}(F^{-1}(a))|^2 = |\widehat{v}(-1)|^2 \leq \frac{(r+1)^2}{2}\int_{-1}^{1} \widehat{v}(\xi)^2 \mathrm{d}\xi = \frac{(r+1)^2}{2}\frac{2}{|b-a|}\|v\|^2_{L^2([a,b])}$$

where in the last equality we have applied a change of variable $x = F(\xi)$ to the integral. Applying the same line of reasoning to $|v(b)|^2$ proves both inequalities and so we are done.

□

Recall the in previous sections established notations, let $r \in \mathbb{N}$ denote the polynomial degree and $V_h^r(\mathcal{T}_h)$ be the discrete subspace.

Definition 1.11. *We define the* ***energy norm*** *on V_h by*

$$\|v\|_h^2 := \sum_{n=0}^{N} \int_{I_n} c(x)v'(x)^2\,dx + \sum_{n=0}^{N+1} \mathtt{a}_n[v(x_n)]^2 \tag{1.16}$$

where $\mathtt{a}$ *denotes the penalization term in (1.8).*

Lemma 1.12. *$\|\cdot\|_h$ defines a norm on V_h.*

Proof. Clearly we have $\|\lambda v\|_h = |\lambda|\,\|v\|_h$ for all $\lambda \in \mathbb{R}, v \in V_h$.

By definition we have $\mathtt{a}, c > 0$ and by extension $\|v\|_h \geq 0$ for all $v \in V_h$. Suppose now that $\|v\|_h = 0$ for some $v \in V_h$, then we must have $v|_{I_n} \equiv \text{const}$ and $[v(x_n)] = 0$ for all n. So v must be constant on all elements and have a jump of zero at the element boundaries. These two facts combined imply that v is constant on all of Ω. By the definition of the jump at the boundary nodes of Ω it immediately follows that $v = 0$. Clearly $\|0\|_h = 0$, therefore $\|\cdot\|_h$ is positive definite.

Using $[v(x_n) + w(x_n)] = [v(x_n)] + [w(x_n)] \quad \forall v, w \in V_h, n = 0, \ldots, N+1$ we find

$$\begin{aligned}\|v + w\|_h &\leq \Big(\sum_{n=0}^{N} (\|\sqrt{c}v'\|_{L^2(I_n)} + \|\sqrt{c}w'\|_{L^2(I_n)})^2 + \sum_{n=0}^{N+1} (\sqrt{\mathtt{a}_n}([v(x_n)] + [w(x_n)]))^2 \Big)^{1/2} \\ &\leq \|v\|_h + \|w\|_h\end{aligned}$$

where in the last inequality we have used the triangle inequality of the euclidian vector norm on $\mathbb{R}^{2N+3}$, with the vector given as

$$\mathbf{v} = \big[\|\sqrt{c}v'\|_{L^2(I_0)}, \ldots, \|\sqrt{c}v'\|_{L^2(I_N)}, \sqrt{\mathtt{a}_0}[v(x_0)], \ldots, \sqrt{\mathtt{a}_{N+1}}[v(x_{N+1})]\big]^T$$

this shows the triangle inequality for $\|\cdot\|_h$ and hence it is a norm. □

Theorem 1.13. *Let $r \in \mathbb{N}$, the bilinear form b_h in (1.9) is continuous on $V_h^r(\mathcal{T}_h)$ and if furthermore $\sigma \geq \frac{6(r+1)^2 c_{\max}}{c_{\min}}$, b_h is also coercive on $V_h^r(\mathcal{T}_h)$. The coercivity and continuity constants are given by*

$$C_{coer} = \frac{1}{2}, \qquad C_{cont} = (3 + \frac{5}{4}C_\sigma)$$

where $C_\sigma = \frac{(r+1)^2 c_{\max}}{\sigma c_{\min}}$

Proof. **Step 1** (*Coercivity*)**.**
Let $w \in V_h$. Note that

$$b_h(w, w) = \|w\|_h^2 - 2\sum_{n=0}^{N+1} \{c(x_n)w'(x_n)\}[w(x_n)] \tag{1.17}$$

To derive the coercivity of b_h we will estimate the term $2\sum_{n=0}^{N+1}\{c(x_n)w'(x_n)\}[w(x_n)]$ from above applying Lemma 1.10 and additional smaller tools:
Using the general fact $2ab \le a^2+b^2, \forall a,b\in\mathbb{R}$ we estimate

$$\begin{aligned}
&2\sum_{n=0}^{N+1}\{c(x_n)w'(x_n)\}[w(x_n)] = 2\sum_{n=0}^{N+1}\{c(x_n)w'(x_n)\}\Big(\frac{\mathtt{a}_n}{2}\Big)^{-1/2}\Big(\frac{\mathtt{a}_n}{2}\Big)^{1/2}[w(x_n)]\\
&\le 2\sum_{n=0}^{N+1}\frac{\{c(x_n)w'(x_n)\}^2}{\mathtt{a}_n}+\frac{1}{2}\sum_{n=0}^{N+1}\mathtt{a}_n[w(x_n)]^2 \qquad (1.18)
\end{aligned}$$

Recalling $\mathtt{a}_n=\sigma\mathtt{c}_n\mathtt{h}_n^{-1}$ from (1.8) and noting the relations $\mathtt{h}_n\le h_n, \mathtt{c}_n^{-1}\le c(x_n^-)^{-1}, c(x_n^+)^{-1}$ we find

$$\begin{aligned}
&\mathtt{a}_n^{-1}c(x_n^+)\le\frac{h_n}{\sigma},\quad \mathtt{a}_n^{-1}c(x_n^-)\le\frac{h_{n-1}}{\sigma},\quad \forall n=1,\dots,N\\
&\mathtt{a}_0^{-1}c(x_0^+)=\frac{h_0}{\sigma},\quad \mathtt{a}_{N+1}^{-1}c(x_{N+1}^-)=\frac{h_N}{\sigma}
\end{aligned}$$

applying this and the usefull inequality $(a+b)^2\le 2a^2+2b^2$ yields

$$\begin{aligned}
&2\sum_{n=0}^{N+1}\frac{\{c(x_n)w'(x_n)\}^2}{\mathtt{a}_n}\\
&=2\sum_{n=1}^{N}\frac{1}{4\mathtt{a}_n}\Big(c(x_n^-)w'(x_n^-)+c(x_n^+)w'(x_n^+)\Big)^2+\frac{2}{\mathtt{a}_0}\Big(c(x_0^+)w'(x_0^+)\Big)^2+\frac{2}{\mathtt{a}_{N+1}}\Big(c(x_{N+1}^-)w'(x_{N+1}^-)\Big)^2\\
&\le 2\sum_{n=1}^{N}\frac{1}{2\sigma}\Big(h_{n-1}c(x_n^-)w'(x_n^-)^2+h_nc(x_n^+)w'(x_n^+)^2\Big)+\frac{2h_0}{\sigma}c(x_0^+)w'(x_0^+)^2+\frac{2h_N}{\sigma}c(x_{N+1}^-)w'(x_{N+1}^-)^2\\
&\le\frac{c_{\max}}{\sigma}\sum_{n=1}^{N}\Big(h_{n-1}w'(x_n^-)^2+h_nw'(x_n^+)^2\Big)+\frac{2c_{\max}h_0}{\sigma}w'(x_0^+)^2+\frac{2c_{\max}h_N}{\sigma}w'(x_{N+1}^-)^2 \qquad (1.19)
\end{aligned}$$

Since $w\in V_h$ is a (broken) polynomial, we can apply Lemma 1.10 elementwise and find

$$w'(x_n^+)^2, w'(x_{n+1}^-)^2\le\frac{(r+1)^2}{h_n}\|w'\|^2_{L^2(I_n)}\quad\forall n=0,\dots,N \qquad (1.20)$$

By combining (1.19), (1.20) and inserting $1=c_{\min}c_{\min}^{-1}\le c(x)c_{\min}^{-1}\quad\forall x\in\Omega$ we find

$$2\sum_{n=0}^{N+1}\frac{\{c(x_n)w'(x_n)\}^2}{\mathtt{a}_n}\le 3C_\sigma\sum_{n=0}^{N}\|\sqrt{c}w'\|^2_{L^2(I_n)} \qquad (1.21)$$

for $C_\sigma:=\dfrac{(r+1)^2c_{\max}}{\sigma c_{\min}}>0$.

Finally putting together (1.17), (1.18) and (1.21) yields

$$\begin{aligned}
b_h(w,w)&\ge\|w\|_h^2-3C_\sigma\sum_{n=0}^{N}\|\sqrt{c}w'\|^2_{L^2(I_n)}-\frac{1}{2}\sum_{n=0}^{N+1}\mathtt{a}_n[w(x_n)]^2\\
&=(1-3C_\sigma)\sum_{n=0}^{N}\|\sqrt{c}w'\|^2_{L^2(I_n)}+\frac{1}{2}\sum_{n=0}^{N+1}\mathtt{a}_n[w(x_n)]^2\\
&\ge\frac{1}{2}\|w\|_h^2
\end{aligned}$$

for $\sigma\ge\frac{6(r+1)^2c_{\max}}{c_{\min}}$, which proves the coercivity of b_h on V_h.

Step 2 (*Continuity*)**.**

The proof the continuity of b_h uses similar ideas as the coercivity proof. Let $u, v \in V_h$, by using Cauchy-Schwarz we immediately get

$$\begin{aligned}|b_h(u,v)| &\leq \sum_{n=0}^{N} \|\sqrt{c}u'\|_{L^2(I_n)} \|\sqrt{c}v'\|_{L^2(I_n)} + \sum_{n=0}^{N+1} \big|\{c(x_n)u'(x_n)\}[v(x_n)]\big| \\ &+ \sum_{n=0}^{N+1} \big|\{c(x_n)v'(x_n)\}[u(x_n)]\big| + \sum_{n=0}^{N+1} \mathtt{a}_n \big|[u(x_n)][v(x_n)]\big| \\ &=: T_{\mathrm{ell}} + T_{\mathrm{cons}}^{(u)} + T_{\mathrm{cons}}^{(v)} + T_{\mathrm{penal}} \end{aligned} \tag{1.22}$$

The goal is now to estimate the consistency terms T_{cons} from above by something of the form $\sum_{n=0}^{N+1} t_n(u)s_n(v) + \sum_{n=0}^{N+1} t_n(v)s_n(u)$, such that together with the terms $T_{\mathrm{ell}}, T_{\mathrm{penal}}$ we can use discrete Cauchy-Schwarz on the sums and hence separate them into a product of the two energy norms $C_{\mathrm{cont}}\|u\|_h\|v\|_h$ scaled by a positive constant.

We will show the estimate of $T_{\mathrm{cons}}^{(u)}$, the procedure to estimate $T_{\mathrm{cons}}^{(v)}$ is analogous.
First rewrite

$$T_{\mathrm{cons}}^{(u)} = \sum_{n=0}^{N+1} \big|\{c(x_n)u'(x_n)\}\mathtt{a}_n^{-1/2}\mathtt{a}_n^{1/2}[v(x_n)]\big| \tag{1.23}$$

Next again using the definition of $\mathtt{a}$ and estimates as in Step 1 we find for interior faces $n = 1, \dots, N$

$$\big|\{c(x_n)u'(x_n)\}\big|\mathtt{a}_n^{-1/2} \leq \frac{1}{2}\sqrt{\frac{\mathtt{h}_n}{\sigma}}\sqrt{c_{\max}}\left(|u'(x_n^-)| + |u'(x_n^+)|\right)$$

and for the boundary faces

$$\big|\{c(x_0)u'(x_0)\}\big|\mathtt{a}_0^{-1/2} \leq \sqrt{\frac{\mathtt{h}_0}{\sigma}}\sqrt{c_{\max}}|u'(x_0^+)|, \quad \big|\{c(x_{N+1})u'(x_{N+1})\}\big|\mathtt{a}_{N+1}^{-1/2} \leq \sqrt{\frac{\mathtt{h}_N}{\sigma}}\sqrt{c_{\max}}|u'(x_{N+1}^-)|$$

Applying Lemma (1.10) yields for $\beta_n(u) := \sqrt{C_\sigma}\|\sqrt{c}u'\|_{L^2(I_n)}, n = 0, \dots, N$

$$\begin{aligned}&\big|\{c(x_n)u'(x_n)\}\big|\mathtt{a}_n^{-1/2} \leq \frac{\beta_{n-1}(u)}{2} + \frac{\beta_n(u)}{2} \quad \text{for } n = 1, \dots, N \\ &\big|\{c(x_0)u'(x_0)\}\big|\mathtt{a}_0^{-1/2} \leq \beta_0(u) \\ &\big|\{c(x_{N+1})u'(x_{N+1})\}\big|\mathtt{a}_{N+1}^{-1/2} \leq \beta_N(u)\end{aligned}$$

which we can now plug back into (1.23) to get

$$T_{\mathrm{cons}}^{(u)} \leq \beta_0(u)\gamma_0(v) + \beta_N(u)\gamma_{N+1}(v) + \sum_{n=1}^{N} \frac{\beta_{n-1}(u)}{2}\gamma_n(v) + \sum_{n=1}^{N} \frac{\beta_n(u)}{2}\gamma_n(v) \tag{1.24}$$

for $\gamma_n(v) := \sqrt{\mathtt{a}_n}\big|[v(x_n)]\big| \, \forall n = 0, \dots, N+1$. By furthermore denoting $\alpha_n(u) := \|\sqrt{c}u'\|_{L^2(I_n)}$ we can represent

$$T_{\mathrm{ell}} = \sum_{n=0}^{N} \alpha_n(u)\alpha_n(v), \qquad T_{\mathrm{penal}} = \sum_{n=0}^{N+1} \gamma_n(u)\gamma_n(v)$$

and in total for

$$\begin{aligned}\mathbf{u} &:= [\alpha_0(u),\dots,\alpha_N(u),\beta_0(u),\beta_N(u),\frac{\beta_0(u)}{2},\dots,\frac{\beta_{N-1}(u)}{2},\frac{\beta_1(u)}{2},\dots,\frac{\beta_N(u)}{2},\\ &\qquad \gamma_0(u),\gamma_{N+1}(u),\gamma_1(u),\dots,\gamma_N(u),\gamma_1(u),\dots,\gamma_N(u),\gamma_0(u),\dots,\gamma_{N+1}(u)]^T \in \mathbb{R}^{6N+7}\\ \mathbf{v} &:= [\alpha_0(v),\dots,\alpha_N(v),\gamma_0(v),\gamma_{N+1}(v),\gamma_1(v),\dots,\gamma_N(v),\gamma_1(v),\dots,\gamma_N(v),\\ &\qquad \beta_0(v),\beta_N(v),\frac{\beta_0(v)}{2},\dots,\frac{\beta_{N-1}(v)}{2},\frac{\beta_1(v)}{2},\dots,\frac{\beta_N(v)}{2},\gamma_0(v),\dots,\gamma_{N+1}(v)]^T \in \mathbb{R}^{6N+7}\end{aligned}$$

we get

$$\begin{aligned}&T_{\text{ell}} + T_{\text{cons}}^{(u)} + T_{\text{cons}}^{(v)} + T_{\text{penal}} \le \mathbf{u}^T\mathbf{v} \le |\mathbf{u}|\,|\mathbf{v}|\\ &\le \Big(\sum_{n=0}^{N}(1+\frac{5}{4}C_\sigma)\|\sqrt{c}u'\|^2_{L^2(I_n)} + 3\sum_{n=0}^{N+1}\mathtt{a}_n[u(x_n)]^2\Big)^{1/2}\Big(\sum_{n=0}^{N}(1+\frac{5}{4}C_\sigma)\|\sqrt{c}v'\|^2_{L^2(I_n)} + 3\sum_{n=0}^{N+1}\mathtt{a}_n[v(x_n)]^2\Big)^{1/2}\\ &\le C_{\text{cont}}\|u\|_h\|v\|_h\end{aligned}$$

where $C_{\text{cont}} := (3+\frac{5}{4}C_\sigma)$. This last estimate together with 1.22 proves the continuity of b_h. □

1.9 Numerical Results

1.9.1 Rate of Convergence

To replicate the theoretical rate of convergence we first consider a sequence of uniform meshes. Let $h_l = 2^{-l}$ denote the global (uniform) meshsize and $\mathcal{T}_h^{(l)}$ denote the partitions of Ω for $l = 2,\dots,9$. As an example we have $\mathcal{T}_h^{(2)} = \{(0,0.25),(0.25,0.5),(0.5,0.75),(0.75,1)\}$. To begin we worked with the constant coefficient $c \equiv 1$. In this case choosing $\sigma = 10(r+1)^2$ was more than enough to guarantee positive definiteness of the system matrix. We tested our methods programmed in `MATLAB` using some very simple exact solutions $u \in C^1(\Omega)$, approximated them numerically finding some $u_h \in V_h^r(\mathcal{T}_h^{(l)})$ (for an a priori fixed polynomial degree r) and calculated the L^2-norm

$$\|u-u_h\|_{L^2(\Omega)} = \Big(\int_\Omega |u(x)-u_h(x)|^2\mathrm{d}x\Big)^{1/2} = \Big(\sum_{n=0}^{N_l}\int_{I_n}|u(x)-u_h(x)|^2\mathrm{d}x\Big)^{1/2}$$

and the broken H^1-norm

$$\|u-u_h\|_{H^1(\mathcal{T}_h^{(l)})} = \Big(\sum_{n=0}^{N_l}\int_{I_n}|u'(x)-u_h'(x)|^2\mathrm{d}x\Big)^{1/2}.$$

Note that even though u_h is not in $H^1(\Omega)$ and we can therefore not calculate the normal H^1-norm, we do have

$$u, u_h \in H^1(\mathcal{T}_h^{(l)}) = \{v \in L^2(\Omega) | v|_{I_n} \in H^1(I_n) \text{ for } n = 0,\dots,N_l\}$$

so taking the broken Sobolev norm makes sense also for V_h functions.

Firstly to ensure the exactness of the method we chose the exact solution $u(x) = x^r$ and approximated it using $\mathcal{P}^r$ elements. As expected we got a minimal floating-point error of around 10^{-15} in both L^2-, and $H^1(\mathcal{T}_h^{(l)})$-norms, slightly increasing for increasing l due to the growing condition number of the system matrix. This indicates exactness of the method.

Next we considered the exact solution $u(x) = e^{-x}\sin(x)$, we observed the with the theory conforming convergence rates $\mathcal{O}(h^r), \mathcal{O}(h^{r+1})$, in the $H^1(\mathcal{T}_h)$- and L^2-norm respectively for $\mathcal{P}^r$-elements. As visible in figures 1.1, 1.2

We plotted the numerical solution for $\mathcal{P}^1, \mathcal{P}^2$-elements in the figures 1.3, 1.4 respectively on the partition $\mathcal{T}_h^{(3)}$, i.e. for 8 elements with meshsize $h = \frac{1}{8}$. By eye the solution seems continuous, this is due to the high

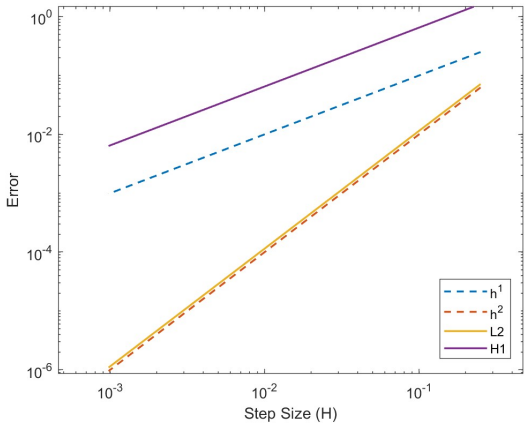

Figure 1.1: Errors of SIPG for P^1-elements

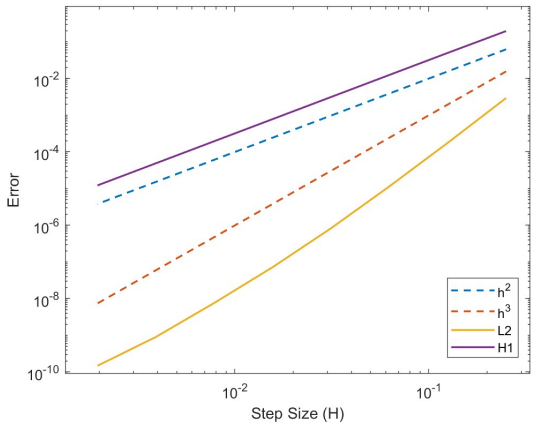

Figure 1.2: Errors of SIPG for P^2-elements

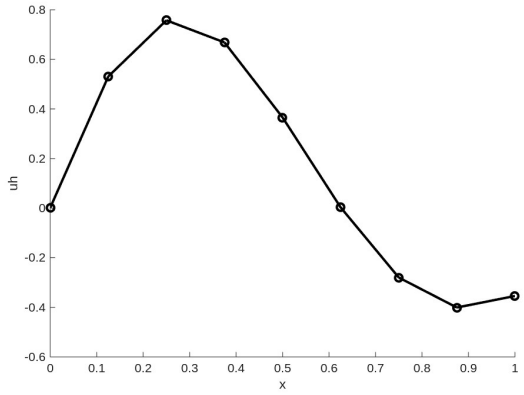

Figure 1.3: numerical SIPG-approximation for P^1-elements

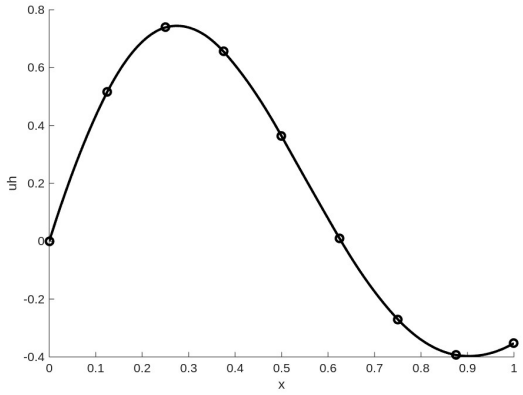

Figure 1.4: numerical SIPG-approximation for P^2-elements

penalization parameter $\sigma = 10(r+1)^2$. In fact the solution is not continuous and the boundary conditions are not exact either, the error is just not noticeable by eye.

If we reduce the penalization parameter enough such that the bilinear form is no longer coercive we can observe the jumps growing and the discontinuity becomes apparent as is visible in figures 1.5, 1.6 where we have set $\sigma = 1.1$. Maybe interesting to observe is the effect of enforcing boundary conditions weakly. Clearly here the Dirichlet boundary condition $u(0) = 0$ is not fulfilled by the numerical solutions, which would be the case for strongly enforced boundary conditions.

Finally we repeated the tests above using a non constant yet still smooth coefficient $c(x) = \sin(10x) + 2$. We observed the same convergence rates but slightly higher errors overall. This was to be expected since for a non-constant c we introduce a quadrature error.

1.9.2 Influence of Quadrature Rule on the Convergence Rate

As noted in section 1.6 by using $r+1$ Gauss-Lobatto nodes (in the context of $\mathcal{P}^r$-elements) as quadrature nodes and basis nodes at the same time introduces an error when assembling the mass matrix of the system. In this subsection we experimentally compare the results of using $r+1$ Gauss-Lobatto quadrature nodes to approximate the integrals versus using the exact integration values. To be precise we fix a polynomial degree r and our Lagrangian basis with $r+1$ Gauss-Lobatto nodes as specified in section 1.6 and approximate all

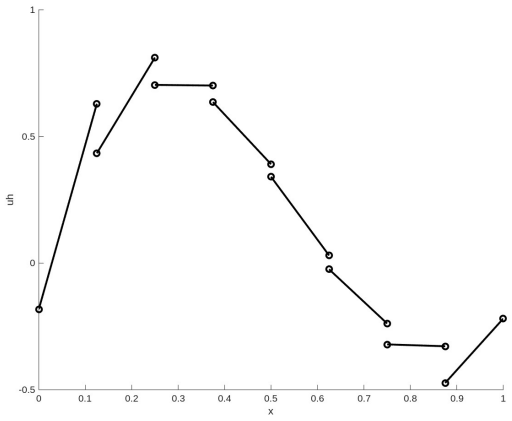

Figure 1.5: numerical SIPG-approximation for P^1-elements with small penalty

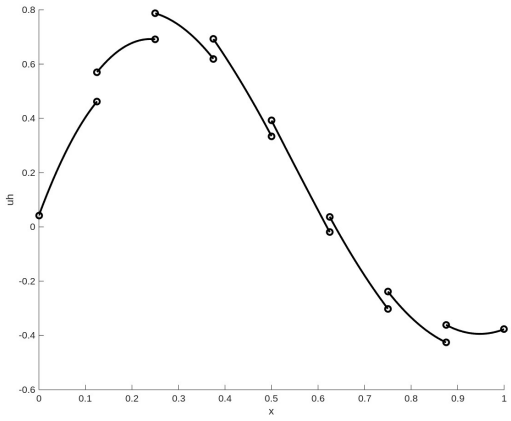

Figure 1.6: numerical SIPG-approximation for P^2-elements with small penalty

integrals first using the Gauss-Lobatto quadrature rule with $r+1$ nodes, then with $r+2$ nodes and compare the two.
First we have to consider a slightly different elliptical problem, which requires a mass matrix.

$$-(c(x)u'(x))' + u(x) = f(x) \qquad \forall x \in \Omega$$

$$u(0) = g_0, u(1) = g_1$$

We can apply the exact same tools as in the derivation of the variational formulation in section 1.3 and get the discrete SIPG variational formulation.
Find $u_h \in V_h$ such that:

$$b_h(u_h, v) + (u_h, v)_{L^2(\Omega)} = \ell(v), \qquad \forall v \in V_h \tag{1.25}$$

Now analogously to section 1.5 we write $u_h = \sum_{m=0}^{N} \sum_{j=0}^{r} \alpha_j^m \Phi_j^m \in V_h$ and find that (1.25) is equivalent to

$$\sum_{m=0}^{N} \sum_{j=0}^{r} \alpha_j^m \Big(b_h(\Phi_j^m, \Phi_i^n) + (\Phi_j^m, \Phi_i^n)_{L^2(\Omega)} \Big) = \ell(\Phi_i^n) \qquad \forall i \in \{0, \dots, r\}, \{n = 0, \dots, N\}$$

which in turn is equivalent to the Matrix-Vector system

$$\Big(\mathbf{B} + \mathbf{M} \Big) \mathbf{u} = \mathbf{l} \tag{1.26}$$

where as before $[\mathbf{B}]_{T(n,i),T(m,j)} = b_h(\Phi_j^m, \Phi_i^n)$, $[\mathbf{u}]_{T(m,j)} = \alpha_j^m$, $[\mathbf{l}]_{T(n,i)} = \ell(\Phi_i^n)$ and furthermore $[\mathbf{M}]_{T(n,i),T(m,j)} = (\Phi_j^m, \Phi_i^n)_{L^2(\Omega)}$.

Next we considered the same setting as described in subsection 1.9.1 and tested the convergence rates for $\mathcal{P}^1, \mathcal{P}^2$-elements, where we chose the exact solution $u(x) = e^{-x}\sin(x)$ and $c(x) = \sin(10x) + 2$.

Method A: "Error Quad"

This method corresponds to exactly doing what we have been doing before. We assembled the matrices of the system (1.26) as described in section 1.7 and approximated the integrals appearing in the entries of $\mathbf{A}$, $\mathbf{M}$, $\mathbf{l}$ using the $r+1$ node Gauss-Lobatto quadrature rule, where $r \in \{1, 2\}$, meaning the quadrature nodes coincide with the basis nodes. We calculated the $L^2, H^1(\mathcal{T}_h)$-error between the numerical solution and the exact solution on all meshes and called it `L2 quad`, `H1 quad` respectively.

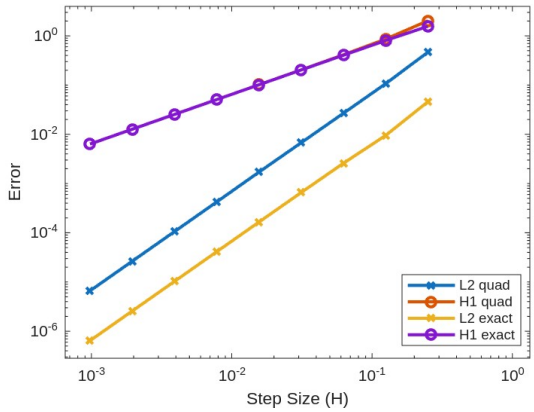

Figure 1.7: Comparison of convergence rates of higher order vs lower order quadrature for $\mathcal{P}^1$-elements

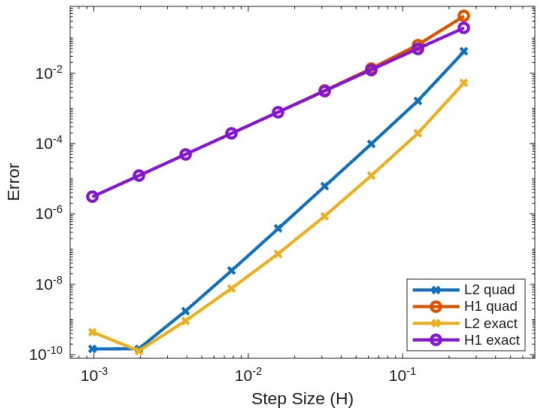

Figure 1.8: Comparison of convergence rates of higher order vs lower order quadrature for $\mathcal{P}^2$-elements

Method B: "Error Exact"

For this method we assembled the matrices $\mathbf{A}$, $\mathbf{M}$, $\mathbf{l}$ for the system 1.26 using $r+2$ Gauss-Lobatto quadrature nodes. This quadrature is exact for polynomials of order $2r+1$, meaning the entries of the mass matrix $\mathbf{M}$ are calculated exactly. Next we solved the system yielding the numerical solution and in turn calculated the $L^2, H^1(\mathcal{T}_h)$-errors on all meshes and called it `L2 exact`, `H1 exact` respectively.

In the figures 1.7, 1.8 we compared the error rates of Method A and B and observed no significant difference but a slight downward shift of `L2 exact` on a logarithmic scale in comparison to `L2 quad`.

Appendix A

Prerequisites

Bibliography

[1] L. C. Evans, *Partial Differential Equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, 2 ed., 2010.

[2] E. H. Georgoulis, *Discontinuous galerkin methods for linear problems: An introduction*, in Approximation Algorithms for Complex Systems, E. Georgoulis, A. Iske, and J. Levesley, eds., vol. 3 of Springer Proc. Math., Springer, Berlin, Heidelberg, 2011, pp. 91–126.

[3] M. Grote, A. Schneebeli, and D. Schötznau, *Discontinuous method for the wave equation*, SIAM Journal on Numerical Analysis, 44 (2006), pp. 2408–2431.

[4] D. A. D. Pietro and A. Ern, *Mathematical Aspects of Discontinuous Galerkin Methods*, Springer, Berlin, Heidelberg, 1 ed., 2012.

[5] A. Quarteronia, R. Sacco, and F. Saleri, *Numerical Mathematics*, vol. 2, Springer, Berlin, Heidelberg, 2007.

[6] B. Rivière, *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation*, vol. 35 of Frontiers in Applied Mathematics, SIAM, Philadelphia, 2008.

[7] T. Warburton and J. S. Hesthaven, *On the constants in hp-finite element trace inverse inequalities*, Computer Methods in Applied Mechanics and Engineering, 192 (2003), pp. 2765–2773.