MATHEMATICAL THEORY OF FINITE ELEMENTS

Institute for Computational Engineering and Sciences
The University of Texas at Austin
Austin, 2014
# Contents

1 Preliminaries

Variational Formulations

1.1 Classical Calculus of Variations ........................................... 1
1.2 Abstract Variational Formulation ........................................ 5
1.3 Classical Variational Formulations ..................................... 9
  1.3.1 Diffusion-Convection-Reaction Problem .......................... 9
  1.3.2 Linear Elasticity ........................................................ 13
1.4 Variational Formulations For First Order Systems ................. 15
  1.4.1 Linear Acoustics Equations ....................................... 15
  1.4.2 Maxwell Equations ................................................... 21
  1.4.3 Maxwell Equations. A Deeper Look .............................. 24

Coercive Problems

1.5 Minimization Principle and the Ritz Method .......................... 26
1.6 Lax-Milgram Theorem and Cea’s Lemma ................................ 29
1.7 Examples of Problems Fitting the Lax-Milgram Theory ............. 31
  1.7.1 A General Diffusion-Convection-Reaction Problem .............. 31
  1.7.2 Linear Elasticity ..................................................... 32
  1.7.3 Model Curl-Curl Problem ......................................... 33

Conforming Finite Elements and the Exact Sequence

1.8 Conforming Finite Elements ............................................. 35
  1.8.1 Classical $H^1$-Conforming Elements ............................ 35
  1.8.2 $H$ (curl)- and $H$ (div)-Conforming Elements. Exact Sequence ..................................................... 38
  1.8.3 Projection Based Interpolation ................................... 44
  1.8.4 Parametric Elements and Piola Transforms (Pullback Maps) .............. 47
1.9 Classical Interpolation Theory ......................................... 56
1

Preliminaries

Variational Formulations

1.1 Classical Calculus of Variations

See the book by Gelfand and Fomin [21] for a superb exposition of the subject.

The classical calculus of variations is concerned with the solution of the constrained minimization problem:

\[
\begin{cases}
\text{Find } u(x), x \in [a, b], \text{ such that:} \\
u(a) = u_a \\
u = \arg \min_{w(a)=u_a} J(w)
\end{cases}
\]  

(1.1)

where the cost functional \( J(w) \) is given by,

\[
J(w) = \int_a^b F(x, w(x), w'(x)) \, dx.
\]  

(1.2)

Integrand \( F(x, u, u') \) may represent an arbitrary scalar-valued function of three arguments*: \( x, u, u' \). Boundary condition (BC): \( u(a) = u_a \), with \( u_a \) given, is known as the essential BC.

In the following discussion we sweep all regularity considerations under the carpet. In other words, we assume whatever is necessary to make sense of the considered integrals and derivatives.

Assume now that \( u(x) \) is a solution to problem (1.1). Let \( v(x), x \in [a, b] \) be an arbitrary test function. Function

\[
w(x) = u(x) + \epsilon v(x)
\]

satisfies the essential BC if and only if (iff) \( v(a) = 0 \), i.e. the test function must satisfy the homogeneous essential BC. Consider an auxiliary function,

\[
f(\epsilon) := J(u + \epsilon v)
\]

*Note that, in this classical notation, \( x, u, u' \) stand for the arguments of the integrand. We could have used any other three symbols, e.g. \( x, y, z \).
If functional $J(w)$ attains a minimum at $u$ then function $f(\epsilon)$ must attain a minimum at $\epsilon = 0$ and, consequently,

$$\frac{df}{d\epsilon}(0) = 0$$

It remains to compute the derivative of function,

$$f(\epsilon) = J(u + \epsilon v) = \int_a^b F(x, u(x) + \epsilon v(x), u'(x) + \epsilon v'(x)) \, dx$$

By Leibnitz formula (see [22], p.17),

$$\frac{df}{d\epsilon}(\epsilon) = \int_a^b \frac{d}{d\epsilon} F(x, u(x) + \epsilon v(x), u'(x) + \epsilon v'(x)) \, dx$$

so, utilizing the chain formula, we get,

$$\frac{df}{d\epsilon}(\epsilon) = \int_a^b \left\{ \frac{\partial F}{\partial u}(u(x), u'(x) + \epsilon v'(x)) v(x) + \frac{\partial F}{\partial u'}(u(x), u'(x) + \epsilon v'(x)) v'(x) \right\} \, dx$$

Setting $\epsilon = 0$, we get,

$$\frac{df}{d\epsilon}(0) = \int_a^b \left\{ \frac{\partial F}{\partial u}(u(x), u'(x)) v(x) + \frac{\partial F}{\partial u'}(u(x), u'(x)) v'(x) \right\} \, dx \quad (1.3)$$

Again, remember that $u, u'$ in $\partial F/\partial u, \partial F/\partial u'$ denote simply the second and third arguments of $F$. Derivative (1.3) is identified as the *directional derivative* of functional $J(w)$ in the direction of test function $v(x)$.

The linear operator,

$$v \rightarrow \langle (\partial J)(u), v \rangle := \frac{df}{d\epsilon}(0) = \int_a^b \left\{ \frac{\partial F}{\partial u}(u(x), u'(x)) v(x) + \frac{\partial F}{\partial u'}(u(x), u'(x)) v'(x) \right\} \, dx \quad (1.4)$$

is identified as the *Gâteaux differential* of $J(w)$ at $u$.

The necessary condition for $u$ to be a minimizer reads now as follows,

$$\begin{cases}
  u(a) = u_a \\
  \langle (\partial J)(u), v \rangle = \int_a^b \left( \frac{\partial F}{\partial u}(x, u, u') v + \frac{\partial F}{\partial u'}(x, u, u') v' \right) \, dx = 0, \quad \forall v : v(a) = 0.
\end{cases} \quad (1.5)$$

Integral identity (1.5) that has to be satisfied for any eligible test function $v$, is identified as the *variational formulation* corresponding to the minimization problem.

In turns out that the variational formulation is equivalent to the corresponding *Euler-Lagrange (E-L) differential equation* and an additional *natural BC* at $x = b$. The key tool to derive both of them is the following Fourier’s lemma.

**LEMMA 1.1.1**

*(Fourier)*

Let $f \in C[a, b]$ such that

$$\int_a^b f(x)v(x) \, dx = 0$$
for every test function \( v \in C[a,b] \) such that \( v(a) = v(b) = 0 \). Then \( f(x) = 0, x \in [a,b] \).

**Proof**  See [28], p.531.

In order to apply Fourier’s argument, we need first to move the derivative from the test function in the second term in (1.5). We get,

\[
\int_a^b \left( \frac{\partial F}{\partial u}(x,u,u') - \frac{d}{dx} \frac{\partial F}{\partial u'}(x,u,u') \right) v \, dx + \frac{\partial F}{\partial u'}(x,u(x),u'(x))v(x)|_a^b = 0
\]

But \( v(a) = 0 \) so the boundary terms reduce only to the term at \( x = b \) (we do not test at \( x = a \)),

\[
\int_a^b \left( \frac{\partial F}{\partial u}(x,u,u') - \frac{d}{dx} \frac{\partial F}{\partial u'}(x,u,u') \right) v \, dx + \frac{\partial F}{\partial u'}(b,u(b),u'(b))v(b) = 0 \tag{1.6}
\]

We can follow now with the Fourier argument.

**Step 1:** Assume additionally that we test only with test functions that vanish both at \( x = a \) and \( x = b \). The boundary term in (1.6) disappears and, by Fourier’s lemma, we can conclude that

\[
\frac{\partial F}{\partial u}(x,u(x),u'(x)) - \frac{d}{dx} \frac{\partial F}{\partial u'}(x,u(x),u'(x)) = 0 \tag{1.7}
\]

We say that we have recovered the differential equation.

**Step 2:** Once we know that the function above vanishes, the integral term in (1.6) must vanish for any test function \( v \). Consequently,

\[
\frac{\partial F}{\partial u'}(b,u(b),u'(b))v(b) = 0,
\]

for any \( v \). Choose such a test function that \( v(b) = 1 \) to learn that the solution must satisfy the natural BC at \( x = b \),

\[
\frac{\partial F}{\partial u'}(b,u(b),u'(b)) = 0. \tag{1.8}
\]

We have recovered the natural BC. The Euler-Lagrange equation (1.7) along with the essential and natural BCs constitute the Euler-Lagrange Boundary-Value Problem (E-L BVP),

\[
\begin{cases}
\frac{\partial F}{\partial u}(x,u,u') - \frac{d}{dx} \left( \frac{\partial F}{\partial u'}(x,u,u') \right) = 0 \quad \text{(Euler-Lagrange equation)} \\
u(a) = u_a \quad \text{(essential BC)} \\
\frac{\partial F}{\partial u'}(b,u(b),u'(b)) = 0 \quad \text{(natural BC)}.
\end{cases} \tag{1.9}
\]

Neglecting the regularity issues, we can say that the E-L BVP and variational formulations are in fact equivalent to each other. Indeed, we have already shown that the variational formulation implies the E-L BVP. To show the converse, we multiply the E-L equation with a test function \( v(x) \), integrate it over interval \( (a,b) \) and add to it the natural BC multiplied by \( v(b) \). We then integrate (back) by parts, to arrive at the variational formulation. We say that the variational formulation and the E-L BVP are formally equivalent, formally meaning w/o paying attention to regularity assumptions.
The E-L BVP provides a foundation for Finite Difference (FD) discretizations, whereas the variational formulation is a starting point for the Galerkin method and Finite Elements.

**Exercises**

**Exercise 1.1.1** Consider a slight variation of Fourier’s lemma:

**LEMMA 1.1.2**

Let \( f \in C[a,b] \) such that

\[
\int_a^b f(x)v(x) \, dx = 0
\]

for every test function \( v \in C[a,b] \). Then \( f(x) = 0, \, x \in [a,b] \).

Which of the two lemmas: Lemma 1.1.1 or the lemma above is stronger? Prove the lemma above (one line argument!).

(1 point)

**Exercise 1.1.2** Derive the variational formulation and the corresponding Euler-Lagrange boundary-value problem for the minimization problem:

\[
\begin{aligned}
&\begin{cases}
  u(a) = u_a, \, u'(a) = d_a \\
  \int_a^b F(x,u,u',u'') \, dx \to \min
\end{cases}
\end{aligned}
\]

Discuss other possible essential BC. (3 points)

**Exercise 1.1.3** Derive the variational formulation and the corresponding Euler-Lagrange boundary-value problem for the two-dimensional minimization problem:

\[
\begin{aligned}
&\begin{cases}
  u = u_0 \text{ on } \Gamma_1 \\
  \int_\Omega F(x,y,u(x,y), \frac{\partial u}{\partial x}(x,y), \frac{\partial u}{\partial y}(x,y)) \, dxdy \to \min
\end{cases}
\end{aligned}
\]

Here \( \Omega \subset \mathbb{R}^2 \) is a bounded two-dimensional domain with boundary \( \Gamma \) split into two disjoint parts, \( \Gamma = \Gamma_1 \cup \Gamma_2 \). (3 points)

**Exercise 1.1.4** (An interface problem) Consider the elastic beam depicted in Fig. 1.1. Deflection \( w(x) \) of the beam minimizes the total potential energy given by the functional

\[
J(w) = \frac{1}{2} \int_0^{3l/2} EI(w'')^2 - \left[ \int_0^{3l/2} qw + P_0 w(\frac{3l}{2}) + M_0 w'(\frac{3l}{2}) \right]
\]
among all possible displacements that satisfy the *kinematic BC*:

\[ w(0) = w'(0) = w(l) = 0 \]

- Derive the Gâteaux derivative of cost functional \( J(w) \) and the corresponding variational formulation for the problem.
- Use integration by parts (twice) and the Fourier’s Lemma argument to derive the corresponding E-L equation(s) in subintervals \((0, l)\) and \((l, 3l/2)\), boundary conditions at \( x = 3l/2 \) and interface conditions at \( x = l \).
- Show the (formal) equivalence between the variational formulation and the E-L interface boundary-value problem.

![An elastic beam example](image)

**Figure 1.1**  
An elastic beam example  

(3 points)

### 1.2 Abstract Variational Formulation

We begin our study on Galerkin and FE methods with the *abstract variational formulation*.

**Abstract variational formulation** reads as follows.

\[
\begin{cases}
  u \in U \\
  b(u, v) = l(v) \quad \forall v \in V .
\end{cases}
\]

Here \( U \) is a *trial* space, and \( V \) is a *test space*. In this monograph, we shall restrict to Hilbert spaces only. The two spaces come with inner products and the corresponding (Euclidean) norms,

\[
\|u\|^2_U = (u, u)_U, \quad \|v\|^2_V = (v, v)_V .
\]
On the left we have a bilinear (or sesquilinear) form $b : U \times V \to \mathbb{R}(U)$ defining the operator, and on the right, we have a linear (antilinear) form $l : V \to \mathbb{R}(U)$ specifying the load. It goes without saying that both forms must be continuous. It is easy to show (see Exercise 1.2.1 and Exercise 1.2.2) that the continuity of forms $b$ and $l$ is equivalent to their boundedness, i.e.,

$$|b(u, v)| \leq M\|u\|_U\|v\|_V \quad \forall u \in U, v \in V,$$

(1.11)

and,

$$|l(v)| \leq C\|v\|_V \quad \forall v \in V,$$

(1.12)

for some $M, C > 0$.

Make sure that, for each variational formulation discussed in the next section, you are able to specify energy spaces $U, V$, and the forms $b(u, v), l(v)$.

**Accounting for non-homogeneous BCs.** In the case of non-homogeneous essential BCs, we may have to consider a more general abstract variational problem:

$$\begin{cases}
    u \in \tilde{u}_0 + U \\
    b(u, v) = l(v) \quad \forall v \in V.
\end{cases}
$$

(1.13)

Here $U$ is a subspace of a larger energy space $X$, and $\tilde{u}_0$ is an element of $X$. Symbol $\tilde{u}_0 + U$ denotes the algebraic sum or $\tilde{u}_0$ and $U$ known also as an *affine subspace of affine submanifold of $X$*,

$$\tilde{u}_0 + U := \{ \tilde{u}_0 + w : w \in U \} .$$

In practice the non-homogeneous boundary data $u_0$ is known only on the boundary of the domain. The tilde over $u_0$ denotes a *finite energy lift of $u_0$*, i.e. an extension of $u_0$ to the whole domain that lives in the energy space $X$. In this discussion though, $\tilde{u}_0$ is simply an arbitrary element of $X$ that does not\(^1\) live in $U$. The moral of this abstract notation is that solution $u$ can be sought in the form $u = \tilde{u}_0 + w$ where $w \in U$. Substituting this representation of $u$ into the variational formulation, using linearity of form $b$ wrt the first argument, and moving known terms to the right-hand side, we obtain,

$$\begin{cases}
    w \in U \\
    b(w, v) = l(v) - b(\tilde{u}_0, v) \quad \forall v \in V.
\end{cases}$$

The case of non-homogeneous BCs can thus be studied within the framework of original formulation (1.10) provided we replace the linear form $l(v)$ with the *modified linear form $l^{\text{mod}}(v)$*. This explains also why the essential BC data $u_0$ is classified as part of the load.

\(^1\)Otherwise, $\tilde{u}_0 + U = U$. 
Galerkin approximation. It is not too early to introduce the fundamental concept of the Galerkin approximation of the abstract variational problem. We approximate solution \( u \) and test functions \( v \) with finite linear combinations:

\[
u \approx u_h := \sum_{j=1}^{N} u_j e_j, \quad v \approx v_h := \sum_{i=1}^{N} v_i g_i
\]

where trial basis functions \( e_j \in U \) live in the trial space, the test basis functions \( g_i \in V \) live in the test space, coefficients \( u_j \in \mathbb{R} \) (\( \mathbb{C} \)) are the unknown degrees-of-freedom (dof) to be determined, and coefficients \( v_i \) are arbitrary real (complex) numbers. Notice that we use the same number of terms in both approximating combinations (explain, why?). Symbol \( h \) here is a general, abstract discretization symbol. In context of finite elements, it may be interpreted as mesh size. We simply replace now \( u \) with \( u_h \) and \( v \) with \( v_h \) and request the resulting system to be satisfied for any test function coefficients \( v_i \). We end up with the following system of linear algebraic equations:

\[
\sum_{j=1}^{N} b(e_j, g_i) u_j = l(g_i) \quad i = 1, \ldots, N
\]

(1.15)

Vector \( l_i \) and matrix \( b_{ij} \) are known as load vector and stiffness matrix. The Galerkin method can now be summarized in three steps:

1. Select trial and test basis functions, and compute stiffness matrix and load vector.

2. Solve the resulted system of linear equations.

3. Compute the approximate solution (1.14) using the (now) known dof and postprocess it as necessary.

The collection of all \( u_h \) and \( v_h \) of form (1.14), for arbitrary dof \( u_j, v_i \) is identified as the finite dimensional trial space \( U_h \subset U \) and test space \( V_h \subset V \). The approximate problem can written thus in the more concise form:

\[
\begin{align*}
\{ & u_h \in U_h \\
& b(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h
\}
\end{align*}
\]

(1.16)

The difference \( e_h := u - u_h \) is identified as the Galerkin error. The main purpose of this monograph is to study the evolution (convergence) of the Galerkin error

\[
\|u - u_h\|_U \rightarrow 0 \quad \text{as} \quad h \rightarrow 0
\]

Stability of discretization. We shall say that the Galerkin method is stable if there exists a stability constant \( C > 0 \) such that

\[
\|u - u_h\|_U \leq C \inf_{u_h \in U_h} \|u - u_h\|_U
\]

\( := \) best approximation error (BAE)

If the method is stable, and the best approximation error converges to zero, then so does the error and it converges with the same rate as the BAE. We say then also that the discretization is optimal. Note that \( C \)
need not be independent of $h$. It it blows up with $h$, the BAE should converge faster to zero than $C_h \to \infty$ in order to see the (non-optimal) convergence.

Try to remember the phrase:

Approximability and stability imply convergence .

Exercises

**Exercise 1.2.1** Equivalence of continuity and boundedness for linear(antilinear) forms. Let $V$ be a normed vector space and $l$ be a linear (antilinear) functional defined on $V$. Prove that the following conditions are equivalent to each other. (3 points)

(i) $l$ is continuous on $V$,
(ii) $l$ is sequentially continuous on $V$,
(iii) $l$ is continuous at 0 (zero vector),
(iv) $l$ is sequentially continuous at 0,
(v) $l$ is bounded, i.e. there exists $C > 0$ such that

$$|l(v)| \leq C \|v\|_V$$

where $\|v\|_V$ is the norm in $V$.

**Exercise 1.2.2** Equivalence of continuity and boundedness for bilinear(sesquilinear) forms. Let $U, V$ be normed vector spaces and $b$ be a bilinear (sesquilinear) functional defined on $U \times V$. Prove that the following conditions are equivalent to each other. (3 points)

(i) $b$ is continuous on $U \times V$,
(ii) $b$ is sequentially continuous on $U \times V$,
(iii) $b$ is continuous at $(0, 0)$,
(iv) $b$ is sequentially continuous at $(0, 0)$,
(v) $b$ is bounded, i.e. there exists $C > 0$ such that

$$|b(u, v)| \leq C \|u\|_U \|v\|_V .$$

**Exercise 1.2.3** Dual norm. Let $V$ be a normed vector space and $l$ be a continuous (bounded) linear (antilinear) functional defined on $V$. Let $\|l\|$ be the “smallest” constant that we can use in the boundedness condition,

$$\|l\| := \inf \{C : |l(v)| \leq C \|v\|_V\}$$
1. Prove equivalent characterizations for \( \|l\| \),

\[
\|l\| = \sup_{v \neq 0} \frac{|l(v)|}{\|v\|} = \sup_{\|v\|=1} |l(v)|
\]

2. Let \( V' \) be the collection of all bounded linear (antilinear) functionals defined on \( V \). Argue that \( V' \) is close wrt the standard operations on functions and, therefore, constitutes a subspace of algebraic dual \( V^* \) consisting of all linear (antilinear) functionals on \( V \). Prove that \( \|l\| \) satisfies the axioms for a norm, i.e \( V' \) is a normed space (called the topological dual of space normed space \( V \)).

(3 points)

**Exercise 1.2.4** Let \( V \) be a Hilbert space. Prove that the infimum and the suprema in Exercise 1.2.5 are actually attained, i.e. the inf and sup symbols can be replaced with min and max. (3 points)

**Exercise 1.2.5** Space of bounded bilinear functionals. Generalize the concept of the norm of a linear functional to the bilinear (sesquilinear) functionals. Let \( U, V \) be normed vector spaces and \( b \) be a continuous (bounded) bilinear (sesquilinear) functional defined on \( U \times V \). Let \( \|b\| \) denote the “smallest” constant that we can use in the boundedness condition,

\[
\|b\| := \inf \{ C : |b(u, v)| \leq C \|u\|_U \|v\|_V \}
\]

1. Prove equivalent characterizations for \( \|b\| \),

\[
\|b\| = \sup_{u,v \neq 0} \frac{|b(u, v)|}{\|u\|_U \|v\|_V} = \sup_{\|u\|_U = \|v\|_V = 1} |b(u, v)|
\]

2. Prove that the collection of all bounded bilinear (sesquilinear) functionals defined on \( U \times V \) forms a normed space with norm \( \|b\| \). Show that the infimum and the suprema above are attained if \( u, V \) are Hilbert.

(3 points)

### 1.3 Classical Variational Formulations

#### 1.3.1 Diffusion-Convection-Reaction Problem

Let \( \Omega \in \mathbb{R}^N, N = 1, 2, 3 \) be a bounded domain (:= open, connected set). Let boundary \( \Gamma = \partial \Omega \) be split in two disjoint parts \( \Gamma_1, \Gamma_2 \). More precisely, \( \Gamma_1, \Gamma_2 \) are assumed to be (relatively) open in \( \Gamma \) and

\[
\Gamma = \Gamma_1 \cup \Gamma_2, \quad \Gamma_1 \cap \Gamma_2 = \emptyset
\]
where the overbar denotes the closure in $\Gamma$.

Consider a general diffusion-convection-reaction boundary-value (BV) problem,

\[
\begin{align*}
\text{Find } u = u(x), & \quad x \in \Omega, \text{ such that:} \\
-(a_{ij}u_{,j})_{,i} + b_ju_{,j} + cu = f & \text{ in } \Omega \\
u = u_0 & \text{ on } \Gamma_1 \\
a_{ij}u_{,j}n_i = g & \text{ on } \Gamma_2
\end{align*}
\] (1.17)

Here $a_{ij}(x), b_j(x), c(x), x \in \Omega$ are diffusion, convection and reaction coefficients (material data), and functions $f(x), x \in \Omega, u_0(x), x \in \Gamma_1, g(x), x \in \Gamma_2$ are part of the load data, all assumed to be given. We are using the Einstein summation convention.

**Elementary integration by parts formula.** The following formula generalizes the classical 1D integration by parts to multispace dimension and it is the workhorse for deriving all variational formulations.

\[
\int_{\Omega} \frac{\partial u}{\partial x_i} v = -\int_{\Omega} u \frac{\partial v}{\partial x_i} + \int_{\partial \Omega} uv n_i
\] (1.18)

where $\Omega \subset \mathbb{R}^N, N = 2, 3$, and $n_i$ is the $i$-th component of the outward normal unit vector $n$. For $N = 2$, the domain integral is a double integral, and the boundary integral is the line integral of the first type. For $N = 3$, we are dealing with a triple integral and the surface integral of the first type. The line and surface integrals, and the formula, can be generalized to any $N$ dimension.

The elementary integration by parts formula can be used to derive more complicated integration by parts formulas for different differential operators. The most classical ones involve operators of gradient, curl and divergence.

\[
\int_{\Omega} \text{div } u q = -\int_{\Omega} u \nabla q + \int_{\Gamma} u_n q
\]

where $u_n := u_i n_i$ denotes the normal component of vector $u$. Similarly,

\[
\int_{\Omega} \nabla \times E F = \int_{\Omega} E \nabla \times F + \int_{\Gamma} n \times EF.
\]

Note that we do not use boldface for vectors (and tensors) and you have to deduce from context what type of functions are we dealing with and wheter we mean product of two numbers, scalar product of two vectors or contraction of two tensors. Talking about tensors, we have the formula:

\[
\int_{\Omega} \text{div } \sigma v = -\int_{\Omega} \sigma \nabla v = \int_{\Gamma} \sigma n v.
\]

If $\sigma$ is the stress tensor then $t := \sigma n$ is the traction vector.

**Classical variational formulation.** We take an arbitrary test function $v = v(x), x \in \Omega$, multiply PDE (1.17) with $v(x)$, integrate over $\Omega$, and integrate the first term by parts using the elementary integration by parts formula, to obtain:

\[
\int_{\Omega} a_{ij}u_{,j}v_{,i} + b_ju_{,j}v + cuv - \int_{\Gamma} a_{ij}u_{,j}n_i v = \int_{\Omega} f v
\]
We can split now the boundary integral into two parts corresponding to $\Gamma_1$ and $\Gamma_2$. On $\Gamma_2$ the co-normal derivative $a_{ij}u_{,j}n_i$ is known and we can replace it with the given load data $g$. On $\Gamma_1$, the derivative is unknown a-priori, and we eliminate this part of the boundary integral by assuming $v = 0$ on $\Gamma_1$. We simply do not test on $\Gamma_1$. This is also consistent with the concept of essential BC in the classical calculus of variations: test functions satisfy always the homogeneous version of the essential BC.

Contrary to the BC on $\Gamma_2$ which has been built in into the formulation, the first BC has to be simply rewritten. The classical formulation reads now as follows.

\[
\begin{aligned}
\text{Find } u = u(x), & \quad x \in \Omega, \text{ such that:} \\
\int_{\Omega} a_{ij}u_{,j}v_{,i} + b_ju_{,j}v + cuv &= \int_{\Omega} fv + \int_{\Gamma_2} gv \\
& \quad \text{for all } v \text{ such that } v = 0 \text{ on } \Gamma_1
\end{aligned}
\]

(1.19)

**Regularity assumptions.** We have now to start paying attention to making appropriate assumptions to guarantee that all terms in the variational formulation are well defined. The first critical tool is the Cauchy-Schwarz inequality,

\[
|\int_{\Omega} uv| \leq \left(\int_{\Omega} |u|^2\right)^{1/2} \left(\int_{\Omega} |v|^2\right)^{1/2}
\]

(1.20)

where

\[
||u|| := \left(\int_{\Omega} |u|^2\right)^{1/2}
\]

is identified as the $L^2$-norm of function $u$. The $L^2$ space will be denoted by $L^2(\Omega)$ and the symbol for the space will be omitted in the symbol for the norm, i.e.

\[
||u|| = ||u||_{L^2(\Omega)}
\]

Recall that the $L^2$-space is a Hilbert space with the inner product,

\[
(u, v)_{L^2(\Omega)} := \int_{\Omega} uv, \quad ||u||^2 = (u, u)
\]

In the discussed case, all functions are real-valued so the complex conjugate over function $v$ is redundant. We shall skip the space symbol in the inner product notation as well.

If we assume now that the reaction coefficient is bounded,

\[
|c(x)| \leq c_{\max} < \infty, \quad x \in \Omega,
\]

and functions $u, v \in L^2(\Omega)$, Cauchy-Schwarz inequality implies that the integral corresponding to the reaction term is bounded as well. Indeed,

\[
|\int_{\Omega} c(x)uv| \leq \int_{\Omega} |c(x)||u||v| \leq c_{\max} \int_{\Omega} |u||v| \leq c_{\max}||u||||v||.
\]

By the same argument, if we assume that diffusion matrix $a_{ij}$ and the advection vector $b_j$ are bounded,

\[
||a(x)|| \leq a_{\max}, \quad ||b(x)|| \leq b_{\max}
\]
we can bound the first two terms on the left-hand side as well,

\[ | \int_{\Omega} a_{ij} u_{i,j} v_j | \leq a_{\text{max}} \left( \sum_i \| u_{i,j} \|^2 \right)^{1/2} \left( \sum_j \| v_{j} \|^2 \right)^{1/2} \]

\[ | \int_{\Omega} b_j u_j v | \leq b_{\text{max}} \left( \sum_i \| u_{i,j} \|^2 \right)^{1/2} \| v \| \]

Notice that by \( \| b \| \) we mean the norm of the vector,

\[ \| b \| = \left( \sum_i | b_i |^2 \right)^{1/2} \]

and by \( \| a \| \) the norm of a matrix. Typically, we assume that the diffusion matrix is symmetric. The norm of \( a \) is then,

\[ \| a \| = \max_j | \lambda_j | \]

where \( \lambda_j \) are the (real-valued) eigenvalues of \( a \). If \( a \) is not assumed to be symmetric then the norm of \( a \) is equal to the maximum value of singular value of \( a \).

These considerations lead to the introduction of our first energy space - the Sobolev space of the first order,

\[ H^1(\Omega) := \{ u \in L^2(\Omega) : \nabla u \in L^2(\Omega) \} \quad (1.21) \]

This is a Hilbert space with inner product,

\[ (u, v)_{H^1(\Omega)} = (u, v) + \sum_i (u_{i}, v_{i}) \]

and the norm,

\[ \| u \|^2_{H^1(\Omega)} := \| u \|^2 + \sum_i \| u_{i} \|^2 \]

Summing up, we can claim the estimate:

\[ | \int_{\Omega} a_{ij} u_{i,j} v_j + b_j u_j v + c u v | \leq (a_{\text{max}} + b_{\text{max}} + c_{\text{max}}) \| u \|_{H^1(\Omega)} \| v \|_{H^1(\Omega)} \quad (1.22) \]

Proceeding along similar lines, we can also estimate the right-hand side,

\[ | \int_{\Omega} f v + \int_{\Gamma_2} g v | \leq \| f \| \| v \| + \| g \|_{L^2(\Gamma_2)} \| v \|_{L^2(\Gamma_2)} \]

with the implicit assumption that \( \| f \|, \| g \|_{L^2(\Gamma_2)} \) are bounded. It follows from the famous Trace Theorem (to be discussed later) that there exists a positive constant \( C > 0 \) such that

\[ \| v \|_{L^2(\Gamma_2)} \leq C \| v \|_{H^1(\Omega)} \]

This leads to our final estimate of the right-hand side,

\[ | \int_{\Omega} f v + \int_{\Gamma_2} g v | \leq (\| f \|^2 + C^2 \| g \|_{L^2(\Gamma_2)})^{1/2} \| v \|_{H^1(\Omega)} \quad (1.23) \]
1.3.2 Linear Elasticity.

**Unknowns:** displacement: $u_i$, strains $\epsilon_{ij}$, stresses $\sigma_{ij}$, $i,j = 1, \ldots, N$.

**Equations:**

- strain-displacement relations: 
  $$\epsilon_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i}),$$

- equilibrium (conservation of linear momentum) equations: 
  $$-\sigma_{ij,j} = f_i,$$

- conservation of angular momentum: 
  $$\sigma_{ij} = \sigma_{ji},$$

- constitutive equations: 
  $$\sigma_{ij} = E_{ijkl} \epsilon_{kl} \quad \text{or} \quad \epsilon_{ij} = C_{ijkl} \sigma_{kl},$$

where elasticities satisfy the following conditions:

- Minor symmetries: 
  $$E_{ijkl} = E_{jikl} = E_{ijlk} \quad (\text{minor symmetries})$$

- Major symmetry: 
  $$E_{ijkl} = E_{klij} \quad (\text{major symmetry})$$

- Positive definiteness: 
  $$E_{ijkl} \xi_{ij} \xi_{kl} > 0 \quad \forall \xi_{ij} = \xi_{ji} \neq 0 \quad (\text{positive definiteness}).$$

**Cauchy stress vector - stress tensor relation:**

$$t_i = \sigma_{ij} n_j.$$

**Standard boundary conditions (BC):**

- Displacement BC: 
  $$u_i = 0$$

- Traction BC: 
  $$t_i = \sigma_{ij} n_j = g_i$$

**Lamé equations.**

$$-(E_{ijkl} u_{k,l})_j = f_i$$

**Classical variational formulation: Principle of Virtual Work**

$$\begin{cases} 
  u_i \in H^1(\Omega), u_i = 0 \text{ on } \Gamma_1 \\
  \int_{\Omega} E_{ijkl} u_{k,l} v_{i,j} = \int_{\Omega} f_i v_i + \int_{\Gamma_2} g_i v_i \quad \forall v_i \in H^1(\Omega) : v_i = 0 \text{ on } \Gamma_1 
\end{cases}$$

**Exercises**
Exercise 1.3.1 (Calculus II refresher) Define line and surface integrals of the first type. Discuss why they are identified as geometrical quantities. Consider a unit circle with center at origin, and density $\rho(x) = |x_2|$. Compute the mass of the circle. Similarly, consider a unit sphere centered at the origin, with density $\rho(x) = |x_3|$. Compute its mass. (2 points)

Exercise 1.3.2 Use whatever source you need to prove the elementary integration by parts formula (1.18) in both two and three space dimensions. (2 points)

Exercise 1.3.3 Follow the discussion for the diffusion-convection-reaction problem to prove continuity of bilinear and linear forms corresponding to the classical formulation (Principle of Virtual Work) for linear elastostatics. (3 points)

Exercise 1.3.4 Hooke’s law. For an isotropic material, the elasticities tensor depends only upon two material (Lamé) constants,

$$E_{ijkl} = \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) + \lambda\delta_{ij}\delta_{kl},$$

and the constitutive equations reduce to the Hooke’s law:

$$\sigma_{ij} = 2\mu\epsilon_{ij} + \lambda\delta_{ij}\epsilon_{kk}.$$

Derive the corresponding formula for the compliance tensor $C_{ijkl}$ and express strains in terms of stresses. (1 point)

Exercise 1.3.5 Specialize the Lamé equations and the corresponding Principle of Virtual Work to the case of an isotropic (but not necessarily homogeneous) material. (2 points)

Exercise 1.3.6 Derive the Principle of Virtual Work for the case of more general BC:

$$u_t = 0 \quad t_n = g_n \quad \text{or} \quad u_n = 0 \quad t_t = g_t$$

where $u_t, u_n$ denote tangential and normal components of vector $u$:

$$u_n = (u_k n_k) n_k, \quad u_t = u - u_n.$$

Use the Fourier’s lemma argument to show formally the equivalence of classical and variational formulations. (3 points)

Exercise 1.3.7 The Principle of Virtual Work involves summation in test functions $v_i$. Argue that the variational formulation is equivalent to a system of three variational identities where we test with just one component $v_i$ at the time. Summing those three variational identities looks arbitrary until you consider more general BC like those in Exercise 1.3.6. (2 points)
1.4 Variational Formulations For First Order Systems

In this section we discuss two new model problems: linear acoustics and Maxwell equations formulated as systems of first order equations. As we will see, starting with the first order system, we open up the possibility of multiple variational formulations for the same problem. It becomes also clear which of the equations are relaxed and which are not. We begin also to use the simplified notation for the domain and boundary integrals replacing them with more compact $L^2(\Omega)$ and $L^2(\Gamma)$ symbols,

$$(u, v) = \int_{\Omega} u \bar{v}, \quad \langle u, v \rangle := \int_{\Gamma} u \bar{v}.$$  

If there is need to indicate a more specific domain of integration, we enhance the brackets with an additional symbol, e.g.,

$$(u, v)_K = \int_{K} u \bar{v}, \quad \langle u, v \rangle_{\Gamma_1} := \int_{\Gamma_1} u \bar{v}.$$  

In the case of complex-valued problems, our default choice will be to complex-conjugate test functions, leading to the formalism of antilinear and sesquilinear forms. It goes without saying that, in case of vector- or tensor-valued functions, we use the proper dot products in place of the standard product of two numbers. In the next section, we will revisit the diffusion-convection-reaction and elastodynamics problems reformulated as first order systems as well.

1.4.1 Linear Acoustics Equations

The classical linear acoustics equations are obtained by linearizing the isentropic form of the compressible Euler equations expressed in terms of density $\rho$ and velocity vector $u_i$, around the hydrostatic equilibrium position $\rho = \rho_0, u_i = 0$. Perturbing the solution around the equilibrium position,

$$\rho = \rho_0 + \delta \rho, \quad u_i = 0 + \delta u_i,$$

and linearizing the Euler equations, see e.g. [25], we obtain a system of $N + 1$ first order equations in terms of unknown perturbations of density $\delta \rho$ and velocity $\delta u_i$, 

$$
\begin{align*}
(\delta \rho)_{,t} + \rho_0 (\delta u_j)_{,j} &= 0 \\
\rho_0 (\delta u_i)_{,t} + (\delta p)_{,i} &= 0,
\end{align*}
$$

with $\delta p$ denoting the perturbation in pressure. For the isentropic\(^\dagger\) flow, the pressure is simply an algebraic function of density,

$$p = p(\rho).$$

\(^\dagger\)The entropy is assumed to be constant throughout the whole domain.
Linearization around the equilibrium position leads to the relation between the perturbation in density and the corresponding perturbation in pressure,

\[
p = p(\rho_0) + \frac{dp}{d\rho}(\rho_0)\delta\rho .
\]

Here \( p_0 \) is the hydrostatic pressure, and the derivative \( \frac{dp}{d\rho}(\rho_0) \) is interpreted \textit{a posteriori} as the sound speed squared, and denoted by \( c^2 \). Consequently, the perturbation in pressure and density are related by the simple linear equation,

\[
\delta p = c^2 \delta \rho .
\]

It is customary to express the equations of linear acoustics in pressure rather than density. Dropping deltas in the notation, we obtain,

\[
\frac{\hat{c}}{\rho_0} p_{,t} + \rho_0 u_{j,j} = 0
\]

\[
\rho_0 i\omega \hat{u}_i + p_{,i} = 0.
\]

**Time-harmonic equations.** Assuming ansatz,

\[
p(t, x) = e^{i\omega t} p(x), \quad u_i(t, x) = e^{i\omega t} u_i(x),
\]

we reduce the acoustics equations to,

\[
\frac{\hat{c}}{\rho_0} i\omega p + \rho_0 u_{j,j} = 0
\]

\[
\rho_0 i\omega \hat{u}_i + \nabla \hat{p} = 0.
\]

**Non-dimensionalization.** Choosing reference length \( l_0 \), pressure \( p_0 \), velocity (speed) \( u_0 \), and angular frequency \( \omega_0 \), we introduce non-dimensional coordinates \( \hat{x}_i \), pressure \( \hat{p} \), velocity \( \hat{u}_i \) and angular frequency \( \hat{\omega} \),

\[
\hat{x}_i = \frac{x_i}{l}, \quad \hat{p} = \frac{p}{p_0}, \quad \hat{u}_i = \frac{u_i}{u_0}, \quad \hat{\omega} = \frac{\omega}{\omega_0}.
\]

Substituting the formulas into the equations, we get:

\[
\frac{\omega_0 p_0}{\hat{c}} i\omega \hat{p} + \frac{\rho_0 u_0}{l} \nabla \hat{u} = 0
\]

\[
\rho_0 i\omega_0 \hat{u} + \frac{p_0}{l} \nabla \hat{p} = 0.
\]

Acoustics is a pure mechanical problem so we can choose only three independent scales (units), typically for mass (or force), length, and time (frequency in our case). For the unit of length \( l \) we can choose the size of domain. For instance, if we are solving our problem in a square domain (2D), after non-dimensionalization, this will be a \( \text{unit} \) square domain. Typically, we want the non-dimensional frequency \( \hat{\omega} \) to coincide with the non-dimensional wave number,

\[
k := \frac{\omega}{\hat{c}}.
\]
which leads to the choice of reference angular frequency, \( \omega_0 = c/l \). Finally, we want to minimize the number of coefficients in our equations. Setting the scaling factors in the first or second equations to be equal, we obtain the relation,

\[
p_0 = \rho_0 c u_0 .
\]

This means that we can choose \( p_0 \) with \( u_0 \) being derived from the equation above of, vice versa, choose \( u_0 \) and obtain \( p_0 \). Dropping the “hats”, we obtain the final non-dimensional equations in the form:

\[
\begin{align*}
\mathbf{i}\omega p + \text{div} u &= 0 \\
\mathbf{i}\omega u + \nabla p &= 0.
\end{align*}
\]

(1.24)

Mixed formulation I and reduction to a second order equation in terms of pressure. Eliminating the velocity, we obtain the Helmholtz equation for the pressure,

\[-\Delta p - \omega^2 p = 0 .\]

Having obtained the second order problem, we can proceed now with the derivation of the weak formulation, as discussed in the previous sections.

It is a little more illuminating to obtain the same variational formulation starting with the first order system. First of all, we make a clear choice in a way we treat the two equations. The equation of continuity (conservation of mass) is going to be satisfied only in the weak sense, i.e. we multiply it with a test function \( q \), integrate over domain \( \Omega \) and integrate the second term by parts to obtain,

\[
(i\omega p, q) - (u, \nabla q) + \langle u_n, q \rangle = 0 \quad \forall q
\]

where \( u_n = u \cdot n = u_j n_j \) denotes the normal component of the velocity on the boundary. At this point we introduce three different boundary conditions:

- a soft boundary \( \Gamma_p \):
  \[
p = p_0 ,
\]

- a hard boundary \( \Gamma_u \):
  \[
u_n = u_0 ,
\]

- and an impedance condition boundary \( \Gamma_i \):
  \[
u_n = dp + u_0 .
\]

where impedance constant \( d > 0 \).

We can now built-in the second and third BCs into the variational formulation to obtain

\[
(i\omega p, q) - (u, \nabla q) + (dp, q)_{\Gamma_i} = -\langle u_0, q \rangle_{\Gamma_u} \cup_{\Gamma_i} , \quad \forall q : q = 0 \text{ on } \Gamma_p .
\]
We say that we have relaxed the first equation. The second equation (conservation of momentum) is also be multiplied with a test function $v$ and integrated over domain $\Omega$ but we do not integrate it by parts,

$$(i\omega u, v) + (\nabla p, v) = 0 \quad \forall v.$$  

If the scalar product of an $L^2$-function $u$ with an arbitrary $L^2$ test function, vanishes,

$$(u, v) = 0 \quad \forall v \in L^2(\Omega),$$

substituting $v = u$, we conclude that $u$ must vanish almost everywhere,

$$\|u\|^2 = 0 \quad \Rightarrow \quad u = 0 \quad \text{a.e.}.$$  

Thus, except for the ‘‘a.e.” symbol nothing has changed, and the equation is still satisfied pointwise, i.e. in the strong sense.

The relaxed continuity equation and strong form of the conservation of momentum equations constitute our Mixed formulation I:

$$
\begin{aligned}
& p \in H^1(\Omega), \; p = p_0 \text{ on } \Gamma_p \\
& u \in L^2(\Omega) \\
& (i\omega p, q) - (u, \nabla q) + (dp, q)_{\Gamma'}, = -(u_0, q)_{\Gamma_i} \quad q \in H^1(\Omega), \; q = 0 \text{ on } \Gamma_p \\
& (i\omega u, v) + (\nabla p, v) = 0, \quad v \in L^2(\Omega).
\end{aligned}
$$

(1.25)

As in the previous section, choice of the energy spaces follows from the assumption on continuity (boundedness) of the sesquilinear form and Cauchy-Schwartz inequality. Pressure $p$ enters the formulation with gradient and, therefore, both $p$ and $\nabla p$ must be square integrable. This leads to the assumption that $p \in H^1(\Omega)$. Similarly, no derivatives of velocity $u$ are present in the formulation and, therefore, $u \in L^2(\Omega)$. It goes without saying that for vectors, we mean the $L^2$-space of vector valued functions. Equivalently, $u \in (L^2(\Omega))^N$, see Exercise 1.4.1. It turns out that, with this choice of energy spaces, all remaining contributions to the sesquilinear form are continuous as well.

In order to fit the formulation into the abstract framework discussed in Section 1.2, we need to introduce group variables,

$$
\begin{aligned}
u := (u, p), \quad \nu := (v, q).
\end{aligned}
$$

Test and trial spaces are identical,

$$
\begin{aligned}
U = V := \{(v, q) \in L^2(\Omega) \times H^1(\Omega) : v = 0 \text{ on } \Gamma_u\},
\end{aligned}
$$

and the antilinear and sesquilinear form are obtained by summing up right- and left sides of the formulation,

$$
\begin{aligned}
l(v) := - (u_0, q)_{\Gamma_i} \\
b(u, \nu) := (i\omega p, q) - (u, \nabla q) + (dp, q)_{\Gamma'}, + (i\omega u, v) + (\nabla p, v).
\end{aligned}
$$
The abstract formulation has the form:
\[
\begin{align*}
  u & \in \tilde{\mathbf{u}}_0 + \mathbf{U} \\
  b(u, v) &= l(v), \quad v \in V
\end{align*}
\]
where \( \tilde{\mathbf{u}}_0 = (0, \tilde{\mathbf{p}}_0) \in L^2(\Omega) \times H^1(\Omega) \) is a finite energy lift of the BC data.

Using the (strong) conservation of momentum equation, we can represent the velocity in terms of pressure,
\[
u = 1 i\omega \nabla p.
\]
(1.26)
In particular, the normal component of the velocity is related to the normal derivative of the pressure,
\[
u_n = 1 i\omega \frac{\partial p}{\partial n}.
\]
Multiplying (1.25) \(_1\) with \( i\omega \), and eliminating the velocity in the domain integral term using formula (1.26), we get the classical variational formulation of the Helmholtz equation. We can classify it as our **Reduced Formulation I**.

\[
\begin{align*}
  p \in H^1(\Omega), \quad p &= p_0 \text{ on } \Gamma_p, \\
  (\nabla p, \nabla q) - \omega^2(p, q) + i\omega \langle dp, q \rangle_{\Gamma_i} &= -\langle u_0, q \rangle_{\Gamma_i}, \quad q \in H^1(\Omega), \quad q = 0 \text{ on } \Gamma_p.
\end{align*}
\]
(1.27)
Note that we have obtained the weak formulation without introducing the second order problem at all. We have a clear understanding which of the starting equations is understood in the weak, and which in a strong sense. We mention only that all these considerations can be made more precise by introducing the language of distributions and Sobolev spaces.

**Mixed formulation II and reduction to a second order equation in terms of velocity.** Eliminating pressure from the first order system, we get the second order equation for the velocity,
\[-\nabla (\text{div } v) - \omega^2 u = 0.\]
As with the Helmholtz equation, we can proceed directly with the second order equation, to derive the corresponding variational formulation. But again, we prefer to work with the first order system. Keeping the conservation of mass equation in the strong form and relaxing the conservation of momentum, we obtain **Mixed Formulation II**.

\[
\begin{align*}
  u & \in H(\text{div}, \Omega), \quad u_n = u_0 \text{ on } \Gamma_u \\
  p & \in L^2(\Omega) \\
  i\omega(p, q) + (\text{div } u, q) &= 0, \quad q \in L^2(\Omega) \\
  i\omega(u, v) - (p, \text{div } v) + \langle d^{-1}u_n, v_n \rangle_{\Gamma_i} &= -\langle p_0, v_n \rangle_{\Gamma_p} + \langle d^{-1}u_0, v_n \rangle_{\Gamma_i}, \quad v \in H(\text{div}, \Omega), \quad v_n = 0 \text{ on } \Gamma_u.
\end{align*}
\]
(1.28)
If we use the first equation to eliminate the pressure, we arrive at the **Reduced Formulation II**.

\[
\begin{align*}
  u_n &= u_0 \text{ on } \Gamma_u \\
  (\text{div } u, \text{div } v) - \omega^2(u, v) + i\omega(d^{-1}u_n, v_n)_{\Gamma_i} &= -i\omega(p_0, v_n)_{\Gamma_p} + \langle d^{-1}u_0, v_n \rangle_{\Gamma_i}, \quad \forall v : v_n = 0 \text{ on } \Gamma_u.
\end{align*}
\]
(1.29)
Above we have arrived at a new energy space,

\[ H(\text{div}, \Omega) := \{ u \in L^2(\Omega) : \text{div} u \in L^2(\Omega) \} \quad (1.30) \]

where, similarly, to the definition of \( H^1(\Omega) \), divergence is understood in the sense of distributions. Note that we avoid using the names of Dirichlet or Neumann BCs. The condition on pressure (soft boundary) is a Dirichlet (essential) BC for the Reduced Formulation I but it becomes the Neumann BC in Reduced Formulation II. The same comment applies to the hard boundary BC.

There are two more variational formulation to go. Before we discuss them, it is convenient to introduce even more abstract notation useful for systems of first order equations. With the group variable \( u := (u, p) \) in place, we introduce the operator corresponding to strong formulation (1.24),

\[ Au := (i\omega p + \text{div} v, i\omega u + \nabla p) . \]

Consistently with the theory of closed operators \[28\], we specify the domain of the operator as,

\[ D(A) := \{ u \in L^2(\Omega) : Au \in L^2(\Omega), p = 0 \text{ on } \Gamma_p, u = 0 \text{ on } \Gamma_u, u_n = dp \text{ on } \Gamma_i \} . \]

By assumption thus, the operator takes values in \( L^2(\Omega) \). With the assumption that both \( p \) and \( u \) are \( L^2 \)-functions, assumption \( Au \in L^2(\Omega) \) is equivalent to conditions: \( \nabla p \in L^2(\Omega) \), \( \text{div} u \in L^2(\Omega) \). The domain of the operator can thus be written in a more concrete form:

\[ D(A) := \{ u = (u, p) \in H(\text{div}, \Omega) \times H^1(\Omega) : p = 0 \text{ on } \Gamma_p, u = 0 \text{ on } \Gamma_u, u_n = dp \text{ on } \Gamma_i \} . \]

The adjoint operator \( A^*v, v \in D(A^*) \) is defined as the operator that satisfies the equation:

\[ (Au, v) = (u, A^*v), \quad u \in D(A), v \in D(A^*) \]

where domain \( D(A^*) \) is the maximum set for which the equality holds. Integration by parts reveals that \( A \) is formally skew-adjoint, \( A^* = A \), with

\[ D(A^*) = \{ v = (v, q) \in H(\text{div}, \Omega) \times H^1(\Omega) : q = 0 \text{ on } \Gamma_p, v = 0 \text{ on } \Gamma_u, v_n = -dq \text{ on } \Gamma_i \} . \]

Note the change of sign in the impedance BC. This is the reason why the operator is only formally skew-adjoint.

**Strong (trivial) variational formulation.** Multiplying equations (1.24) with test functions and integrating over \( \Omega \), we obtain the **Strong (Trivial) Variational Formulation**:

\[
\begin{align*}
(u, p) & \in H(\text{div}, \Omega) \times H^1(\Omega) \\
p & = p_0 \text{ on } \Gamma_p \\
u_n & = u_0 \text{ on } \Gamma_u \\
p & = du_n + u_0 \text{ on } \Gamma_i \\
i\omega(p, q) + (\text{div} u, q) & = 0, \quad q \in L^2(\Omega) \\
i\omega(u, v) + (\nabla p, v) & = 0, \quad v \in L^2(\Omega) .
\end{align*}
\]
Using the formalism of closed operators, we can write it in a more compact form,
\[
\begin{align*}
\{ & u = \tilde{u}_0 + D(A) \\
&Au, v = 0, \quad v \in L^2(\Omega).
\end{align*}
\]
where, as usual, \(\tilde{u}_0\) is a lift of the BC data.

**Ultraweak variational formulation.** Integrating by parts both equation and building soft and hard BCs in, we obtain
\[
\begin{align*}
i\omega(p, q) - (u, \nabla q) &= -(u_0, q)_{\Gamma_u} - (u_n, q)_{\Gamma_i} \quad \forall q : q = 0 \text{ on } \Gamma_p \\
i\omega(u, v) - (p, \text{div } v) &= -(p_0, v_n)_{\Gamma_p} - (p, v_n)_{\Gamma_i}, \quad \forall v : v_n = 0 \text{ on } \Gamma_u.
\end{align*}
\]
We still have to figure out how to build in the impedance BC. This is where the adjoint operator comes in. Limiting ourselves to test functions satisfying condition \(v_n = -dq\) on \(\Gamma_i\), summing up the equations, and building the impedance BC in, we obtain:
\[
\begin{align*}
\{ & u \in L^2(\Omega) \\
&(u, A^*v) = -(u_0, q)_{\Gamma_u \cup \Gamma_i} - (p_0, v_n)_{\Gamma_p}, \quad v \in D(A^*). \quad (1.32)
\end{align*}
\]
So, what are the lessons of this section? As we have learned, the same boundary-value problem can admit many variational formulations. We will demonstrate in Section?? that all of them are simultaneously well-posed. They differ in energy setting corresponding to subtle regularity assumptions on the solution. Each of them can be used as a starting point for developing a separate FE method. The functional setting will translate into convergence in different (trial) norms. The two mixed formulations along with the corresponding reduced formulations enjoy a symmetric functional setting, and are eligible for Bubnov-Galerkin method (not a must though...). The strong and ultraweak variational formulations, with their non-symmetric functional setting, must be discretized with a Petrov-Galerkin scheme. Finally, we have introduced two more classical energy spaces: the \(L^2(\Omega)\) and \(H(\text{div}, \Omega)\).

### 1.4.2 Maxwell Equations

For a short introduction to Maxwell equations, we refer to [1]. We shall consider the time-harmonic Maxwell equations:

- **Faraday’s law,**
  \[
\frac{1}{\mu} \nabla \times E = -\frac{1}{\mu} K^{imp} - i\omega H , \quad (1.33)
\]
  - and **Ampere’s law,**
  \[
\nabla \times H = J^{imp} + \sigma E + i\omega \epsilon E . \quad (1.34)
\]

Here \(\epsilon, \mu, \sigma\) denote the material constants: permittivity, permeability and conductivity, and \(J^{imp}\) and \(K^{imp}\) stand for a prescribed impressed electric or magnetic current, respectively. The ultimate variational formulation can be obtained in terms of either the electric field \(E\), or the magnetic field \(H\). Depending upon the
choice, one of the equations is going to be satisfied in a weak sense, and the other one pointwise. If we choose
to solve for the electric field, we multiply the Ampere's law with $-i\omega$, then with a test function $F$, integrate
over $\Omega$ and integrate by parts to obtain,
\[
\int_{\Omega} \left( -i\omega \mathbf{H} \nabla \times F - (\omega^2 \epsilon - i\omega \sigma) \mathbf{E} \mathbf{F} \right) dx + i\omega \int_{\Gamma} \mathbf{n} \times \mathbf{H} \mathbf{F} dS = -i\omega \int_{\Omega} \mathbf{J}^{\text{imp}} \mathbf{F} dx, \quad \forall F. \tag{1.35}
\]
To simplify the notation, we have dropped the symbol for the dot product.

We introduce now the boundary conditions:

- Perfectly Conducting Boundary (PEC) on $\Gamma_D$:
  \[
  \mathbf{n} \times \mathbf{E} = \mathbf{n} \times \mathbf{E}_D,
  \]

- prescribed electric surface current on $\Gamma_N$:
  \[
  \mathbf{n} \times \mathbf{H} = \mathbf{n} \times \mathbf{H}_D =: \mathbf{J}^{\text{imp}}_S,
  \]

- an impedance boundary condition on $\Gamma_C$:
  \[
  \mathbf{n} \times \mathbf{H} + \gamma \mathbf{E}_t =: \mathbf{J}^{\text{imp}}_S.
  \tag{1.36}
  \]

Here $\mathbf{E}_t = -\mathbf{n} \times (\mathbf{n} \times \mathbf{E})$ stands for the tangential component of $\mathbf{E}$, $\gamma$ is a prescribed impedance, and $\mathbf{J}^{\text{imp}}_S$ is a prescribed electric surface current. Notice that the impressed surface current is tangent to the boundary. The impressed surface current on $\Gamma_C$ has a different interpretation than on $\Gamma_N$.

Introducing the boundary conditions into Equation 1.35, we obtain,
\[
\int_{\Omega} \left( -i\omega \mathbf{H} \nabla \times F - (\omega^2 \epsilon - i\omega \sigma) \mathbf{E} \mathbf{F} \right) dx + i\omega \int_{\Gamma_C} \gamma \mathbf{E}_t \mathbf{F} dS
= -i\omega \int_{\Omega} \mathbf{J}^{\text{imp}} \mathbf{F} + i\omega \int_{\Gamma_N \cup \Gamma_C} \mathbf{J}^{\text{imp}}_S \mathbf{F} dS,
\quad \forall F : \mathbf{n} \times \mathbf{F} = 0 \text{ on } \Gamma_D.
\]
Notice that $\mathbf{E}_t \mathbf{F} = \mathbf{E}_t \mathbf{F}_t$ and $\mathbf{J}^{\text{imp}}_S \mathbf{F} = \mathbf{J}^{\text{imp}}_S \mathbf{F}_t$.

The final variational formulation is obtained by using the Faraday equation to eliminate the magnetic field. We obtain,
\[
\left\{ \begin{array}{l}
\mathbf{n} \times \mathbf{E} = \mathbf{n} \times \mathbf{E}_D \text{ on } \Gamma_D, \\
\int_{\Omega} \left( \frac{1}{\mu} \nabla \times \mathbf{E} \nabla \times F - (\omega^2 \epsilon - i\omega \sigma) \mathbf{E} \mathbf{F} \right) dx + i\omega \int_{\Gamma_C} \gamma \mathbf{E}_t \mathbf{F} dS
= -i\omega \int_{\Omega} \mathbf{J}^{\text{imp}} \mathbf{F} + i\omega \int_{\Gamma_N \cup \Gamma_C} \mathbf{J}^{\text{imp}}_S \mathbf{F} dS, \\
\forall F : \mathbf{n} \times \mathbf{F} = 0 \text{ on } \Gamma_D.
\end{array} \right. \tag{1.37}
\]
\[\mathbf{n} \times \mathbf{E} = \mathbf{n} \times \mathbf{E}_t\] is a "rotated" tangential component of $\mathbf{E}$.
Preliminaries

Behind the choice of the primary variable are subtle regularity assumptions. In order for the electric field $E$ to “live” in energy space $H(\text{curl}, \Omega)$, the impressed magnetic current must be square integrable but the electric impressed current need not. We record the final formulas for the bilinear and linear forms.

$$X = H(\text{curl}, \Omega)$$

$$b(E, F) = \int_\Omega \left( \frac{1}{\mu} \nabla \times E \nabla \times F - (\omega^2 \varepsilon - i \omega \sigma)EF \right) \, dx + i \omega \int_{\Gamma_C} \gamma E_t F \, dS \quad (1.38)$$

$$l(F) = -i \omega \int_\Omega J^{\text{imp}} F - \int_\Omega \frac{1}{\mu} K^{\text{imp}} \nabla \times F \, dx + i \omega \int_{\Gamma_N \cup \Gamma_C} J^{\text{imp}}_S F \, dS$$

Recall that, as long as the shape functions are real-valued, there is no difference between the Galerkin methods based on bilinear or sesquilinear forms formulations.

**Formulation in terms of the magnetic field.** If we choose to work with the magnetic field, we treat the Faraday equation in the weak form. Since permeability $\mu$ may be a function of $x$, we multiply first the equation with $\mu$, and only then test it with a test function $F$ to obtain,

$$\int_\Omega (E \nabla \times F + i \omega \mu HF) \, dx + \int_{\Gamma} n \times E F \, dS = -\int_\Omega K^{\text{imp}} F \, dx \quad \forall F. \quad (1.39)$$

We discuss now the boundary conditions,

- prescribed electric surface current on $\Gamma_D$:

  $$n \times H = n \times H_D,$$

- Perfectly Conducting Boundary (PEC) on $\Gamma_N$, i.e. a prescribed magnetic surface current:

  $$n \times E = n \times E_D =: -K^{\text{imp}}_S,$$

- impedance boundary condition on $\Gamma_C$:

  $$n \times E - \frac{1}{\gamma} H_t = -\frac{1}{\gamma} n \times J^{\text{imp}} =: -K^{\text{imp}}_S.$$

Notice that the definition of the Dirichlet or Neumann part of the boundary depends upon the formulation. The Dirichlet data for the $E$-formulation has become now a Neumann data, and vice versa. The new form of the Cauchy boundary condition has been obtained by multiplying Equation 1.36 on the left by $n \times$ and dividing by impedance constant $\gamma$. Substituting the boundary conditions data into the boundary term in formulation 1.39, and restricting ourselves to test functions satisfying the homogeneous Dirichlet boundary condition we get,

$$\int_\Omega (E \nabla \times F + i \omega \mu HF) \, dx + \int_{\Gamma_C} \frac{1}{\gamma} n H_t F \, dS = -\int_\Omega K^{\text{imp}} F \, dx + \int_{\Gamma_N \cup \Gamma_C} K^{\text{imp}}_S F \, dS, \quad \forall F : n \times F = 0 \text{ on } \Gamma_D.$$
The final variational formulation is obtained by using the Ampere’s law to eliminate the electric field:

\[
\begin{aligned}
  n \times H &= n \times H_D \\
  \int_{\Omega} \left( \frac{1}{i\omega + \sigma} \nabla \times H \nabla \times F + i \omega \mu H F \right) \, dx + \int_{\Gamma_C} \frac{1}{\gamma} H_t F \, dS \\
  &= -\int_{\Omega} K^{imp} F \, dx + \int_{\Omega} \frac{1}{i\omega + \sigma} J^{imp} \nabla \times F \, dx + \int_{\Gamma_N \cup \Gamma_C} K^{imp} F \, dS,
\end{aligned}
\]

\(\forall F : n \times F = 0\) on \(\Gamma_D\).

The formulas for the bilinear and linear forms are as follows.

\[
\begin{aligned}
  X &= H(\text{curl}, \Omega) \\
  b(H, F) &= \int_{\Omega} \left( \frac{1}{i\omega + \sigma} \nabla \times H \nabla \times F + i \omega \mu H F \right) \, dx + \int_{\Gamma_C} \frac{1}{\gamma} H_t F \, dS \\
  l(F) &= -\int_{\Omega} K^{imp} F \, dx + \int_{\Omega} \frac{1}{i\omega + \sigma} J^{imp} \nabla \times F \, dx + \int_{\Gamma_N \cup \Gamma_C} K^{imp} F \, dS
\end{aligned}
\]

### 1.4.3 Maxwell Equations. A Deeper Look

The story behind Maxwell’s equations goes much deeper behind the need for a new energy space \(H(\text{curl}, \Omega)\). Complete (time harmonic) Maxwell’s equations include not only the Faraday and Ampère Laws but also the two Gauss laws and the conservation of (free) charge equation.

\[
\begin{aligned}
  \nabla \times E &= -i\omega(\mu H) \quad \text{Faraday’s Law} \\
  \nabla \times H &= J^{imp} + \sigma E + i\omega(\epsilon E) \quad \text{Ampère’s Law} \\
  \nabla \cdot (\mu H) &= 0 \quad \text{Gauss’ Magnetic Law} \\
  \nabla \cdot (\epsilon E) &= \rho^{imp} + \rho \quad \text{Gauss’ Electric Law} \\
  i\omega \rho + \nabla \cdot J &= 0 \quad \text{Conservation of charge}
\end{aligned}
\]

We have a total of seven scalar unknowns (three components of \(E, H\) each and \(\rho\), and a total of nine scalar equations. Obviously, the equations are linearly dependent. To simplify the discussion, we can eliminate the free charge density by combining the last two equations into one (we will call it the “continuity equation”),

\[
\begin{aligned}
  \nabla \times E &= -i\omega(\mu H) \quad \text{Faraday’s Law} \\
  \nabla \times H &= J^{imp} + \sigma E + i\omega(\epsilon E) \quad \text{Ampère’s Law} \\
  \nabla \cdot (\mu H) &= 0 \quad \text{Gauss’ Magnetic Law} \\
  -i\omega \rho^{imp} + \nabla \cdot J + i\omega \nabla \cdot (\epsilon E) &= 0
\end{aligned}
\]

The algebraic dependence structure is now clearly visible. The Gauss’ Magnetic Law is obtained by applying the divergence operator to both sides of the Faraday’s law, and the continuity equation is obtained by taking the divergence of the Ampère’s Law. The last two equations are thus automatically satisfied once the first two hold. Note that once the electric field \(E\) is known, either the Gauss’ electric law or the conservation of
charge equation, can be used to compute the free charge density \( \rho \). Notice also that the prescribed impressed current and charge must be compatible with each other (satisfy the conservation of charge equation).

Critical to the discretization of Maxwell equations is the fact that this automatic satisfaction of the Gauss’ Magnetic Law and the continuity equations carries over to the weak form of the equations, and then to the discrete level as well.

We shall focus on the formulation 1.37 in terms of electric field \( E \). Analogous results hold for the other formulation as well. First of all, once the electric field is known, the corresponding magnetic field is computed using the strong form of the Faraday’s law:

\[
-i\mu \omega \mathbf{H} = \nabla \times \mathbf{E}.
\]

Taking the divergence of both sides, we verify easily the Gauss’ magnetic Law.

In order to recover the continuity equation from variational formulation 1.37, we employ a special test function \( \mathbf{F} = \nabla q \) where \( q \in H^1(\Omega) \), \( q = 0 \) on \( \Gamma_D \) to obtain:

\[
-\int_{\Omega} (\omega^2 \epsilon - i \omega \sigma) \mathbf{E} \cdot \nabla q + i \omega \int_{\Gamma_C} \gamma \mathbf{E}_\tau \cdot \nabla q = -i \omega \int_{\Omega} \mathbf{J}^{imp} \cdot \nabla q + i \omega \int_{\Gamma_N \cup \Gamma_C} \mathbf{J}^{imp}_S \nabla q \quad \forall q \quad (1.44)
\]

The equation represents not only a weak form of the continuity equation but also additional (automatically satisfied) boundary conditions on \( \Gamma_N \) and \( \Gamma_C \).

The critical point here is the fact that we could make the substitution \( \mathbf{F} = \nabla q \), i.e. that the gradients \( \nabla q \) live in the energy space \( H(\text{curl}, \Omega) \).

**Exercises**

**Exercise 1.4.1** Explain why space of vector-valued \( L^2 \)-functions,

\[
L^2(\Omega) := \{ u : \Omega \to \mathbb{C}^N : \int_{\Omega} |u|^2 < \infty \}
\]

is isomorphic and isometric with \( N \) copies of scalar-valued functions,

\[
(L^2(\Omega))^N.
\]

(1 point)

**Exercise 1.4.2** Write down explicitly trial and test spaces, and formulas for sesquilinear and antilinear forms for all six variational formulations for the acoustic problem. (1 point)

**Exercise 1.4.3** Consider the Faraday and Ampère Laws:

\[
\nabla \times E = -i\omega \mu \mathbf{H} \quad \text{Faraday’s Law}
\]

\[
\nabla \times \mathbf{H} = \mathbf{J}^{imp} + \sigma \mathbf{E} + i \omega \mathbf{E} \quad \text{Ampère’s Law}
\]
accompanied with BCs:

\[ n \times E = n \times E_0 \quad \text{on } \Gamma_1 \]

\[ n \times H = n \times H_0 \quad \text{on } \Gamma_2 \]

Proceed along exactly the same lines as for acoustics equations, to derive trivial, mixed, ultraweak and reduced variational formulations (a total of six) for the Maxwell equations. Introduce the \( H(\text{curl}, \Omega) \) energy space. (5 points)

Exercise 1.4.4 Integration by parts formulas. Let \( \Omega \subset \mathbb{R}^3 \) be a domain with boundary \( \partial \Omega \). Use elementary integration by parts to derive the following integration by parts formulas.

\[
\begin{align*}
\int_{\Omega} \nabla u \cdot v &= - \int_{\Omega} u \nabla \cdot v + \int_{\partial \Omega} n u v \\
\int_{\Omega} (\nabla \times E) \cdot F &= \int_{\Omega} E \cdot (\nabla \times F) + \int_{\partial \Omega} (n \times E) \cdot F \\
\int_{\Omega} (\nabla \cdot u) v &= - \int_{\Omega} u \cdot (\nabla v) + \int_{\partial \Omega} u \cdot n v \\
\end{align*}
\]

(3 points)

Exercise 1.4.5 Maxwell problem. Repeat discussion from Section 1.4.3 on the implicit satisfaction of the Gauss’ Magnetic Law and continuity equations for the variational formulation in terms of magnetic field \( H \). (5 points)

Coercive Problems

1.5 Minimization Principle and the Ritz Method

Abstract minimization principle. The real case. Assume the symmetric functional setting with trial and test spaces coinciding with each other, \( U = V \). Assume additionally that the spaces are real, and consider bilinear and linear forms corresponding to the abstract variational formulation. Define the quadratic energy functional (total potential energy):

\[
J(u) := \frac{1}{2} b(u, u) - l(u)
\]

and derive the corresponding Gateaux derivative,

\[
\langle \delta J(u), v \rangle = \frac{1}{2} [b(u, v) + b(v, u)] - l(v)
\]

If we additionally assume that form \( b \) is symmetric, i.e.

\[ b(u, v) = b(v, u) \quad u, v \in U \]
Preliminaries

the formula reduces to:

$$\langle \delta J(u), v \rangle = b(u, v) - l(v)$$

The abstract variational formulation:

$$\begin{cases} u \in U \\ b(u, v) = l(v) \quad v \in U \end{cases}$$  \quad (1.45)

represents thus a necessary condition for \( u \) to be a minimizer (or maximizer as well).

Conversely, a simple computation reveals that,

$$J(u + v) - J(u) = b(u, v) - l(v) + \frac{1}{2} b(v, v)$$

If form \( b(v, v) \) is positive definite over \( U = V \), i.e.

$$b(v, v) > 0 \quad v \in V, v \neq 0$$  \quad (1.46)

then solution \( u \) to the variational problem is seen to be the unique minimizer of the total potential energy functional \( J(u) \).

The minimization problem:

$$u = \arg \min_{w \in U} J(w)$$  \quad (1.47)

and the variational formulation (1.45) are thus equivalent to each other.

**Well posedness.** Equivalence of the minimization and the variational problems does not prove that either of them is well-posed. The positive-definitness of form \( b(u, v) \) implies that \( b(u, v) \) may be identified as an inner product with the corresponding energy norm

$$\|u\|_E^2 = b(u, u)$$  \quad (1.48)

The well-posedness of the variational problem is implied then by the Riesz Representation Theorem [28], provided we can show that form \( l(v) \) is continuous in the energy norm, and the space \( U \) equipped with the energy norm, is complete. In order to guarantee these properties, we upgrade the positive definitness of form \( b(u, v) \) to the coercivity condition. We say that form \( b(u, v) \) is \( U \)-coercive if there exists a constant \( \alpha > 0 \) such that

$$\alpha \|u\|_U^2 \leq b(u, u) \quad u \in U$$  \quad (1.49)

Note that the coercivity indeed implies positive-definitness. With the coercivity assumption in place, the original and energy norms are equivalent,

$$\alpha \|u\|_U^2 \leq \|u\|_E^2 \leq M \|u\|_U^2$$

Consequently, if \((U, \| \cdot \|_U)\) is complete then so is \((U, \| \cdot \|_E)\). By the same token, if \( l(v) \) is continuous wrt norm \( \| \cdot \|_U \) then it is also continuous wrt to the energy norm.
The Ritz method. Assume \( b(u,v) \) is hermitian and positive-definite. Let \( U_h \subset U \) be a finite-dimensional subspace of \( U \). The following problems are equivalent to each other.

(i) Minimization of energy over the approximate space \( U_h \):

\[
J(u_h) = \min_{w_h \in U_h} J(w_h) .
\]

(ii) Galerkin approximation of the variational problem:

\[
\begin{aligned}
& \{ u_h \in U_h \\
& b(u_h,v_h) = l(v_h) \quad \forall v_h \in U_h .
\end{aligned}
\]

(iii) Minimization of the distance between the exact and approximate solutions in the energy norm:

\[
\| u - u_h \|_E = \min_{w_h \in U_h} \| u - w_h \|_E
\]

where \( \|v\|_E^2 := b(v,v) \).

(iv) Minimization of the residual in the norm dual to the energy norm,

\[
\sup_{v \in U} \frac{|b(u_h,v) - l(v)|}{\|v\|_E} = \min_{w_h \in U} \sup_{v \in U} \frac{|b(w_h,v) - l(v)|}{\|v\|_E}
\]

**Proof** Equivalence of (i) and (ii) has already been proved for space \( U \). As \( U \) was an arbitrary inner product space, the result holds also for the finite dimensional space \( U_h \).

To see the equivalence of (i) and (iii), expand the formula for the energy norm,

\[
\frac{1}{2} \| u - u_h \|_E^2 = \frac{1}{2} b(u - u_h, u - u_h) = \frac{1}{2} b(u,u) + \frac{1}{2} b(u_h,u_h) - \underbrace{b(u,u_h)}_{=l(u_h)}
\]

Equivalence with the fourth condition is left as an exercise, comp. Exercise 1.5.5.

In terms of the energy norm, Ritz method delivers the orthogonal projection (the best approximation error).

Equivalence of the original and energy norms implies stability of the discretization in the original norm. Indeed,

\[
\alpha \| u - u_h \|_U^2 \leq \| u - u_h \|_E^2 = \inf_{w_h \in U_h} \| u - w_h \|_E^2 \leq M \inf_{w_h \in U_h} \| u - w_h \|_U^2
\]

which implies that

\[
\| u - u_h \| \leq \sqrt{\frac{M}{\alpha}} \inf_{w_h \in U_h} \| u - w_h \|_U
\]

**Exercises**
**Preliminaries**

**Exercise 1.5.1** Use the abstract minimization framework to identify energy functionals for the Poisson and elasticity problems. Verify positive definiteness of bilinear forms. (5 points)

**Exercise 1.5.2** Consider the diffusion-reaction problem with \( a_{ij} = \delta_{ij}, b_j = 0 \) and \( c > 0 \) with arbitrary BCs. Identify the energy functional and verify positive definiteness of bilinear form. (3 points)

**Exercise 1.5.3** Consider again the diffusion-reaction problem discussed in Exercise 1.5.2 but with a relaxed condition for the reaction coefficient \( c \geq 0 \) (in particular, the reaction term may vanish) and the Cauchy BC imposed on the whole boundary \( \Gamma \):

\[
\frac{\partial u}{\partial n} + \beta u = g
\]

Derive the corresponding classical variational formulation and identify condition(s) for coefficient \( \beta \) for the bilinear form to positive definite. (5 points)

**Exercise 1.5.4** Extend the minimization principle to the complex case. Form \( b(u, v) \) must be hermitian and positive definite, and the energy functional is given by:

\[
J(u) := \frac{1}{2} b(u, u) - \Re(l(u))
\]

(4 points)

**Exercise 1.5.5** Prove that the Ritz method is equivalent to the minimization of the residual measured in the norm dual to the energy norm. (3 points)

---

### 1.6 Lax-Milgram Theorem and Cea’s Lemma

**Coercive bilinear form.** A sesquilinear form \( b(u, v) \) defined on a Hilbert space \( U \) is **coercive** if there exists a positive (coercivity) constant \( \alpha > 0 \) such that

\[
|b(u, u)| \geq \alpha \|u\|^2_U, \quad \forall u \in U.
\]  

(1.50)

**THEOREM 1.6.1 (Lax-Milgram Theorem)**

Let \( U \) be a Hilbert space. Let \( b(u, v) \) be a continuous and coercive sesquilinear form defined on \( U \times U \). Let \( l \in U' \). The (abstract) variational problem

\[
\begin{cases}
    u \in U \\
    b(u, v) = l(v) \quad \forall v \in U 
\end{cases}
\]
is then well-posed, i.e. it admits a unique solution \( u \) that depends continuously upon the data, namely:

\[
\|u\|_U \leq \frac{1}{\alpha} \|l\|_{U'}
\]

where \( \alpha \) is the coercivity constant.

**PROOF** Lax Milgram Theorem is a corollary to the Babuška-Nečas Theorem which in turn is a reformulation of Banach Closed Range Theorem to variational problems. The following is an elementary proof reproduced from [6], p.62. The proof relies on two theorems: Riesz Representation Theorem, and Banach Contractive Map Theorem. Both of these results are more elementary than the Closed Range Theorem.

Consider the map:

\[
Tu = u - \rho R^{-1}(Bu - l)
\]

where \( B : U \to U' \) is the operator corresponding to bilinear form \( b(u, v) \), and \( R : U \to U' \) is the Riesz operator corresponding to the scalar product in \( U \). We shall prove that, with a proper choice of constant \( \rho > 0 \), map \( T : U \to U \) is a contraction, i.e. there exists a contraction constant \( 0 < k < 1 \) such that

\[
\|Tu\|_U \leq k\|u\|_U
\]

By the Contractive Map Theorem, map \( T \) has then a unique fixed point \( u \), i.e. \( Tu = u \), which is equivalent to \( Bu = l \). Stability estimate follows directly from the coercivity assumption,

\[
\alpha\|u\|^2 \leq b(u, u) = \|l(u)\| \leq \|l\|_{U'} \|u\|_U
\]

We have now,

\[
\|Tu\|^2_U = (u - \rho R^{-1}Bu, u - \rho R^{-1}Bu)
\]

\[
= \|u\|^2_U - \rho(R^{-1}Bu, u) + \rho^2\|R^{-1}Bu\|_U
\]

\[
= \|u\|^2_U - \rho(Bu, u) + \rho^2\|R^{-1}Bu\|_U
\]

\[
= \|u\|^2_U - 2\rho b(u, u) + \rho^2\|R^{-1}Bu\|_U
\]

\[
\leq (1 - 2\rho \alpha + \rho^2 M^2) \|u\|^2_U
\]

since \( \|R\| = 1 \) and \( \|B\| \leq M \). Selecting \( \rho \in (0, 2\alpha/M^2) \), we get \( k < 1 \) which finishes the proof.

**Galerkin orthogonality.** Let \( U_h \subset U \) and \( V_h \subset V \) be approximate trial and test spaces. Let \( u_h \in U_h \) be the Galerkin approximation to the variational problem,

\[
\begin{cases}
  u_h \in U_h \\
  b(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h
\end{cases}
\]

(1.51)
Then the Galerkin orthogonality condition holds:

\[ b(u - u_h, v_h) = 0 \quad \forall v_h \in V_h. \quad (1.52) \]

**THEOREM 1.6.2 (Cea's Lemma)**

Let \( b(u, v) \) be a continuous and coercive sesquilinear form defined on a Hilbert space \( U \),

\[ |b(u, v)| \leq M \|u\| \|v\|, \]
\[ |b(v, v)| \geq \alpha \|v\|^2, \quad \alpha > 0. \]

Let \( U_h \subset U \) and let \( u_h \in U_h \) be the Bubnov-Galerkin projection of some \( u \in U \) onto subspace \( U_h \), i.e.

\[ b(u - u_h, v_h) = 0 \quad \forall v_h \in U_h. \]

Then the following stability result holds:

\[
\frac{\|u - u_h\|_U}{\text{approximation error}} \leq \frac{M}{\alpha} \inf_{w_h \in U_h} \frac{\|u - w_h\|_U}{\text{the best approximation error}}. \quad (1.53)
\]

Note that the Cea’s result does not provide an optimal stability constant for the hermitian problems (compare with the Ritz method).

**Exercises**

**Exercise 1.6.1** (3 points)

1.7 Examples of Problems Fitting the Lax-Milgram Theory

1.7.1 A General Diffusion-Convection-Reaction Problem

\[
\begin{cases}
- \frac{\partial}{\partial x_1} \left( a_{ij} \frac{\partial u}{\partial x_j} \right) + b_j \frac{\partial u}{\partial x_j} + cu = f & \text{in } \Omega \\
\quad u = 0 & \text{on } \Gamma_1 \\
\quad a_{ij} \frac{\partial u}{\partial x_j} = g & \text{on } \Gamma_2 \\
\quad a_{ij} \frac{\partial u}{\partial x_j} - \beta u = g & \text{on } \Gamma_3
\end{cases}
\quad (1.54)
\]
The corresponding forms:

\[
\begin{align*}
    b(u,v) &= \int_{\Omega} \left\{ a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + b_j \frac{\partial u}{\partial x_j} v + cuv \right\} \, dx + \int_{\Gamma_3} \beta uv \, dS, \\
    l(v) &= \int_{\Omega} fv \, dx + \int_{\Gamma_2 \cup \Gamma_3} gv \, dS.
\end{align*}
\]

(1.55)

**Sobolev space** \(H^1(\Omega)\). Concept of distributional derivative. Elementary example:

\[
\frac{d}{dx} R u = R u' + [u(x_0)] \delta_{x_0}
\]

### 1.7.2 Linear Elasticity

**THEOREM 1.7.1 (Korn’s inequality)**[24]

Let \(\Omega\) be a bounded Lipschitz domain in \(\mathbb{R}^N\), \(N=2,3\). There exists a positive constant \(C_K > 0\) such that:

\[
C_K \|u\|_{H^1(\Omega)}^2 \leq \|u\|_{L^2(\Omega)}^2 + \sum_{i,j} \|\epsilon_{ij}(u)\|_{L^2(\Omega)}^2 \quad \forall u \in (H^1(\Omega))^N
\]

(1.56)

where \(\epsilon_{ij}(u) = \frac{1}{2}(u_{i,j} + u_{j,i})\) is the symmetric part of \(u\) (linearized strain). Constant \(C_K\) depends upon the domain but it is independent of \(u\).

**THEOREM 1.7.2 (Coercivity for linear elasticity)**

Let the assumptions of Korn’s inequality hold. Let \(\Gamma_1\) be a subset of boundary \(\partial \Omega\) with non-zero measure. There exists then a coercivity constant \(\alpha > 0\) such that:

\[
\alpha \|v\|^2 \leq \sum_{i,j} \|\epsilon_{ij}(v)\|_{L^2(\Omega)}^2 \quad \forall v \in (H^1(\Omega))^N : v = 0 \text{ on } \Gamma_1.
\]

(1.57)

**PROOF** We proceed by contradiction. Let \(v_n\) be a sequence such that \(\|v\|_{L^2(\Omega)} = 1\) and the right-hand side above converges to zero. By Korn’s inequality, sequence \(v_n\) is bounded in \(H^1(\Omega)\). Consequently, we can extract from \(v_n\) a subsequence, denoted with the same symbol, converging weakly to a limit \(v\), \(v_n \rightharpoonup v\) in \(H^1(\Omega)\). Next we observe that the \(L^2\) norm of the strain is positive definite. Indeed, if it vanishes that \(v\) must be a rigid body motion and the kinematic boundary condition sets it to zero. Positive definiteness implies strict convexity. In turn, strict convexity and (strong) continuity implies weak lower semi-continuity. Consequently,

\[
\sum_{i,j} \|\epsilon_{ij}(v)\|_{L^2(\Omega)}^2 \leq \liminf \sum_{i,j} \|\epsilon_{ij}(v_n)\|_{L^2(\Omega)}^2 = 0
\]

and, therefore, the weak limit must also be a rigid body motion. The kinematic BC implies then again that \(v = 0\). Finally, by the Rellich Embedding Theorem, \(v_n \to 0\) in the \(L^2\) norm. This is
a contradiction with the assumption that \( \|v_n\|_{L^2(\Omega)} = 1 \) (the limit should have a unit \( L^2 \) norm as well).

### 1.7.3 Model Curl-Curl Problem

\[
\begin{aligned}
\left\{ \begin{array}{l} 
n \times E = 0 \text{ on } \Gamma_1 \\
\int_{\Omega} \nabla \times E \cdot \nabla \times F + \epsilon \int_{\Omega} E \cdot F = \int_{\Omega} f \cdot F + \int_{\Gamma_2} g \cdot F \\
\forall F : n \times F = 0 \text{ on } \Gamma_1
\end{array} \right.
\end{aligned}
\]

(1.58)

### Exercises

**Exercise 1.7.1**

Distributional derivatives (comp. Exercise 1.4.5). Let a domain \( \Omega \subset \mathbb{R}^N, N = 2,3 \), be split into two subdomains \( \Omega_1, \Omega_2 \) with a smooth interface \( \Gamma \). Let \( u, E, v \) be functions consisting of two smooth branches \( u^I, E^I, v^I, I = 1,2 \) defined in the subdomains. By “smooth” we understand \( u^I \in C_1(\Omega^I) \) etc. Let \( n \) be the unit vector on interface \( \Gamma \) pointing from subdomain \( \Omega_1 \) into subdomain \( \Omega_2 \).

(i) Let \( \phi \in C_0^\infty(\Omega) \) be a Schwartz test function (scalar- or vector-valued). Use elementary integration by parts to derive the following formulas:

\[
- \int_{\Omega} u \nabla \phi = \sum_I \int_{\Omega^I} \nabla u^I \phi + \int_{\Gamma} [u] n \phi ,
\]

\[
\int_{\Omega} E \nabla \times \phi = \sum_I \int_{\Omega^I} \nabla \times E^I \phi + \int_{\Gamma} [n \times E] \phi ,
\]

\[
\int_{\Omega} v \nabla \cdot \phi = \sum_I \int_{\Omega^I} \nabla \cdot v^I \phi + \int_{\Gamma} [n \cdot v] \phi
\]

where

\[
[u] = u^2 - u^1, \quad [n \times E] = n \times (E^2 - E^1), \quad [n \cdot v] = n(v^2 - v_1).
\]

(ii) Interpret the formulas above in the language of distributions using the definition of regular distributions, distributional derivatives and corresponding operators of grad, curl and div understood in the distributional sense. You will have to introduce a multidimensional equivalent of Dirac’s delta.

(iii) Conclude that functions \( u, E, v \) belong to energy spaces \( H^1(\Omega), H(\text{curl}, \Omega), H(\text{div}, \Omega) \) if and only if the corresponding continuity conditions across the interface \( \Gamma \) are satisfied:

\[
[u] = 0, \quad [n \times E] = 0, \quad [n \cdot v] = 0.
\]

(5 points)

**Exercise 1.7.2**

Poincaré inequality.
(i) Use elementary means to prove the 1D version of Poincaré inequality:

\[ \alpha \int_0^1 |u|^2 \leq \int_0^1 |u'|^2 \quad \forall u \in H^1(0,1) : u(0) = 0 \quad \alpha > 0. \]

Provide a concrete estimate for \( \alpha \). 
*Hint:* Apply the Second Fundamental Theorem of Differential Calculus to interval \((0, x)\),

\[ u(x) = \int_0^x u'(s) \, ds \]

and take it from there.

(ii) Interpret the best (largest) Poincaré constant \( \alpha \) as the minimum eigenvalue of the 1D Laplace operator with appropriate BC. Use Sturm-Liouville Theorem to compute \( \alpha \) and compare it with the estimate obtained in the previous step.

(iii) Use scaling arguments to derive the best Poincaré constant for an interval of length \( l \) to see how \( \alpha \) changes with the size of the domain.

(iv) Repeat the first three steps for an elementary 2D scenario with \( \Omega = (0,1)^2 \) and \( u \) vanishing on west boundary: \( u(0, y) = 0, y \in (0,1) \) (you will need refresh your skills on separation of variables).

(5 points)

**Exercise 1.7.3** Coercivity of elasticity bilinear form.

(i) Consider the elasticity problem in a square domain \((0,1)^2\) with kinematic BC on the south and west boundaries,

\[ u(x,0) = 0, \quad x \in (0,1) \quad \text{and} \quad u(0,y) = 0, \quad y \in (0,1) \, . \]

Use elementary means similar to those in Exercise 1.7.2 to prove that there exists a positive constant \( C > 0 \) such that

\[ C \int_{\Omega} |u|^2 \leq \int_{\Omega} \sum_{i,j} |\epsilon_{ij}(u)|^2 \quad \text{for every kinematically admissible} \ v \in (H^1(\Omega))^2. \]

(ii) Use the standard assumptions on the elasticities to conclude that the elastic bilinear form,

\[ b(u, v) = \int_{\Omega} E_{ijkl} u_{k,l} v_{i,j}, \]

is coercive on the space of kinematically admissible displacements.

(iii) Interpret the best (largest) coercivity constant as the smallest elastic eigenfrequency (with density \( \rho = 1 \)),

\[ \begin{cases} 
  u \in V_0, & \lambda \in \mathbb{R} \\
  b(u, v) = \lambda(u, v) & \forall v \in V_0, 
\end{cases} \]

where \( V_0 \) is the space of kinematically admissible displacements. Use a scaling argument to estimate \( \alpha = \lambda_{\text{min}} \) in terms of the size of the domain.
Exercise 1.7.4 Maxwell model problem. Discuss well-posedness of model problem (1.58) using either Riesz or Lax-Milgram Theorems. Under what additional assumptions on data \( f, g \), the solution is bounded uniformly in \( \epsilon \)? *Hint:* use a Helmholtz decomposition for \( f \) and assume appropriate scalings in \( \epsilon \).

(5 points)

Conforming Finite Elements and the Exact Sequence

1.8 Conforming Finite Elements

1.8.1 Classical \( H^1 \)-Conforming Elements

**Ciarlet’s definition of a finite element.** In order to define a finite element, we must introduce:

- a bounded domain (usually a polygon or polyhedral) \( K \),
- a space of FE shape functions (usually polynomials) \( X(K) \) contained in the appropriate energy space, \( \dim X(K) = n \),
- a set of linear and continuous functionals \( \psi_j \), called *degrees-of-freedom* (d.o.f.) defined on a subset \( \mathcal{X}(K) \) (of sufficiently regular functions) of the energy space containing the FE space \( X(K) \),

\[
\psi_j : \mathcal{X}(K) \to \mathbb{R}(\mathcal{E}), \quad j = 1, \ldots, n, \tag{1.59}
\]

such that restrictions of \( \psi_j \) to \( X(K) \) are linearly independent, i.e. they form a basis in the algebraic dual of \( X(K) \).

The linear independency condition is known as the *unisolvence condition*. The corresponding dual basis in \( X(K) \),

\[
\phi_i \in X(K), \quad \langle \psi_j, \phi_i \rangle = \delta_{ij}, \quad i, j = 1, \ldots, n, \tag{1.60}
\]

is identified as *FE shape functions*.

The definition should be treated rather informally. It tells only a part of the story. In particular, implicit in the construction is an assumption that by equating certain d.o.f. for neighboring elements, we guarantee that the union of the FE shape functions lives in the global energy space. This is best explained starting with examples.
Lagrange finite elements. Element: master interval, quad, hehaxedron, master triangle, tetrahedron (simplices), master prism. FE space: $\mathcal{P} = \mathcal{P}^p(0, 1), \mathcal{Q}^{p,q} := \mathcal{P}^p \otimes \mathcal{P}^q, \mathcal{Q}^{p,q,r} := \mathcal{P}^p \otimes \mathcal{P}^q \otimes \mathcal{P}^r, \mathcal{P} = \mathcal{P}^p(T), \mathcal{P}^p(T) \otimes \mathcal{P}^q(0, 1)$. Degrees of freedom: values of shape functions at Lagrangian nodes:

$$\psi_j : X(K) \ni \phi \rightarrow \phi(a_j) \in \mathbb{R}.$$ 

Lagrangian nodes are uniformly distributed over the master element, their number matches the dimension of the corresponding space of element shape functions. I am frequently drawing them to compute the dimension of the space.

**Parametric $H^1$-conforming Lagrange element.** Given a master element $\hat{K}$ and an element map $x_K$ from $\hat{K}$ onto a physical element $K \subset \mathbb{R}^N$,

$$x_K : \hat{K} \rightarrow K, \quad x = x(\xi),$$

we introduce the triple:

- element $K$,
- space of element shape functions:
  $$X(K) := \{ \hat{u} \circ x_K^{-1} : \hat{u} \in X(\hat{K}) \},$$
- element d.o.f.:
  $$\psi_j : X(K) \ni u \rightarrow u(a_j) \in \mathbb{R},$$

  where $a_j$ is the image of Lagrangian node $\hat{a}_j$ in the master element.

Note the commuting property:

$$\langle \psi_j, u \rangle = \langle \hat{\psi}_j, \hat{u} \rangle.$$

For general parametric elements, the commuting property may be enforced by definition, i.e. it defines the d.o.f. on the physical element.

**Interpolation operator.**

$$X(K) \ni u \rightarrow \Pi_K u := \sum_{j=1}^{n} \langle \psi_j, u \rangle \phi_j \in X(K).$$

The commuting property for the d.o.f. implies the corresponding commuting property for interpolation on master and physical elements:

$$\hat{\Pi}(u \circ x_K^{-1}) = \Pi(u \circ x_K^{-1})$$

or, in a more concise form ("breaking the hat" property):

$$\hat{\Pi}(u) = \Pi \hat{u}.$$
Illustration of the commuting property:

\[ u \xrightarrow{\Pi} \Pi u \]

\[ \downarrow x_K^{-1} \quad \downarrow x_K^{-1} \]

\[ \hat{u} \xrightarrow{\hat{\Pi}} \hat{\Pi} \hat{u} = \hat{\Pi} \hat{u}. \]

Finite Element space.

\[ X_h := \{ u \in H^1(\Omega) : u|_K \in X(K) \quad \forall K \in \mathcal{T}_h \}. \]

Enforcement of the global continuity leads to the identification of d.o.f. corresponding to element vertices, edges and faces, their equality for neighboring elements and, eventually, the notion of degrees of freedom as well-defined functionals on the the global FE space \( X_h \),

\[ \psi_j : \mathcal{X}(\Omega) \supset X_h \rightarrow \mathbb{R}. \]

The corresponding dual basis is identified as (Galerkin) basis functions \( e_i \). Degrees of freedom and the corresponding basis functions are naturally classified into vertex, edge, face and element interior shape functions. They are unions of the corresponding element shape functions. Finally, we have the global interpolation operator:

\[ \Pi u = \sum_j \langle \psi_j, u \rangle e_j. \]

Both symbols for the global interpolation operator \( \Pi \) and global d.o.f. \( \psi_j \) are typically overloaded.

Hierarchical shape functions. In the \( p \)-version of the FE method, the mesh is fixed and we converge to the exact solution by raising the polynomial order \( p \) of approximation, hence the name. The order can be raised uniformly or adaptively, i.e. only in some elements. We arrive at the need of meshes combining elements of varying order. The use of such elements takes also place in the \( h \)-version of the FE method. Frequently, we need to employ higher order elements \( (p = 4, 5) \) locally\(^\dagger\) with most of the domain discretized with lower order elements, say \( p = 2 \). Hence the need for building a code that supports variable order elements.

Varying polynomial order with Lagrange elements is practically impossible but it is very natural and straightforward with hierarchical shape functions that have been used in the \( p \)-method from the very beginning.

The rise of the \( p \)-method and hierarchical shape functions revealed limitations of Ciarlet’s formalism for constructing finite elements. Hierarchical shape functions (Szabo called them \textit{modes}) reflected geometry of the mesh and were constructed without defining d.o.f. first. They are classified into vertex, edge, face, and element shape functions (modes). Support of a vertex basis function spans over all elements sharing the vertex, support of an edge basis functions consists of all elements sharing the edge, support of a face basis

\(^\dagger\)E.g. to avoid the so-called locking phenomenon occurring in the discretization of thin-walled structures.
function spans over (at most two) elements sharing the face and, finally, support of an element basis function includes the element only. As for Lagrange elements, the basis functions are unions of the corresponding contributing element shape functions, possibly premultiplied with a sign factor accounting for orientation.

Once the shape functions have been introduced, we may try to identify the corresponding d.o.f. and then proceed with the construction of the interpolation operator. This is not so straightforward as the shape functions imply the uniqueness of the corresponding d.o.f. (the dual basis) only on the FE space of element shape functions $X(K)$ but not the bigger and rather ambiguous subspace $X'(K)$ of the energy space. Recall that, in the Ciarlet definition, the choice of subspace $X'(K)$ is simply driven by necessary regularity assumptions to make the d.o.f. well defined. Early attempts to identify d.o.f. corresponding to hierarchical shape functions led to wrong choices of subspace $X'(K)$ and, most importantly, suboptimal interpolation operators.

The alternative came with the construction of *Projection-Based (PB) interpolation* operators, [27, 16, 13, 8, 17, 15]. Here we construct the interpolation operators without using any d.o.f.. In fact, we do not need to define the d.o.f. at all. For academic reasons, we may try to identify d.o.f. that would result in the PB interpolation using Ciarlet's definition.

### 1.8.2 $H(\text{curl})$- and $H(\text{div})$-Conforming Elements. Exact Sequence

**3D Exact sequence.** Thinking in terms of FE spaces first, degrees-of-freedom next.

$$
\begin{align*}
H^1 & \xrightarrow{\nabla} H(\text{curl}) \xrightarrow{\nabla \times} H(\text{div}) \xrightarrow{\nabla} L^2 \\
\cup & \quad \cup \\
W^p & \xrightarrow{\nabla} Q^p \xrightarrow{\nabla \times} V^p \xrightarrow{\nabla} Y^p
\end{align*}
$$

*(1.61)*

**2D Exact sequence.** Discuss structure of the curl operator for 2D cases:

**Case:** $E = (E_1(x, y), E_2(x, y), 0)$

$$
\nabla \times E = (0, 0, E_{2,1} - E_{1,2})
$$

leads to the definition:

$$
E = (E_1(x, y), E_2(x, y)), \quad \text{curl} E := E_{2,1} - E_{1,2}
$$

**Case:** $E = (0, 0, E_3(x, y))$

$$
\nabla \times E = (E_{3,2}, -E_{3,1})
$$

leads to the definition:

$$
\phi = \phi(x, y), \quad \nabla \times \phi = \left( \frac{\partial \phi}{\partial y}, -\frac{\partial \phi}{\partial x} \right)
$$

**2D exact sequence:**
Preliminaries

\[
H^1 \xrightarrow{\nabla} H(\text{curl}) \xrightarrow{\text{curl}} L^2 \\
\cup \cup \cup
\]

\[
W^p \xrightarrow{\nabla} Q^p \xrightarrow{\text{curl}} Y^p
\]

“Rotated” 2D exact sequence:

\[
H^1 \xrightarrow{\nabla} H(\text{curl}) \xrightarrow{\text{div}} L^2 \\
\cup \cup \cup
\]

\[
W^p \xrightarrow{\nabla} V^p \xrightarrow{\text{div}} Y^p
\]

1D exact sequence:

\[
H^1 \xrightarrow{\partial} L^2 \\
\cup \cup
\]

\[
\mathcal{P}^p \xrightarrow{\partial} \mathcal{P}^{p-1}
\]

Lowest order tetrahedral element of the first type. Let \( K \) be an arbitrary tetrahedron. The FE spaces are defined as follows.

\[
W = W_1 = \mathcal{P}^1(K)
\]

\[
Q = Q_1 = \{ E \in (\mathcal{P}^2(K))^3 : E_i|_e \in \mathcal{P}^1(e), \text{ for each edge } e \} 
\]

\[
V = V_1 = \{ v \in (\mathcal{P}^2(K))^3 : v_n|_f \in \mathcal{P}^1(e), \text{ for each face } f \} 
\]

\[
Y = Y_1 = \mathcal{P}^0(K)
\]

where \( E_i = E \cdot \tau_e \) is the tangential component of vector \( E \), and \( v_n = v \cdot n_f \) is the normal component of \( v \) with \( \tau_e \) denoting a unit tangent vector for edge \( e \), and \( n_f \) a unit normal vector for face \( f \). Note that the definition is independent of the choice of the edge and face unit vectors, and

\[
\begin{align*}
\dim W &= \text{number of vertices} = 4 \\
\dim Q &= \text{number of edges} = 6 \\
\dim V &= \text{number of faces} = 4 \\
\dim Y &= \text{number of elements} = 1
\end{align*}
\]

Argue why the spaces with operators of grad, curl and div, form the exact sequence. The element d.o.f. are defined as follows.

\[
H^1(K) \ni ? \ni u \rightarrow u(v) \in \mathbb{R} \quad \text{for each vertex } v
\]

\[
H(\text{curl}, K) \ni ? \ni E \rightarrow \int_e E_i \in \mathbb{R} \quad \text{for each edge } e
\]

\[
H(\text{div}, K) \ni ? \ni v \rightarrow \int_f v_n \in \mathbb{R} \quad \text{for each face } f
\]

\[
L^2(K) \ni q \rightarrow \int_K q \in \mathbb{R} \quad \text{for element } K
\]
The question marks stand for subspaces of energy spaces, consisting of sufficiently regular functions for which the d.o.f. are well-defined. They are usually characterized in terms of Sobolev spaces $H^s$ with real exponent, $s \in \mathbb{R}$. We shall specify them later after we review fundamental facts about Sobolev spaces.

The interpolation operators $\Pi_{\text{grad}}$, $\Pi_{\text{curl}}$, $\Pi_{\text{div}}$ corresponding to the d.o.f. can be equivalently specified as unique operators satisfying the conditions:

\[
\Pi_{\text{grad}} u - u = 0 \quad \text{at each vertex } v ,
\]
\[
\int_e (\Pi_{\text{curl}} E - E)_t = 0 \quad \text{for each edge } e ,
\]
\[
\int_f (\Pi_{\text{div}} V - V) \cdot n_f = 0 \quad \text{for each face } f .
\]

The interpolation operator for the $L^2$ spaces is simply the $L^2$-projection, and the property:

\[
\int_K (P w - w) = 0 ,
\]

is an equivalent definition of $L^2$-projection onto constants.

Finally, note that the discussed d.o.f. guarantee not only the unisolvence conditions but the conformity of the global discretization as well. If you miss this fact, you are in big trouble.

**Whitney shape functions.** Let $a_0, a_1, a_2, a_3$ denote the vertices of the tetrahedron. Vectors $a_i - a_0$, $i = 1, 2, 3$, are linearly independent and, therefore, for each point $x \mathbb{R}^3$, there exist unique numbers (components) $\lambda_i, i = 1, 2, 3$ such that

\[
x - a_0 = \sum_{i=1}^3 \lambda_i (a_i - a_0)
\]

or, equivalently,

\[
x = (1 - \lambda_1 - \lambda_2 - \lambda_3) a_0 + \lambda_1 a_1 + \lambda_2 a_2 + \lambda_3 a_3
\]

Numbers $\lambda_0, \ldots, \lambda_3$ are identified as the affine (barycentric) coordinates of point $x$ with respect to the vertices of tetrahedron $K$. Once can show that $\lambda_i$ are linear functions of $x$ and they are invariant under affine isomorphisms, comp. Exercise 1.8.8.

The following Whitney shape functions form bases for the lowest order tetrahedron of the first type

\[
\lambda_i \quad i = 0, 1, 2, 3
\]

\[
\lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i
\]

\[
\lambda_i (\nabla \lambda_j \times \nabla \lambda_k) + \lambda_k (\nabla \lambda_i \times \nabla \lambda_j) + \lambda_j (\nabla \lambda_k \times \nabla \lambda_i)
\]

\[
1
\]

corresponding to the following degrees-of-freedom (see Exercise 1.8.9).

- $H^1$ element:

\[
\phi \rightarrow \phi(a_i), \quad i = 0, 1, 2, 3
\]
• **H(curl) element:**

\[ E \rightarrow \frac{1}{m(e_{ij})} \int_{e_{ij}} E \cdot (a_j - a_i), \quad (i, j) = (0, 1), (1, 2), (0, 2), (0, 3), (1, 3), (2, 3) \]

where \( e_{ij} \) denotes the edge from vertex \( a_i \) to vertex \( a_j \), and \( m(e_{ij}) = |a_j - a_i| \) stands for its length.

• **H(div) element:**

\[ V \rightarrow \frac{1}{m(f_{ijk})} \int_{f_{ijk}} V \cdot |(a_j - a_i) \times (a_k - a_i)|, \quad (i, j, k) = (0, 1, 2), (0, 1, 3), (1, 2, 3), (0, 2, 3) \]

where \( f_{ijk} \) denotes the face spanned by vertices \( a_i, a_j, a_k \), and \( m(f_{ijk}) = |(a_j - a_i) \times (a_k - a_i)| \) is the area of the face. Note that face normal unit vector \( n_f \) is given by:

\[ n_f = \frac{(a_j - a_i) \times (a_k - a_i)}{|(a_j - a_i) \times (a_k - a_i)|}. \]

• **L^2 element:**

\[ q \rightarrow \frac{1}{m(K)} \int_K q \]

where \( m(K) = [a_1 - a_0, a_2 - a_0, a_3 - a_0] = (a_1 - a_0) \cdot ((a_2 - a_0) \times (a_3 - a_0)) \) is the volume of the element.

Invariance of affine coordinates with respect to affine isomorphisms implies that the Whitney formulas remain valid for any tetrahedron \( K \).

**De Rham diagram. Commutativity of interpolation operators.** The following de Rham diagram communicates commuting properties of the interpolation operators.

\[
\begin{array}{cccccc}
H^1 & \xrightarrow{\nabla} & H(\text{curl}) & \xrightarrow{\nabla \times} & H(\text{div}) & \xrightarrow{\nabla} & L^2 \\
\downarrow & \Pi^{\text{grad}} & \downarrow & \Pi^{\text{curl}} & \downarrow & \Pi^{\text{div}} & \downarrow & P \\
W^p & \xrightarrow{\nabla} & Q^p & \xrightarrow{\nabla \times} & V^p & \xrightarrow{\nabla} & Y^p
\end{array}
\]

(1.67)

where \( \Pi^{\text{grad}}, \Pi^{\text{curl}}, \Pi^{\text{div}} \) are the interpolation operators and \( P \) denote the \( L^2 \)-projection.

**THEOREM 1.8.1**

The FE spaces corresponding to the lowest order tetrahedron of the first type and the corresponding interpolation operators satisfy the de Rham diagram.

**PROOF** We start with the commutativity of \( \Pi^{\text{grad}} \) and \( \Pi^{\text{curl}} \),

\[ \nabla(\Pi^{\text{grad}} u) \cong \Pi^{\text{curl}}(\nabla u). \]
As both sides live in space $Q_1$, by the shape functions reproducability property, the statement is equivalent to,

$$
\Pi^\text{curl}(\nabla(\Pi^\text{grad}u)) = \Pi^\text{curl}(\nabla u),
$$

or,

$$
\Pi^\text{curl}(\nabla(\Pi^\text{grad}u - u)) = 0.
$$

Consequently, it is sufficient to show that the $H(\text{curl})$ d.o.f. applied to $\nabla(\Pi^\text{grad}u - u)$ are zero. Consider edge $e_{ij}$ connecting vertex $a_i$ with vertex $a_j$. Set $E = \nabla(\Pi^\text{grad}u - u)$. Then

$$
\frac{1}{|a_j - a_i|} \int_{e_{ij}} (\nabla(\Pi^\text{grad}u - u)) \cdot (a_j - a_i) = \frac{1}{|a_j - a_i|} \int_0^1 \nabla(\Pi^\text{grad}u - u)(a_i + t(a_j - a_i)) |a_j - a_i| dt
$$

$$
= \int_0^1 \frac{d}{dt}(\Pi^\text{grad}u - u)(a_i + t(a_j - a_i)) dt
$$

$$
= (\Pi u - u)(a_j) - (\Pi u - u)(a_i) = 0.
$$

Done.

The second commutativity property reads as follows.

$$
\nabla \times (\Pi^\text{curl}E) = \Pi^\text{div}(\nabla \times E).
$$

Again, by the shape functions reproducability property, this is equivalent to

$$
\Pi^\text{div}(\nabla \times (\Pi^\text{curl}E - E)) = 0.
$$

Vanishing of the interpolant is equivalent to vanishing of all d.o.f., i.e., there must be

$$
\int_f \nabla \times (\Pi^\text{curl}E - E) \cdot n_f = 0,
$$

for each face $f$. But, by the Stokes Theorem, the face integral is equal to:

$$
\int_{\partial f} (\Pi^\text{curl}E - E)_t = \sum_e \int_e (\Pi^\text{curl}E - E)_t = 0,
$$

by the definition of operator $\Pi^\text{curl}$.

Finally, we have the third commutativity property,

$$
P(\nabla \cdot v) = \nabla \cdot (\Pi^\text{div}v).
$$

By the shape functions reproducability property, it is equivalent to prove that

$$
P(\nabla \cdot (\Pi^\text{div}v - v)) = 0,
$$

or,

$$
\int_K \nabla \cdot (\Pi^\text{div}v - v) = 0.
$$

But this follows immediately from the Gauss Theorem and definition of operator $\Pi^\text{div}$,

$$
\int_K \nabla \cdot (\Pi^\text{div}v - v) = \int_{\partial K} (\Pi^\text{div}v - v) \cdot n = \sum_f \int_f (\Pi^\text{div}v - v) \cdot n_f = 0.
$$
Hexahedral element of the first type. We proceed now with the definition of spaces of arbitrary order $p$.

$$W^p = \mathcal{P}^p \otimes \mathcal{P}^q \otimes \mathcal{P}^r$$

$$Q^p = (\mathcal{P}^{p-1} \otimes \mathcal{P}^q \otimes \mathcal{P}^r) \times (\mathcal{P}^p \otimes \mathcal{P}^{q-1} \otimes \mathcal{P}^r) \times (\mathcal{P}^p \otimes \mathcal{P}^q \otimes \mathcal{P}^{r-1})$$

$$V^p = (\mathcal{P}^p \otimes \mathcal{P}^{q-1} \otimes \mathcal{P}^{r-1}) \times (\mathcal{P}^{p-1} \otimes \mathcal{P}^q \otimes \mathcal{P}^{r-1}) \times (\mathcal{P}^{p-1} \otimes \mathcal{P}^q \otimes \mathcal{P}^r)$$

$$Y^p = \mathcal{P}^{p-1} \otimes \mathcal{P}^{q-1} \otimes \mathcal{P}^{r-1}$$

or, using Ciarlet notation for tensor products: $Q^{(p,q,r)} := \mathcal{P}^p \otimes \mathcal{P}^q \otimes \mathcal{P}^r$.

$$W^p = Q^{(p,q,r)}$$

$$Q^p = (Q^{(p-1,q,r)} \times Q^{(p,q-1,r)} \times Q^{(p,q,r-1)})$$

$$V^p = (Q^{(p,q-1,r-1)} \times Q^{(p-1,q,r-1)} \times Q^{(p-1,q-1,r)})$$

$$Y^p = Q^{(p-1,q-1,r-1)}$$

Note that the tensor product element allows for a different order of approximation in each direction. It is naturally an anisotropic element as opposed to the tetrahedral elements discussed next which are isotropic.

Shape functions for the lowest order hexahedron are defined as tensor product of 1D affine coordinates $\lambda_i, \mu_j, \nu_k$, $i = 0, 1$ corresponding to the three directions. The $H^1$ shape functions are defined as follows.

$$\lambda_i(x_1) \mu_j(x_2) \nu_k(x_3), \quad (i, j, k) = (0, 0, 0), (1, 0, 0), (0, 1, 0), (1, 1, 0), (0, 0, 1), (1, 0, 1), (0, 1, 1), (1, 1, 1)$$

where we use the lexicographic ordering for the vertices.

Similar expressions can be derived for the $H$ (curl), $H$ (div) and $L^2$ shape functions, comp. Exercise 1.8.11.

We use the same d.o.f. as for the lowest order tetrahedron (with orientations implied by the lexicographic rule). The de Rham diagram holds as well. Note that, due to the invariance of 1D affine coordinates wrt to 1D affine isomorphismm the formulas for the shape functions remain valid for any hexahedron (of arbitrary dimension $a \times b \times c$.

Degrees of freedom for higher order elements.

Tetrahedral element of the first kind. Spaces:

$$W^p = \mathcal{P}^p$$

$$Q^p = (\mathcal{P}^{p-1} \times \mathcal{P}^{p-1} \times \mathcal{P}^p) \oplus \mathcal{N}^p$$

$$V^p = (\mathcal{P}^{p-1} \times \mathcal{P}^{p-1} \times \mathcal{P}^{p-1}) \oplus \mathcal{T}^p$$

$$Y^p = \mathcal{P}^{p-1}$$

where

$$\mathcal{N}^p := \{ E \in \mathcal{P}^p \times \mathcal{P}^p \times \mathcal{P}^p : x \cdot E(x) = 0 \quad \forall x \}$$

$$\mathcal{T}^p := \{ x \phi(x) = \phi(x)(x_1, x_2, x_3) : \phi \in \mathcal{P}^{p-1} \}$$

with $\mathcal{P}^p$ denoting scalar-valued homogeneous polynomials of order $p$. 
Tetrahedral element of the second kind. Spaces:

\[ W^p = \mathcal{P}^p \]
\[ Q^p = \mathcal{P}^{p-1} \times \mathcal{P}^{p-1} \times \mathcal{P}^{p-1} \]
\[ V^p = \mathcal{P}^{p-2} \times \mathcal{P}^{p-2} \times \mathcal{P}^{p-2} \]
\[ Y^p = \mathcal{P}^{p-3} \]

1.8.3 Projection Based Interpolation

Discuss the fundamental principles of the concept: locality, conformity and optimality.

**H**^1 PB Interpolation

\[ H^r(K) \ni u \rightarrow \Pi^{\text{grad}} u = u_1 + u_2 + u_3 + u_4 \in W^p(K) \quad (1.68) \]

where

- \( u_1 \) is the vertex interpolant constructed using vertex shape functions \( \phi_v \):

\[ u_1(x) := \sum_v u(v) \phi_v(x), \]

- \( u_2 := \sum_e u_{2,e} \) is the edge contribution where edge \( e \) bubble \( u_{2,e} \) is a combination of edge shape functions (edge bubbles),

\[ u_{2,e} = \sum_{j=1}^{p-1} u_{2,e,j} \phi_j, \quad \phi_j \in \mathcal{P}_e^p(e), \]

and it is obtained by solving the edge projection problem:

\[ \| \frac{\partial}{\partial t} (u - (u_1 + u_{2,e})) \|_{L^2(e)} \rightarrow \min. \]

- \( u_3 := \sum_f u_{3,f} \) is the face contribution where face \( f \) bubble \( u_{3,f} \) is a combination of face shape functions (face bubbles),

\[ u_{3,f} = \sum_j u_{3,f,j} \phi_j, \]

and it is obtained by solving the face projection problem:

\[ \| \nabla_t (u - (u_1 + u_2 + u_{3,f})) \|_{L^2(f)} \rightarrow \min. \]

- \( u_4 \) is the element bubble obtained by projecting difference \( u - u_1 - u_2 - u_3 \) over the element bubbles,

\[ \| \nabla (u - (u_1 + u_2 + u_3 + u_4)) \|_{L^2(K)} \rightarrow \min. \]
Above, $\partial/\partial t$ denotes the tangential derivative along the edge and $\nabla_t$ stands for the tangential component of the gradient. Equivalent variational statements are:

\[
\int_e \frac{\partial}{\partial t} (u - (u_1 + u_{2,e})) \frac{\partial \varphi}{\partial t} = 0 \quad \text{for each edge bubble } \varphi, \\
\int_f \nabla_t (u - (u_1 + u_2 + u_{3,f})) \cdot \nabla_t \varphi = 0 \quad \text{for each face bubble } \varphi, \\
\int_K \nabla (u - (u_1 + u_2 + u_3 + u_4)) \cdot \nabla \varphi = 0 \quad \text{for each element bubble } \varphi.
\]

Equivalent definition of the interpolant:

\[
(u - u_p)(v) = 0 \quad \text{for each vertex } v, \\
\int_e \frac{\partial}{\partial t} (u - u_p) \frac{\partial \varphi}{\partial t} = 0 \quad \text{for each edge bubble } \varphi, \quad \text{for each edge } e, \\
\int_f \nabla_t (u - u_p) \cdot \nabla_t \varphi = 0 \quad \text{for each face bubble } \varphi, \quad \text{for each face } f, \\
\int_K \nabla (u - u_p) \cdot \nabla \varphi = 0 \quad \text{for each element bubble } \varphi.
\]

\textbf{\textit{H} (curl) PB Interpolation}

\[
H^{r,s}(\text{curl}, K) \ni E \to \Pi^{\text{curl}} E = E_p = E_1 + E_2 + E_3 \in Q^p
\]

Here:

- $E_1 = \sum_e E_{1,e}$ is the edge interpolant. Each edge $e$ contribution $E_{1,e}$ lives in the span of edge $e$ shape functions and it is obtained by solving the edge projection problem:

\[
\|(E - E_{1,e})_t\|_{L^2(e)} \to \min
\]

where $E_t$ denotes the tangential component of vector $E$.

- $E_2 = \sum_f E_{2,f}$, with each face contribution $E_{2,f}$ living in the span of face shape functions (face bubbles) and being the solution of the constrained projection problem:

\[
\begin{cases} 
\|\nabla_t \times (E - E_1 - E_{2,f})\|_{L^2(f)} \to \min \\
((E - E_1 - E_{2,f})_t, \nabla_t \varphi)_{L^2(f)} = 0 \quad \text{for each } H^1 \text{bubble } \varphi.
\end{cases}
\]

- $E_3$ lives in the span of element $H(\text{curl})$ bubbles, and is the solution of the constrained projection problem:

\[
\begin{cases} 
\|\nabla \times (E - E_1 - E_2 - E_3)\|_{L^2(K)} \to \min \\
((E - E_1 - E_2 - E_3), \nabla \varphi)_{L^2(K)} = 0 \quad \text{for each } H^1 \text{bubble } \varphi.
\end{cases}
\]
Equivalent definition of the interpolant:

\[
\int_e (E - E_p) \psi_e = 0 \quad \text{for each edge shape function } \psi, \\
\int_f \nabla \times (E - E_p) \cdot \nabla \times \psi = 0 \quad \text{for each } H(\text{curl}) \text{ face bubble } \psi, \\
\int_f (E - E_p) \cdot \nabla \psi = 0 \quad \text{for each } H^1 \text{ face bubble } \varphi, \\
\int_K \nabla \times (E - E_p) \cdot \nabla \psi = 0 \quad \text{for each element } H(\text{curl}) \text{ bubble } \psi, \\
\int_K (E - E_p) \cdot \nabla \varphi = 0 \quad \text{for each element } H^1 \text{ bubble } \varphi.
\]

(1.71)

**H (div) PB Interpolation**

\[H^{r-s}(\text{div}, K) \ni v \rightarrow \Pi^{\text{div}} v = v_p = v_1 + v_2 \in V^p\]

(1.72)

Here:

- \(v_1 = \sum_f v_{1,f}\) is the face interpolant. Each face contribution \(v_{1,f}\) lives in the span of the face shape functions and is the solution of the projection problem:

\[\| (v - v_{1,f}) \cdot n \|_{L^2(f)} \rightarrow \min\]

- \(E_3\) lives in the span of element \(H(\text{div})\) bubbles and is the solution of the constrained projection problem:

\[
\left\{ \begin{array}{l}
\| \nabla \cdot (v - v_1 - v_2) \|_{L^2(K)} \rightarrow \min \\
((v - v_1 - v_2), \nabla \times \varphi)_{L^2(K)} = 0 \quad \text{for each element } H(\text{curl}) \text{ bubble } \varphi.
\end{array} \right.
\]

Equivalent definition of the interpolant:

\[
\int_f ((v - v_p) \cdot n) \psi \cdot n = 0 \quad \text{for each } H(\text{div}) \text{ face shape function } \psi, \\
\int_K \nabla \cdot (v - v_p) \nabla \cdot \psi = 0 \quad \text{for each element } H(\text{div}) \text{ bubble } \psi, \\
\int_K (v - v_p) \cdot \nabla \times \varphi = 0 \quad \text{for each element } H(\text{curl}) \text{ bubble } \varphi.
\]

(1.73)

**\(L^2\) Projection**

\[L^2(K) \ni f \rightarrow Pf = f_p \in Q^p\]

(1.74)

where

\[\int_K (f - f_p) \psi = 0 \quad \text{for each shape function } \psi.\]
1.8.4 Parametric Elements and Piola Transforms (Pullback Maps)

The idea of parametric element can be generalized to the rest of the exact sequence energy spaces. Given the exact sequence for a master element $\hat{K}$, we seek transforms (pullbacks) for the energy spaces defined over an arbitrary (possibly curvilinear) physical element $K$ that will make the following diagram commute.

$$
\begin{align*}
H^1(\hat{K}) & \xrightarrow{\nabla} H(\text{curl}, \hat{K}) \xrightarrow{\nabla} H(\text{div}, \hat{K}) \xrightarrow{\nabla} L^2(\hat{K}) \\
\downarrow T^{\text{grad}} & \downarrow T^{\text{curl}} \downarrow T^{\text{div}} \downarrow T^{L^2} \\
H^1(K) & \xrightarrow{\nabla} H(\text{curl}, K) \xrightarrow{\nabla} H(\text{div}, K) \xrightarrow{\nabla} L^2(K)
\end{align*}
$$

A general (physical) element $K$ is the image of master element $\hat{K}$ by an element map $x_K: \hat{K} \ni \xi \rightarrow x = x_K(\xi) \in K$ that we assume to be a $C^1(\hat{K})$-diffeomorphism, i.e. the map is a bijection, and derivatives of both $x_K$ and its inverse $x_K^{-1}$ exist and are continuous up to the boundary. The first $T^{\text{grad}}$ map has already been defined,

$$
T^{\text{grad}}: H^1(\hat{K}) \ni \hat{u} \rightarrow u \in H^1(K), \quad u(x) := \hat{u}(x_K^{-1}(x)) \quad \text{or} \quad u = \hat{u} \circ x_K^{-1} \quad \text{or} \quad \hat{u} = u \circ x_K.
$$

Using an engineering notation,

$$
u(x) = \hat{u}(\xi(x)) \quad \text{or} \quad \hat{u}(\xi) = u(x(\xi)).$$

Definition of the remaining maps is a consequence of the commutativity of the pullback maps. The transformation $T^{\text{curl}}$ must apply in particular to gradients so we can find it out by computing $\nabla u$,

$$
\frac{\partial u}{\partial x_j} = \frac{\partial \hat{u}}{\partial \xi_i} \frac{\partial \xi_i}{\partial x_j}.
$$

This leads to the transform for the $H(\text{curl})$ space:

$$
E_j(x) = \hat{E}_i(\xi(x)) \frac{\partial \xi_i}{\partial x_j}(x) \quad \text{or} \quad E = J^{-T} \hat{E} \circ x_K^{-1}
$$

where $J = \frac{\partial x_i}{\partial \xi_j}$ denotes the Jacobian matrix of the element map. The objects with hats are always functions of $\xi$ and the objects without hats depend upon $x$. This leads to the simplified notation:

$$
T^{\text{curl}}: H(\text{curl}, K) \ni \hat{E} \rightarrow E \in H(\text{curl}, K) \quad \text{where} \quad E = J^{-T} \hat{E}.
$$

It goes without saying that the right-hand side must be composed with $x_K^{-1}$ or the left-hand side must be composed with $x_K$.

The next transformation is determined by computing curl $E$.

$$
\begin{align*}
(\text{curl } E)_i &= \epsilon_{ijk} \frac{\partial E_k}{\partial x_j} = \epsilon_{ijk} \frac{\partial}{\partial x_j} (\hat{E}_l \frac{\partial \xi_l}{\partial x_k}) \\
&= \epsilon_{ijk} \frac{\partial \hat{E}_l}{\partial x_j} \frac{\partial \xi_l}{\partial x_k} + \epsilon_{ijk} \hat{E}_l \frac{\partial^2 \xi_l}{\partial x_j \partial x_k} \quad \text{(product of a symmetric and an antisymmetric matrix = 0)} \\
&= \epsilon_{ijk} \frac{\partial \hat{E}_l}{\partial \xi_m} \frac{\partial \xi_l}{\partial x_j} \frac{\partial \xi_k}{\partial x_k} = (*)
\end{align*}
$$
Recall now the definition of inverse Jacobian $j^{-1}$ (determinant of inverse Jacobian matrix $J^{-1}$),

$$
\epsilon_{ijk} \frac{\partial \xi_1}{\partial x_i} \frac{\partial \xi_2}{\partial x_j} \frac{\partial \xi_3}{\partial x_k} = j^{-1}
$$
or, more generally,

$$
\epsilon_{ijk} \frac{\partial \xi_\alpha}{\partial x_i} \frac{\partial \xi_\beta}{\partial x_j} \frac{\partial \xi_\gamma}{\partial x_k} = j^{-1} \epsilon_{\alpha\beta\gamma}.
$$

Multiplying both sides by $\frac{\partial x_l}{\partial \xi_\alpha}$, we get,

$$
\epsilon_{ijk} \frac{\partial x_l}{\partial \xi_\alpha} \frac{\partial \xi_\alpha}{\partial x_i} \frac{\partial \xi_\beta}{\partial x_j} \frac{\partial \xi_\gamma}{\partial x_k} = j^{-1} \epsilon_{\alpha\beta\gamma} \frac{\partial x_l}{\partial \xi_\alpha}
$$
or

$$
\epsilon_{ijk} \frac{\partial \xi_\beta}{\partial x_j} \frac{\partial \xi_\gamma}{\partial x_k} = j^{-1} \epsilon_{\alpha\beta\gamma} \frac{\partial x_l}{\partial \xi_\alpha}.
$$

In particular, differentiating both sides wrt $x_l$, we learn that

$$
\frac{\partial}{\partial x_l} (j^{-1} \frac{\partial x_l}{\partial \xi_\alpha}) = \epsilon_{\alpha\beta\gamma} \frac{\partial^2 \xi_\beta}{\partial x_j \partial x_l} \frac{\partial \xi_\gamma}{\partial x_k} + \epsilon_{ijk} \frac{\partial \xi_\beta}{\partial x_j} \frac{\partial \xi_\gamma}{\partial x_k} \frac{\partial x_l}{\partial \xi_\alpha} = 0
$$
as the product of a symmetric and an unsymmetric matrix must vanish.

Returning to our computation of curl $E$, we get,

$$
(*) = \epsilon_{\alpha\beta\gamma} j^{-1} \frac{\partial x_l}{\partial \xi_\alpha} \frac{\partial \hat{E}_l}{\partial \xi_\beta} \frac{\partial \hat{E}_m}{\partial \xi_\gamma} = j^{-1} \frac{\partial x_l}{\partial \xi_\alpha} \frac{\partial \hat{E}_l}{\partial \xi_\beta} \frac{\partial \hat{E}_m}{\partial \xi_\gamma} = j^{-1} \frac{\partial x_l}{\partial \xi_\alpha} \left( \frac{\partial (\text{curl} \hat{E})}{\partial \xi_\alpha} \right).
$$

This leads to the transformation rule for the $H(\text{div})$ fields,

$$
T_{\text{div}} : H(\text{div}, \hat{K}) \ni \hat{E} \rightarrow E \in H(\text{div}, K) \quad \text{where} \quad H_j = j^{-1} \frac{\partial x_l}{\partial \xi_\alpha} \hat{H}_l \quad \text{or} \quad H = j^{-1} J \hat{H}.
$$

Finally, we need to compute div $H$,

$$
\text{div } H = \frac{\partial H_j}{\partial x_i} = \frac{\partial}{\partial x_i} \left( j^{-1} \frac{\partial x_l}{\partial \xi_\alpha} \right) \hat{H}_l + j^{-1} \frac{\partial x_l}{\partial \xi_\alpha} \frac{\partial \hat{H}_l}{\partial x_i} = j^{-1} \frac{\partial x_l}{\partial \xi_\alpha} \frac{\partial \hat{H}_l}{\partial x_i} \frac{\partial \xi_\beta}{\partial x_i} \frac{\partial \xi_\gamma}{\partial x_k} = j^{-1} \frac{\partial \hat{H}_l}{\partial \xi_\alpha} = j^{-1} \frac{\partial (\text{div} \hat{H})}{\partial \xi_\alpha}
$$
where the underbraced term vanishes by setting $\beta = 2, \gamma = 3$ in (1.76). The last transformation formula for the $L^2$ fields reads thus as follows,

$$
T^{A^2} : L^2(\hat{K}) \ni \hat{E} \rightarrow E \in L^2(K) \quad \text{where} \quad f = j^{-1} \hat{f}.
$$

The pullback map for the $H(\text{div})$ fields is known in mechanics as the Piola transform which has motivated me to extend this name to all of the transforms.

Note that, with the regularity assumptions made on the element map, all Piola transforms are well defined, i.e. they preserve the energy spaces. We make now some crucial observations concerning conformity. Begin with a simple observation that the global $C^0$-continuity of the union of element maps and the continuity of functions $\hat{u}$ in the parametric domain, implies the global continuity of the corresponding functions $u$ in
the physical domain. If two sufficiently regular functions are continuous along a curve, the corresponding tangential derivative must be the same. As the Piola transform $T_{\text{curl}}$ was derived by computing the gradients, we expect that the continuity of tangential components of $H^\text{curl}$ fields will be preserved as well. This is indeed the case. Consider a curve in the parametric domain parametrized with

$$\xi_k = \xi_k(t), \quad t \in [0, 1].$$

The image of the curve through the element map is naturally parametrized with the composition of the parametrization in the parametric domain and the element map,

$$x_j = x_j(\xi_k(t)), \quad t \in [0, 1].$$

Computing the tangent component of $H(\text{curl}) \, E$ field,

$$\frac{\partial x_j}{\partial \xi_k} \frac{\partial \xi_k}{\partial t} E^i_j = \frac{\partial x_j}{\partial \xi_k} \frac{\partial E^i_k}{\partial t} = \frac{\partial E^i_j}{\partial t},$$

we obtain the tangent component of field $\hat{E}$ in the parametric domain. Equivalently,

$$E^i ds = \hat{E}^i ds_0$$

where $ds, ds_0$ stand for the length of the tangent vectors before the normalization. The Piola map preserves tangent components, and the tangential component of $E$ along the curve in the physical domain depends only upon the restriction of the element map to the corresponding curve in the parametric domain. Now comes the main point. If the union of element maps is globally continuous ($C^0$ continuity is enough) then $H(\text{curl})$-conforming functions in the parametric domain are mapped into $H(\text{curl})$-conforming functions in the physical domain.

A similar result holds for the $H(\text{div})$ fields. We begin again with the formula for the determinant,

$$\frac{\epsilon_{ijk}}{\partial \xi_\alpha} \frac{\partial x_j}{\partial \xi_\alpha} \frac{\partial x_k}{\partial \xi_\gamma} = j \epsilon_{\alpha\beta\gamma}.$$

This implies,

$$\frac{\epsilon_{ijk}}{\partial \xi_\alpha} \frac{\partial x_j}{\partial \xi_\alpha} \frac{\partial \xi_\beta}{\partial s} \frac{\partial x_k}{\partial \xi_\gamma} \frac{\partial \xi_\gamma}{\partial t} = j \epsilon_{\alpha\beta\gamma} \frac{\partial \xi_\beta}{\partial s} \frac{\partial \xi_\gamma}{\partial t},$$

where $\xi_\beta(s)$ and $\xi_\gamma(t)$ are parametrization of two curves in a surface $\hat{S}$ in the parametric domain. As $x_j(\xi_\beta(s))$ and $x_k(\xi_\gamma(t))$ are parametrizations of the corresponding surface in the physical domain, and cross product of two tangent vectors to a surface gives a normal to the surface, we obtain the relation between normal vectors for $\hat{S}$ and the corresponding image surface $S$,

$$\frac{\partial x_j}{\partial \xi_\alpha} n_\alpha dS = j \hat{n}_\alpha dS_0,$$

or,

$$n_\alpha dS = j \frac{\partial \xi_\alpha}{\partial x_l} \hat{n}_\alpha dS_0.$$
where \( \hat{n}, n \) are now the unit vectors and \( dS_0, dS \) denote the length of normal vectors before normalization. This implies now the relation between normal components of \( H(\text{div}) \) fields in the parametric and physical domains,

\[
n_l H_l \, dS = j \frac{\partial \xi_\alpha}{\partial x_l} \hat{n}_\alpha \, j^{-1} \frac{\partial x_l}{\partial \xi_\beta} \hat{H}_\beta \, dS_0 = \hat{n}_\alpha \hat{H}_\alpha \, dS_0.\]

Consequently, normal components are preserved which implies that the Piola transform maps \( H(\text{div}) \)-conforming fields in the parametric domain into \( H(\text{div}) \)-conforming fields in the physical domain.

**Exercises**

**Exercise 1.8.1** Lagrange square element.

(i) Draw master quad of order \((3, 4)\) and the corresponding Lagrange nodes. Use elementary means to construct shape functions corresponding to a sample vertex, edge and interior node. Check that they are in the space of element shape functions.

(ii) Use the Lagrange shape functions to prove the unisolvency condition. Note that this technique mathematically awkward as the shape functions are supposed to be defined after the unisolvency is established. Can you think of alternate ways to prove the unisolvency without using the shape functions?

(iii) Think of possible ways to modify the location of the Lagrangian nodes to keep the unisolvency condition intact.

(iv) Consider two master elements sharing an edge and assume the order of the elements in such a way that the restrictions of element shape functions to the common edge live in the same polynomial space. Explain why matching the dof (pointwise values) at the common edge Lagrangian nodes implies global continuity of functions obtained by “gluing” shape functions defined on the two elements (the mathematical term is unions of shape functions).

(v) Going back to question asked in Step (iii), is the location of Lagrangian nodes and their number at vertices, edges and interior essential for enforcing the global continuity? Discuss possible modifications to the location of Lagrangian nodes that would not destroy the global continuity result.

(5 points)

**Exercise 1.8.2** Lagrange triangular element. Repeat the steps of Exercise 1.8.1 for the master triangle of order 4.

(3 points)

**Exercise 1.8.3** Lagrange three-dimensional element. Pick your favorite 3D element and repeat for it the steps of Exercise 1.8.1.

(3 points)
**Exercise 1.8.4** Parametric Lagrange element.

(i) Prove the unisolvency for an arbitrary parametric Lagrange element. Discuss why it is necessary for the element map to remain bijective in the element closure \( \overline{K} \) (this eliminates the possibility of singular maps like Duffy’s map).

(ii) Assume you have two physical 2D Lagrange elements \( K_1, K_2 \) sharing an edge \( e \). Let \( x_{K_i} \) be the corresponding element maps defined on master elements \( \tilde{K}_i, i = 1, 2 \). Discuss sufficient conditions on master element space \( X(\tilde{K}_i) \) and the element maps that would guarantee the global continuity of unions of FE shape functions.

(5 points)

**Exercise 1.8.5** Alternative degrees of freedom. Consider your favorite 3D Lagrangian element of an arbitrary order, and replace the Lagrangian dof with a new set of degrees of freedom defined by using edge, face and element moments:

- **vertex dof:** \( u \rightarrow (v) \quad \forall \text{ vertex } v \)
- **edge dof:** \( u \rightarrow \int_e u f^e_i \quad i = 1, \ldots, ? \quad \forall \text{ edge } e \)
- **face dof:** \( u \rightarrow \int_f u f^f_i \quad i = 1, \ldots, ? \quad \forall \text{ face } f \)
- **interior dof:** \( u \rightarrow \int_K u f^K_i \quad i = 1, \ldots, ? \)

Discuss the number of edge, face and interior moments necessary for enforcing the global continuity. Provide a concrete example of weights \( f^e_i, f^f_i, f^K_i \) with which the element satisfies the unisolvency condition.

(5 points)

**Exercise 1.8.6** Polynomial exact sequences. Prove that the discussed polynomial sequences for the hexahedral and tetrahedral elements are exact.

(5 points)

**Tetrahedron of the first type.**

\[
\mathcal{P}^p \xrightarrow{\nabla} (\mathcal{P}^{p-1})^3 + \mathcal{N}^p \xrightarrow{\nabla} (\mathcal{P}^{p-1})^3 + \mathcal{RT}^p \xrightarrow{\nabla} \mathcal{P}^{p-1}
\]

where

\[
\mathcal{N}^p := \{ E \in (\mathcal{P})^3 : x \cdot E(x) = 0 \quad \forall x \}
\]

and

\[
\mathcal{RT}^p := \mathcal{P}_x = \{ \phi(x)x : \phi \in \mathcal{P} \}.
\]

As the differentiation lowers the (total) polynomial degree by one, the sequence if well defined and it automatically inherits the structure of the differential complex, i.e. the range of each operator is in the null space of the next operator in the sequence.
In the proof of the exactness of the sequence, the right inverses of grad, curl, div operators come handy, see Section ??.

Let \( E = E_{p-1} + \tilde{E}_p \) where \( E_{p-1} \in (P^{p-1})^3 \) and \( \tilde{E}_p \in \mathcal{N}^p \). Assume that \( \nabla E = 0 \).

According to the right inverse formula,

\[
E_{p-1} + \tilde{E}_p = \nabla(GE_{p-1} + G\tilde{E}_p) = \nabla(GE_{p-1}) = E_{p-1}
\]

which proves that \( \tilde{E}_p = 0 \) and \( E = E_{p-1} \) is the gradient of \( GE_{p-1} \in \mathcal{P}^p \).

Similarly, let \( v = v_{p-1} + \tilde{v}_p \) where \( v_{p-1} \in (P^{p-1})^3 \) and \( \tilde{v}_p \in \mathcal{RT}^p \). Assume that \( \nabla \cdot v = 0 \). By the right inverse formula then,

\[
v_{p-1} + \tilde{v}_p = \nabla \times (Kv_{p-1} + K\tilde{v}_p) = \nabla \times (Kv_{p-1}) = v_{p-1}
\]

which proves that \( \tilde{v}_p \) component vanishes and \( v \) is the curl of a polynomial from \( (P^{p-1})^3 \).

Finally, surjectivity of div operator follows directly from the existence of the right inverse.

**Hexahedron of the first type.**

As in the previous case, the sequence if well defined and it automatically inherits the structure of the differential complex. In order to prove that the sequence is exact, you can again utilize the right inverse operators \( G, K, D \). All that you have to notice is that each of them is actually well defined on the whole space, i.e.

\[
G(Q^p) \subset W^p, \quad K(V^p) \subset Q^p, \quad D(Y^p) \subset V^p.
\]

**Exercise 1.8.7** 2D elements. Given the 3D exact polynomial sequences, write out the corresponding two 2D exact polynomial sequences for the square and triangular elements (a total of six sequences) and prove that they are exact. Be concise.

(5 points)

**Exercise 1.8.8** Affine coordinates. Prove the following facts about the affine coordinates:

- The affine coordinates are independent of the enumeration of vertices (in the presented construction, we considered vectors \( x - a_0, a_i - a_0, i = 1, 2, 3 \), so it looks like things might depend upon the choice of vertex \( a_0 \)).

- The affine coordinates are invariant under affine transformations: if \( \lambda_i \) are affine coordinates of a point \( x \) with respect to vertices \( a_i \) then \( \lambda_i \) are also affine coordinates of a point \( Tx \) with respect to vertices \( Ta_i \), for any bijective affine map \( T \).

- In 2D, the affine coordinates may be interpreted as area coordinates. Prove that

\[
\lambda_i = \frac{\text{area of } T_i}{\text{area of } T}, \quad i = 0, 1, 2
\]

where subtriangles \( T_i \) of triangle \( T \) are defined in Fig. 1.2.
Exercise 1.8.9 Whitney shape functions (1.8.2). Prove that the Whitney shape functions indeed represent the dual bases corresponding to the d.o.f. specified in the text.

(3 points)

Exercise 1.8.10 Whitney shape functions (1.8.2).

We start with \( H(\text{curl}) \) shape functions. The \( H^1 \) shape functions are standard linear Lagrange shape functions. Consider, for instance, edge 01,

\[
E_{01} := \lambda_0 \nabla \lambda_1 - \lambda_1 \nabla \lambda_0 ,
\]

and focus on the first term. Factor \( \lambda_0 \) vanishes along all edges that do not contain vertex \( a_0 \), i.e. edges: 12, 13, 23. By the same token, \( \lambda_1 \) vanishes along edges 02, 03, 23. Consequently, tangential component of \( \nabla \lambda_1 \) vanishes along edges 02, 03, 23, as well. The same argument applies to the second term, with the roles of \( \lambda_0, \lambda_1 \) being switched. Let \( x = a_0 + s(a_1 - a_0), s \in [0, 1] \), be now a point on the edge. We have,

\[
(\nabla \lambda_1)(a_0 + s(a_1 - a_0)) \cdot (a_1 - a_0) = (D(a_1 - a_0) \lambda_1)(a_0 + s(a_1 - a_0))
= \frac{d}{dt} \lambda_1(a_0 + (s + t)t(a_1 - a_0))|_{t=0}
= \frac{d}{dt}(s + t) = 1 .
\]

Similarly, \( (\nabla \lambda_0) \cdot (a_1 - a_0) = -1 \) and, consequently,

\[
E_{01} \cdot (a_1 - a_0) = \lambda_0 + \lambda_1 = 1 \quad \text{on edge 01} ,
\]

which proves that the corresponding degree-of-freedom (average value of the tangential derivative) is indeed equal one.
Exercise 1.8.11 Shape functions for the lowest order hexahedron. Write out shape functions for the lowest order hexahedron in terms of 1D affine coordinates and their derivatives. Prove that they provide dual bases to the standard d.o.f. with properly introduced orientations for edges and faces.

(5 points)

Exercise 1.8.12 Characterization of Nedelec’s space. Let $\tilde{P}^k$ denote homogeneous polynomials of order $k$.

Prove the following identity.

$$x \times (\tilde{P}^{p-1})^3 = \{E \in (\tilde{P}^p)^3 : x \cdot E(x) = 0 \ \forall x\}$$

(5 points)

Inclusion “⊂” is trivial. Take $E(x)$ from the right-hand side. We have,

$$E_i(x) = \frac{x_2 E_2(x)}{x_1} - \frac{x_3 E_3(x)}{x_1}$$  \hspace{1cm} (1.77)

Consider the representation,

$$E_i(x) = F_i(x_2, x_3) + x_1 G_i(x), \quad i = 1, 2$$

where $F_i, G_i$ are homogeneous polynomials of order $p$ and $p - 1$, resp. According to representation (1.77), there must exist a homogeneous polynomial $\psi(x_2, x_3)$ of order $p - 1$ such that,

$$F_1(x_2, x_3) = x_3 \psi(x_2, x_3), \quad F_2(x_2, x_3) = -x_2 \psi(x_2, x_3)$$

Consequently,

$$E_1(x) = -x_2 G_2(x) - x_3 G_3(x), \quad E_2(x) = x_3 \psi(x) + x_1 G_2(x), \quad E_3(x) = -x_2 \psi(x) + x_1 G_3(x),$$

i.e.

$$(E_1, E_2, E_3) = (x_1, x_2, x_3) \times (\psi, G_3, -G_2).$$

Exercise 1.8.13 Prismatic element. Given the exact sequences for the triangle and the 1D sequence for a unit interval, construct two exact sequences for the prism starting with $W_p = \mathcal{P}^p(T) \otimes \mathcal{P}^q(I)$ where $T$ is a triangle and $I$ an interval.

(5 points)

Exercise 1.8.14 $H^1$ PB interpolation.

(i) Discuss shortly why the three formulations in the text are equivalent.

(ii) Recall the Ciarlet definition of the interpolation operator defined in terms of d.o.f. $\psi_j$,

$$\Pi u = \sum_{j=1}^{N} \psi_j(u) \phi_j,$$
and prove that it is equivalent to the condition:

$$\Pi u \in X(K), \quad \psi_j(u - \Pi u) = 0, \quad j = 1, \ldots, N.$$ \(\text{Here } N = \dim X(K) \text{ and } \phi_j \text{ are the shape functions corresponding to degrees-of-freedom } \psi_j.\)

(iii) Based on characterization (1.69), write out the formulas for the d.o.f. corresponding to the PB interpolation.

(iv) Write down explicitly systems of linear equations that need to be solved for computing the edge, face and interior contributions to the interpolant on a tetrahedral element of order \(p\).

(v) Discuss in a couple of lines why the definition of the PB interpolation holds for all \(H^1\)-conforming elements including elements of variable order.

(vi) Is the use of hierarchical shape functions necessary for computing the PB interpolant? Discuss.

(vii) While it is natural to use the shape functions to extend \(u_1, u_2, e, u_3, f\) to the whole element, the final interpolant \(u_p\) is independent of particular lifts as long as they live in the FE space \(X(K) = W_p(K)\). Explain, why?

(10 points)

**Exercise 1.8.15** Coding \(H^1\) PB interpolation. The PB interpolant is computed by solving sequentially small systems of linear equations over element edges, faces and interiors. Suppose you would like to simplify the logic of implementation by solving a single system of linear equations for one element at a time. Try to write down such a system of equations for a 2D triangular element of order \(p\).

(5 points)

**Exercise 1.8.16** What are the minimum regularity assumptions for the PB interpolation to be continuous? In other words, what is the minimum \(r\) in (1.68)? *Hint:* Recall Trace and Sobolev Embedding Theorems.

(3 points)

**Exercise 1.8.17** \(H(\text{curl})\) PB interpolation.

(i) Write down the variational form of the constrained projection problems. Are the corresponding Lagrange multipliers equal zero?

(ii) Following the ideas from Exercise 1.8.14, identify the degrees-of-freedom corresponding to the PB interpolation operator.

(5 points)

**Exercise 1.8.18** Commutativity of PB interpolation.

(i) Assume that field \(E\) is a gradient, \(E = \nabla u\) and prove that so must be the PB interpolant, \(E_p = \nabla u_p\) where \(u_p \in W_p(K)\).
(ii) Prove that \( u_p = \Pi \nabla u \). \textit{Hint:} Reduce the definition to the case when both \( E \) and \( E_p \) are gradients and compare it with the definition of \( H^1 \) interpolant. Recall the discussion for the lowest order Whitney elements.

(10 points)

**Exercise 1.8.19** Commutativity of PB interpolation (continued). Prove the commutativity of the remaining two blocks in the diagram.

(10 points)

### 1.9 Classical Interpolation Theory

In this section we develop classical \( h \)-interpolation error estimates for the exact sequence energy spaces. For simplicity, we shall restrict ourselves to the sequences of first type only, i.e.

\[
W^p \xrightarrow{\nabla} Q^p \xrightarrow{\nabla \times} V^p \xrightarrow{\nabla \cdot} Y^p
\]

\( \mathcal{P}^p \subset W^p, \quad (\mathcal{P}^{p-1})^N \subset Q^p, \quad (\mathcal{P}^{p-1})^N \subset V^p, \quad \mathcal{P}^{p-1} \subset Y^p \)

Notice that symbol \( p \) in the notation for the space indicates the order of \( H^1 \) element only. The remaining spaces contain complete polynomials of order less or equal \( p - 1 \) only.

In each case, we assume silently that the interpolation operator commutes with the pullback (Piola) transform (“breaking the hat property”), i.e.

\[
\hat{\Pi} u = \hat{\Pi} \hat{u}
\]

We also assume silently that parameter \( r \) specifying the Sobolev regularity of the interpolated function is sufficiently large to assure the continuity of the interpolation operator.

#### 1.9.1 Bramble-Hilbert Argument

We begin with a version of the Poincaré lemma.

**Lemma 1.9.1**

Let \( \Omega \subset \mathbb{R}^N, N = 1, 2, \ldots \) be a domain. There exists a positive constant \( C = C(\Omega) \) such that

\[
\|u\|^2 \leq C \left\{ \left| \int_\Omega u \right|^2 + \| \nabla u \|^2 \right\} \quad \forall u \in H^1(\Omega)
\]  
(1.78)
**LEMMA 1.9.2**

There exists \( C > 0 \) such that

\[
\| u \|_{H^r(\Omega)}^2 \leq C \left\{ \sum_{|\alpha| \leq r-1} \left| \int_{\Omega} D^\alpha u \right|^2 + \| u \|_{H^r(\Omega)}^2 \right\}
\]  

(1.79)

for any integer \( r > 0 \).

**PROOF** Use Lemma 1.9.1 and mathematical induction.

**LEMMA 1.9.3**

There exists \( C > 0 \) such that

\[
\inf_{\varphi \in P_{r-1}} \| u - \varphi \|_{H^r(\Omega)}^2 \leq C\| u \|_{H^r(\Omega)}^2
\]

(1.80)

for any integer \( r > 0 \).

**PROOF** Apply inequality (1.79) to difference \( u - \varphi \),

\[
\| u - \varphi \|_{H^r(\Omega)}^2 \leq C \left\{ \sum_{|\alpha| \leq r-1} \left| \int_{\Omega} D^\alpha (u - \varphi) \right|^2 + \| u \|_{H^r(\Omega)}^2 \right\}
\]

Note that the \( r \)-order derivatives for \( r - 1 \) order polynomial \( \varphi \) vanish, hence absence of \( \varphi \) in the seminorm on the right-hand side. It remains to show that we can select a polynomial \( \varphi \) in such a way that all averages on the right-hand side vanish. Start by noticing that all derivatives \( D^\alpha \varphi \) of highest order, i.e. \( |\alpha| = r - 1 \) are constants. We can match these constants with the corresponding averages of derivatives of function \( u \),

\[
|\Omega| D^\alpha \varphi = \int_{\Omega} D^\alpha u
\]

Next, represent \( \varphi \) as sum of the monomials,

\[
\varphi = \sum_{|\alpha| \leq r-1} c_\alpha x^\alpha
\]

All constants \( c_\alpha \), for \( |\alpha| = r - 1 \), have been selected and we can apply now the same argument to constants corresponding to monomials of one order less,

\[
|\Omega| D^\alpha x^\alpha = \int_{\Omega} D^\alpha (u - \sum_{|\beta| = r-1} c_\beta x^\beta), \quad |\alpha| = r - 2
\]

Proceed by induction to finish the proof.
COROLLARY 1.9.1

Seminorm $|\cdot|_{H^r(\Omega)}$ provides an equivalent norm for the quotient space $H^r(\Omega)/P^{r-1}$. In particular, the quotient space equipped with that (semi)norm is complete. Following the same line of argument, we can claim also a more general result for any space of shape functions $W_p$ that contains $P^{p-1}$. Replacing $u$ with $u-\varphi$, $\varphi \in W_p$ in inequality (1.80), and taking infimum wrt to $\varphi \in W_p$ on both sides, we get,

$$\inf_{\varphi \in W_p} \|u - \varphi\|_{H^r(\Omega)}^2 \leq C \inf_{\varphi \in W_p} |u - \varphi|_{H^r(\Omega)}^2$$

(1.81)

The right-hand side represents thus a norm equivalent to the standard norm in the quotient space $H^r(\Omega)/W_p$.

We arrive at the fundamental result of Bramble and Hilbert.

THEOREM 1.9.1

(Bramble-Hilbert Argument for $H^r$ norm)

Let $\Omega$ be a domain in $\mathbb{R}^N$, $N = 1, 2, \ldots$ and let $W_p$ be a subspace of $H^1(\Omega)$ such that

$$\mathcal{P}^p \subset W_p$$

(1.82)

for some $p \geq 0$. Let $r > 0$ and let $p + 1 \geq r$. There exists a constant $C > 0$, dependent upon $r$, such that

$$\inf_{\varphi \in W_p} |u - \varphi|_{H^r(\Omega)} \leq C |u|_{H^r(\Omega)}$$

(1.83)

for every $u \in H^r(\Omega)$.

PROOF Notice that

$$\inf_{\varphi \in W_p} |u - \varphi|_{H^r(\Omega)} \leq |u|_{H^r(\Omega)}$$

and apply inequality (1.81).

THEOREM 1.9.2

(Bramble-Hilbert Argument for $H(\text{curl})$ norm)

Let $\Omega$ be a domain in $\mathbb{R}^3$ and let $Q_p$ be a subspace of $H(\text{curl}, \Omega)$ such that

$$P^{p-1} \subset Q_p \quad \text{and} \quad P^{p-1} \subset \nabla \times Q_p.$$  

(1.84)

for some $p > 0$. Let $r > 0$ and let $p \geq r$. There exists a constant $C > 0$, dependent upon $r$, such that

$$\inf_{\varphi \in Q_p} \left( \|E - \varphi\|_{H^r(\Omega)}^2 + \|\nabla \times (E - \varphi)\|_{H^r(\Omega)}^2 \right)^{1/2} \leq C \left( |E|_{H^r(\Omega)}^2 + |\nabla \times E|_{H^r(\Omega)}^2 \right)^{1/2}$$

(1.85)
for every $E \in H^r(\Omega)$ such that $\nabla \times E \in H^r(\Omega)$.

**PROOF** It is sufficient to prove the result for $p = r$. Consider the space

$$H^r(\text{curl}, \Omega) := \{E \in H^r(\Omega) : \nabla \times E \in H^r(\Omega)\}$$

and the corresponding quotient space:

$$H^r(\text{curl}, \Omega)/Q^p.$$ (1.87)

We have:

$$\inf_{\varphi \in Q^p} \left( |E - \varphi|^2_{H^r(\Omega)} + |\nabla \times (E - \varphi)|^2_{H^r(\Omega)} \right)^{1/2} \leq \inf_{\varphi \in Q^p} \left( \|E - \varphi\|^2_{H^r(\Omega)} + \|\nabla \times (E - \varphi)\|^2_{H^r(\Omega)} \right)^{1/2}$$

and both sides represent a norm for the quotient space. Indeed, the right-hand side is the standard norm for a quotient space. Concerning the left-hand side, we need only to prove the definitness, i.e. if the left-hand side vanishes for a function $E$, then $E$ must be in $Q^p$. Since the polynomial space is finite dimensional, the infimum on the left-hand side is attained for some specific $\varphi \in Q^p$.

Both terms are non-negative so they both must vanish. Vanishing of the second term implies\(^1\) that $\nabla \times (E - \varphi) \in \mathcal{P}^{r-1}$. Vanishing of the first term implies that $E - \varphi \in \mathcal{P}^{r-1}$. Consequently, $E - \varphi \in Q^p$ and, therefore, $E \in Q^p$ as well.

Now comes a delicate point. We claim that the quotient space equipped with both norms is complete. For the norm on the right-hand side, this is a standard result for Banach spaces. For the norm on the left-hand side, we need to show it. Let $E_n \in H^r(\text{curl}, \Omega)/Q^p$ be a Cauchy sequence. Then $E_n$ is a Cauchy sequence in $H^r(\Omega)/Q^p$ and also $\nabla \times E_n$ is a Cauchy sequence in $H^r(\Omega)/\nabla \times Q^p$. By Corollary 1.9.1, both spaces equipped with the alternate norm implied by the seminorms, are complete and, therefore, both $E_n$ and $\nabla \times E_n$ converge to some limits, say $E, F$. The touchy point is to show that $F = \nabla \times E$ modulo a polynomial in $Q^p$. Consider any multiindex $\alpha, |\alpha| = r$. We have,

$$\langle D^\alpha E_n, \nabla \times \psi \rangle = \langle D^\alpha \nabla \times E_n, \psi \rangle \quad \forall \psi \in \mathcal{D}(\Omega).$$

For a given $\psi \in \mathcal{D}(\Omega)$, both sides are continuous functional on our quotient space. Passing to the limit, we obtain,

$$\langle D^\alpha E, \nabla \times \psi \rangle = \langle D^\alpha F, \psi \rangle \quad \forall \psi \in \mathcal{D}(\Omega).$$

Consequently,

$$D^\alpha (\nabla \times E - F) = 0 \quad \text{for every } |\alpha| = r$$

which shows that $\nabla \times E - F \in \mathcal{P}^{r-1} \subset \nabla \times Q^p$. Done.

\(^1\) $|u|_{H^r(\Omega)} = 0 \Rightarrow u \in \mathcal{P}^{r-1}(\Omega)$. 
Consequently, the identity map is continuous when the quotient space is equipped with those two norms. By the Banach Theorem, the inverse map (the identity itself) must be continuous as well. Thus the reverse inequality holds with some multiplicative constant $C$. Finally, we have trivially (set $\varphi = 0$),

$$\inf_{\varphi \in Q_{p,q}} \left( \|E - \varphi\|_{H^r(\Omega)}^2 + \|\nabla \times (E - \varphi)\|_{H^s(\Omega)}^2 \right)^{1/2} \leq C \left( \|E\|_{H^r(\Omega)}^2 + \|\nabla \times E\|_{H^s(\Omega)}^2 \right)^{1/2}. \quad (1.89)$$

\[\square\]

In the same way we prove an analogous result for the $H(\text{div})$ spaces.

**THEOREM 1.9.3**

(Bramble-Hilbert Argument for $H(\text{div})$ norm)

Let $\Omega$ be a domain in $\mathbb{R}^3$ and let $V^p$ be a subspace of $H(\text{div}, \Omega)$ such that

$$\mathbb{P}^{p-1} \subset V^p \quad \text{and} \quad \mathbb{P}^{p-1} \subset \nabla \cdot V^p, \quad (1.90)$$

for some $p > 0$. Let $r > 0$ and let $p \geq r$. There exists a constant $C > 0$, dependent upon $r$, such that

$$\inf_{\phi \in V^p} \left( \|v - \phi\|_{H^r(\Omega)}^2 + \|\nabla \cdot (v - \phi)\|_{H^r(\Omega)}^2 \right)^{1/2} \leq C \left( \|v\|_{H^r(\Omega)}^2 + \|\nabla \cdot v\|_{H^r(\Omega)}^2 \right)^{1/2} \quad (1.91)$$

for every $v \in H^r(\text{div}, \Omega)$ where

$$H^r(\text{div}, \Omega) := \{v \in H^r(\Omega) : \nabla \cdot v \in H^r(\Omega)\}. \quad (1.92)$$

1.9.2 $H^1$, $H(\text{curl})$ and $H(\text{div})$ $h$ - Interpolation Estimates

We will discuss the 3D case. The estimate for the 2D case will be left as an exercise.

Let

$$x = h\xi + b \quad (1.93)$$

be the simplest element map with $h = h_K$ being the element size.

The Piola transforms imply the following scalings for the $H^1$-, $H(\text{curl})$-, $H(\text{div})$-, and $L^2$-conforming elements:

$$u = \hat{u}, \quad E = h^{-1} \hat{E}, \quad v = h^{-2} \hat{v}, \quad f = h^{-3} \hat{f}. \quad (1.94)$$
\textbf{L}^2\text{-projection estimate.} Let \( f \in H^r(K) \), and let \( f_p = Pf \) be the \( L^2 \)-projection on space \( Y_p \) such that \( \mathcal{P}_{p-1} \subset Y_p \), \( p \geq r \). We have:

\[
\| f - f_p \|^2 = h^{-6}h^3\| \hat{f} - \hat{f}_p \|^2 \quad \text{(Piola transform and change of coordinates)}
\]

\[
= h^{-3}\inf_{\hat{\phi} \in \hat{Y}_p} \| (I - \hat{P})(\hat{f} - \hat{\phi}) \|^2 \quad \text{(shape functions preserving property)}
\]

\[
\leq h^{-3}\| I - \hat{P} \|^2 \inf_{\hat{\phi} \in \hat{Y}_p} \| \hat{f} - \hat{\phi} \|^2_{L^2(\hat{K})} \quad \text{(continuity of } L^2 \text{-projection, } \| I - \hat{P} \| = 1)\]

\[
\lesssim h^{-3}\| \hat{f} \|^2_{H^r(\hat{K})} \quad \text{(Bramble-Hilbert argument)}
\]

\[
= h^{-3}h^3h^{-3}h^{2r}\| f \|^2_{H^r(K)} = h^{2r}\| f \|^2_{H^r(K)} \quad \text{(scalings)}
\]

(1.95)

\textbf{H (div)-interpolation estimate.} Let \( v \in H^r(\text{div, } K) \) be a given function, and let \( v_p = \Pi^\text{div}v \in V^p \) denote its FE interpolant. We assume that \( \mathcal{P}_{p-1} \subset V^p \), \( \mathcal{P}_{p-1} \subset \nabla \cdot V^p \), \( p \geq r \).

\[
\| v - v_p \|^2 = h^{-4}h^3\| \hat{v} - \hat{v}_p \|^2 \quad \text{(scalings and change of variables)}
\]

\[
= h^{-1}\| (I - \Pi^\text{div})\hat{v} \|^2
\]

\[
= h^{-1}\inf_{\hat{\phi} \in \hat{V}_p} \| (I - \Pi^\text{div})(\hat{v} - \hat{\phi}) \|^2 \quad \text{(shape functions preserving property)}
\]

\[
\lesssim h^{-1}\| I - \Pi^\text{div} \|^2_{L^2(H^{r-r}(\text{div, } K), L^2(\hat{K}))}
\]

\[
\inf_{\hat{\phi} \in \hat{V}_p} \left( \| \hat{v} - \hat{\phi} \|^2_{H^r(\hat{K})} + \| \nabla \cdot (\hat{v} - \hat{\phi}) \|^2_{H^r(\hat{K})} \right) \quad \text{(continuity of interpolation operator)}
\]

\[
\lesssim h^{-1}\| \hat{v} \|^2_{H^r(\hat{K})} + \| \nabla \cdot \hat{v} \|^2_{H^r(\hat{K})} \quad \text{(Bramble-Hilbert argument)}
\]

\[
= h^{-1}(h^{2r+1}|v|^2_{H^r(K)} + h^{2r+3}|
abla \cdot v|^2_{H^r(K)}) \quad \text{(scalings)}
\]

\[
\leq h^{2r}|v|^2_{H^r(\text{div, } K)} \quad \text{(definition of the seminorm)}
\]

(1.96)

Notice that the higher power of \( h \) that we get in the second term is useless as the first term dominates.

The commuting diagram property implies now the estimate in the full \( H(\text{div}) \)-norm. Indeed, for \( f = \nabla \cdot v \), \( \nabla \cdot \Pi^\text{div}v = Pf = f_p \), which implies that

\[
\| \nabla \cdot (v - v_p) \|^2 = \| \nabla \cdot v - f_p \|^2 \leq C h^{2r}|\nabla \cdot v|^2_{H^r(K)} \leq C h^{2r}\| v \|^2_{H^r(\text{div, } K)}
\]

(1.97)

Combining the two estimates above, we obtain,

\[
\| v - v_p \|^2_{H(\text{div, } K)} \leq C h^{2r}\| v \|^2_{H^r(\text{div, } K)}
\]

(1.98)

\textbf{H (curl)-interpolation estimate.} Let \( E \in H^r(\text{curl, } K) \), and let \( E_p = \Pi^\text{curl}E \in Q^p \) denote its FE interpolant. We assume that \( \mathcal{P}_{p-1} \subset Q^p \), \( \mathcal{P}_{p-1} \subset \nabla \times Q^p \), \( p \geq r \) and proceed analogously to the \( H(\text{div}) \)
\[ \| E - E_p \|^2 = h^{-2} h^2 \| \hat{E} - \hat{E}_p \|^2 \] ( scalings and change of variables )

\[ = h \inf_{\hat{\varphi} \in \mathcal{Q}_p} \| (I - \Pi^\text{grad}) (\hat{E} - \hat{\varphi}) \|^2 \] ( FE shape functions preserving property )

\[ \lesssim h \| I - \Pi^\text{curl} \|^2_{L(\mathcal{H}^r(\text{curl}, K), L^2(\bar{K}))} \]

\[ \inf_{\hat{\varphi} \in \mathcal{Q}_p} \left( \| \hat{E} - \hat{\varphi} \|^2_{\mathcal{H}^r(K)} + \| \nabla \times (\hat{E} - \hat{\varphi}) \|^2_{\mathcal{H}^r(K)} \right) \] ( continuity of interpolation operator )

\[ \lesssim h \left( \| E \|^2_{\mathcal{H}^r(K)} + |\nabla \times E|^2_{\mathcal{H}^r(K)} \right) \] ( Bramble-Hilbert argument )

\[ = h \left( h^{2r-1} |E|^2_{\mathcal{H}^r(K)} + h^2 |\nabla \times E|^2_{\mathcal{H}^r(K)} \right) \] ( scalings )

\[ = h^{2r} |E|^2_{\mathcal{H}^r(\text{curl}, K)} \] ( definition of the seminorm )

(1.99)

\[ \| \nabla \times (E - E_p) \|^2 = \| \nabla \times E - v_p \|^2 \leq C h^{2r} |\nabla \times E|^2_{\mathcal{H}^r(K)} \leq C h^{2r} \| E \|^2_{\mathcal{H}^r(\text{curl}, K)} \] (1.100)

Combining the two estimates above, we obtain,

\[ \| E - E_p \|^2_{\mathcal{H}(\text{curl}, K)} \leq C h^{2r} \| E \|^2_{\mathcal{H}^r(\text{curl}, K)} \] (1.101)

\[ H^1\text{-interpolation estimate.} \] Let \( u \in \mathcal{H}^r(K) \), and let \( u_p = \Pi^\text{grad} u \in W^p \) denote its FE interpolant. We assume that \( \mathcal{P}^p \subset W^p, p + 1 \geq r \). We have,

\[ \| u - u_p \|^2 = h^3 \| \hat{u} - \hat{u}_p \|^2 \] ( scalings and change of variables )

\[ = h^3 \inf_{\hat{\varphi} \in W^p} \| (I - \Pi^\text{grad}) (\hat{u} - \hat{\varphi}) \|^2 \] ( FE shape functions preserving property )

\[ \lesssim h^3 \| I - \Pi^\text{grad} \|^2_{L(\mathcal{H}^r(\bar{K}), L^2(\bar{K}))} \]

\[ \inf_{\hat{\varphi} \in W^p} \| \hat{u} - \hat{\varphi} \|^2_{\mathcal{H}^r(K)} \] ( continuity of interpolation operator )

\[ \lesssim h^3 \| \hat{u} \|^2_{\mathcal{H}^r(K)} \] ( Bramble-Hilbert argument )

\[ = h^{2r} \| u \|^2_{\mathcal{H}^r(K)} \] ( scalings )

Assume now that \( p \geq r \). Consequently, \( p + 1 \geq r + 1 \), and we can replace \( r \) with \( r + 1 \) to get,

\[ \| w - \Pi^\text{grad} w \|_{L^2(K)} \lesssim h^{r+1} \| w \|_{\mathcal{H}^{r+1}(K)} \] (1.103)

Moreover, applying the \( H(\text{curl}) \) estimate to gradient \( \nabla w \), and using the commutativity argument, we get,

\[ \| \nabla w - \Pi^\text{curl} \nabla w \| = \| \nabla (w - \Pi^\text{grad} w) \| \lesssim h^r \| \nabla w \|_{\mathcal{H}^r(K)} \leq h^r \| w \|_{\mathcal{H}^{r+1}(K)} \] (1.104)

which yields the final estimate in the full norm,

\[ \| w - \Pi^\text{grad} w \|_{\mathcal{H}^1(K)} \lesssim h^r \| w \|_{\mathcal{H}^{r+1}(K)} \] (1.105)
Note that the $L^2$ interpolation error converges one order faster than the $H^1$ error. This is not the case for the $H(\text{curl})$ and $H(\text{div})$ estimates where the $L^2$-estimates and the corresponding energy estimates are of the same order.

**Limited regularity case.** We explain the issue for the $H(\text{curl})$ case only. The other cases are fully analogous. Two situations are possible:

- The interpolated function is (relative to $p$) regular, i.e. $p < r$. We use the estimate above with $p$ in place of $r$ to obtain:

  \[
  \| E - E_p \|_{H(\text{curl}, K)} \leq C h^p \| E \|_{H^r(\text{curl}, K)} \leq C h^p \| E \|_{H^r(\text{curl}, K)} \quad \text{(1.106)}
  \]

  The rate of convergence is dictated by the polynomial order $p$.

- The interpolated function is less regular, $p > r$. We use the original estimate to obtain

  \[
  \| E - E_p \|_{H(\text{curl}, K)} \leq C h^r \| E \|_{H^r(\text{curl}, K)} \quad \text{(1.107)}
  \]

  In this case, the rate of convergence is dictated by the regularity of the solution.

We usually combine the two estimates into one by writing:

\[
\| E - E_p \|_{H(\text{curl}, K)} \leq C h^{\min\{p, r\}} \| E \|_{H^r(\text{curl}, K)} . \quad \text{(1.108)}
\]

**REMARK 1.9.1** All the estimates above have been carried out for an integer $r$. An interpolation argument for Hilbert spaces can be used to generalize the results to real values of $r$.

**Minimum regularity of interpolated functions.** What is the minimum value of $r$ for which the standard interpolation operators are continuous on Sobolev spaces? By the Uniform Boundedness Theorem (Banach-Steinhaus Theorem), it is sufficient to determine the minimum $r$ for which the interpolation operator is well-defined. In $H^1$-interpolation, we use point values (e.g. at vertices or Lagrange nodes). The answer comes then from the Sobolev Embedding Theorem: $r > 1/2$ in 1D, $r > 1$ in 2D and $r > 3/2$ in 3D. As $\nabla H^r \subset H^{r-1}(\text{curl})$, the commuting diagram property implies that we must have $r > 0$ in 2D and $r > 1/2$ in 3D for computing the edge averages of $E_t$. This is indeed the case, the estimate comes this time from Trace Theorems. In 2D, for $r > 0$, trace of functions from $H^r(\text{curl}, K)$ to an edge $e$ lives in $H^{r-1/2}(e)$ and this is a sufficient regularity to compute the edge average. Indeed, the edge average of tangential component $E_t$ can be viewed as action of $E_t$ on the unity function,

\[
\int_e E_t = \langle E_t, 1 \rangle ,
\]

and we need only to argue that the unity function lives in the dual of $H^{r-1/2}(e)$ for $r > 0$. This is indeed the case although the proof of this innocent statement requires a working knowledge of Sobolev spaces. In
3D we need to apply the Trace Theorem twice. For \( r > 1/2 \), trace \( E_t \) of a function \( E \) from \( H^r(\text{curl}, K) \) to a face \( f \), lives in \( H^{r-1/2}(\text{curl}, f) \). Applying the 2D result to the space on the face finishes then the reasoning. Finally, the Trace Theorem for \( H(\text{div}) \) spaces implies that the face averages are well defined for functions \( v \in H^r(\text{div}, K) \), for \( r > 0 \). Indeed, the normal trace \( v \cdot n \) to a face \( f \) lives then in \( H^{r-1/2}(f) \), and this is again sufficient to interpret the face average of \( v_n := v \cdot n \) as the action of \( v_n \) on the unity function.

The presented Projection-Based (PB) Interpolation increases the regularity assumptions. For instance, the edge projections in 3D, require the edge trace of a function \( u \) to be in \( H^r(e) \). The Trace Theorem implies then that function \( u \) must come from \( H^r(K) \) with \( r > 2 \). This may be too demanding from the point of view of expected regularity of functions to be interpolated (exact solutions) and has led to a modification of the PB interpolation using fractional norms, compare [19] with [17, 15]. The version of the PB Interpolation using projections in fractional norms requires the same regularity assumptions as the classical interpolation operators for the lowest order elements.

Estimates for general affine elements. Shape regularity assumptions. The presented interpolation error estimates generalize easily to the case of a general affine isomorphism,

\[
x_K : \hat{K} \ni \xi \rightarrow x = A\xi + x_0.
\]

Here \( A \) is a non-singular matrix, \( \det A \neq 0 \), and \( x_0 \) is a point. Obviously, both may depend upon the element \( K \). Note that the inverse of an affine isomorphism is an affine isomorphism as well. Typically, we request \( \det A > 0 \).

In place of simple scalings, we need now more careful estimates for the Piola maps. We have,

\[
\begin{align*}
j &= \det A, \quad j^{-1} = \det A^{-1}, \\
\| E \| &\leq \| J^{-T} \| \| \hat{E} \| = \| A^{-1} \| \| \hat{E} \|, \\
\| H \| &\leq \| \det A^{-1} \| \| A \| \| \hat{H} \|, \\
| f | &\leq \| \det A^{-1} \| | \hat{f} |,
\end{align*}
\]

where all norms of vectors and matrices are Euclidean norms. We may estimate them in terms of geometrical quantities. For instance,

\[
\| A \| \leq \frac{h}{\hat{\rho}} \tag{1.109}
\]

where \( h = h_K \) is the elemet size defined as

\[
h_K := \sup_{x,y \in K} \| x - y \|
\]

and \( \hat{\rho} \) is the radius of the largest sphere contained in the corresponding master element \( \hat{K} \), comp. Exercise 1.9.2.

In the same way we can estimate higher order Sobolev seminorms. Start with the transformation rule for first order differential,

\[
d_{\xi} \hat{u}(\hat{\xi}) = d_x u(A\hat{\xi}).
\]
Analogously, for a differential of order \( r \),

\[
d_{\xi}^{r} \hat{u}(\hat{e}_1, \ldots, \hat{e}_r) = d_{\xi}^{r} u(A \hat{e}_1, \ldots, A \hat{e}_r). \tag{1.110}
\]

Consequently,

\[
\left\| d_{\xi}^{r} \hat{u} \right\| := \sup_{\hat{e}_1, \ldots, \hat{e}_r} \frac{d_{\xi}^{r} \hat{u}(\hat{e}_1, \ldots, \hat{e}_r)}{\left\| \hat{e}_1 \right\|, \ldots, \left\| \hat{e}_r \right\|} = \sup_{\hat{e}_1, \ldots, \hat{e}_r} \frac{d_{\xi}^{r} u(A \hat{e}_1, \ldots, A \hat{e}_r)}{\left\| A \hat{e}_1 \right\|, \ldots, \left\| A \hat{e}_r \right\|} \leq \left\| d_{\xi}^{r} u \right\| \| A \|^r.
\]

Consequently, if we can bound \( \| A \|, \| A^{-1} \|, |j^{-1}| \) uniformly for all elements in the mesh, all the discussed interpolation error estimates hold as well at the expense of introducing additional constants reflecting shape regularity, comp. Exercise 1.9.3.

Note that formula (1.110) does not hold for a non-constant Jacobian. In the case of a general element map, \( r \)-th derivative in \( \xi \) will depend upon not only \( r \)-th derivative in \( x \) but also all derivatives of lower order \( 1, \ldots, r - 1 \). Consequently, the scaling argument fails and the estimates do not generalize to non-affine elements.

**Discretization in the parametric domain.** This is very important. The element maps need not be random (as it happens for instance in the case of unstructured mesh generators). In the case of CAD defined geometries, they come from a predefined global geometry map in a reference domain, see Fig. 1.3, where element map \( x_K \) is the composition of an affine reference map \( \eta = \eta(\xi) \) and the CAD parametrization \( x = x(\eta) \). The entire problem can be redefined in the reference domain. The geometry maps contribute then to the redefined material data, and the original problem is effectively solved in the reference domain where all elements are shape regular affine elements. The CAD parametrizations can be used directly (exact geometry elements) or they can be approximated (interpolated) with polynomials, usually coming from \( H^1 \) space of element shape functions \( W^p \) (isoparametric element). There is no problem with convergence then.

**Figure 1.3**
Reference geometry map.
1.9.3 *hp*- Interpolation Estimates.

If the interpolation operator preserves FE shape functions and we have in our disposal a $p$-interpolation estimates on the master element, we can immediately use the discussed scaling arguments to obtain the corresponding $hp$-interpolation error estimates. For instance, for the $H(\text{curl})$ case, we have,

$$\| \tilde{E} - \hat{\Pi}_\text{curl} \tilde{E} \|_{H(\text{curl}, \hat{K})} \leq C(r) \ln p \, p^{-r} \| \hat{E} \|_{H^r(\text{curl}, \hat{K})}. \quad (1.111)$$

Instead of using the continuity of the interpolation operator, we can use now the $p$-estimate:

$$h^1 \| \tilde{E} - \hat{\Pi}_\text{curl} \tilde{E} \|_{H(\text{curl}, \hat{K})}^2 = h^1 \inf_{\hat{\psi} \in \hat{Q}_p} \| (\hat{E} - \hat{\psi}) - \hat{\Pi}_\text{curl} (\hat{E} - \hat{\psi}) \|_{H(\text{curl}, \hat{K})}^2 \quad (\text{shape functions preservation}) \quad (1.112)$$

$$\lesssim h^1 \ln^2 p \, p^{-2r} \inf_{\hat{\psi} \in \hat{Q}_p} \| \hat{E} - \hat{\psi} \|_{H^r(\text{curl}, \hat{K})} \quad (p\text{-interpolation error estimate})$$

with the rest of the argument remaining identical as in the $h$-case. The ultimate estimate reads as follows:

$$\| E - \Pi \text{curl} E \|_{H(\text{curl}, K)} \leq C(r) \ln p \, \frac{h^\min\{p, r\}}{p^r} \| E \|_{H^r(\text{curl}, K)} \quad (1.113)$$

**Exercises**

**Exercise 1.9.1** Norm of a matrix induced by Euclidean norm. Let $A \in L(\mathbb{R}^n, \mathbb{R}^n)$ be a linear map. Let $\| \cdot \|$ be the standard $l^2$ (Euclidean) norm in $\mathbb{R}^n$. The corresponding induced norm for map $A$ is defined as:

$$\| A \| := \sup_{x \neq 0} \frac{\| Ax \|}{\| x \|}.$$

(i) Demonstrate that the norm of $A$ equals the maximum characteristic value of $A$:

$$\| A \| = \max_{i=1, \ldots, n} \lambda_i,$$

where $\lambda_i \geq 0$, $\lambda_i^2$ are eigenvalues of $AA^T$ or, equivalently, $A^T A$.

(ii) Extend the formula to the complex case.

(5 points)

**Exercise 1.9.2** Estimate of the Euclidean norm of a linear map (Jacobian of an affine map). Prove estimate (1.109).

(3 points)

**Exercise 1.9.3** Interpolation error estimates for an affine element. Rederive all four interpolation error estimates for a general affine element. Use geometrical estimate (1.109) for Jacobians.

(10 points)
Exercise 1.9.4 Fractional Sobolev spaces. Consider the infinite L-shape domain shown in Fig. 1.4.

(i) Switch to polar coordinates and use separation of variables to determine a family of solutions to the Laplace equation with homogenous BC \( u = 0 \) on the reentrant edges.

(ii) Determine the most singular solution that belongs to the energy space \( H^1_{\text{loc}} \) (it is in \( H^1 \) in any compact neighborhood of the reentrant corner). It should be in the form of

\[ u(r, \theta) = r^\alpha f(\theta). \]

where \( f(\theta) \) is a smooth function.

(iii) Determine values of exponent \( \alpha \) for which the function above lives in \( H^1_{\text{loc}} \) or \( H^2_{\text{loc}} \). Guess the fractional Sobolev space in which the actual solution lives.

This “guessing” procedure may be made very precise using the interpolation theory for Sobolev spaces. Solution to a corresponding Laplace problem in a bounded domain containing the reentrant corner (with same BC along the reentrant edges but arbitrary BC on the remaining part of the boundary) will have the same singularity at the corner. The solution you have developed is commonly used as a manufactured solution for a bounded domain to verify the expected convergence rates.

Figure 1.4
Infinite L-shape domain.

(5 points)
1.10 Numerical Experiments

Exercises

Exercise 1.10.1 Effect of the regularity of the solution on the convergence rates. Repeat the experiment from the class: use the code to study the convergence rates corresponding to uniform $h$-refinements for the L-shape domain problem and order of approximation $p = 1, 2, 3, 4, 5$. Plot $H^1$ error vs. element size $h$ on the log-log scale determining the experimental convergence rates. Compare the results with those predicted by the theory. (8 points)

Exercise 1.10.2 Comparison of different refinement strategies. Repeat the experiment from the class for the L-shape domain problem:

(i) Prepare a new input with quads only.

(ii) Starting with a mesh of quadratic elements, perform convergence study for uniform $h$- and $p$- refinements, and then for $h$-adaptive and $hp$-adaptive refinements provided by the code. Compute the experimental convergence rates.

(iii) Plot $H^1$ error vs. number of dof on the log-log scale. Compute experimental convergence rates and illustrate them on the plot. You may modify the existing plotting routine within the code or (easier) use a third party software for plotting.

Attempt to discuss the results.

(10 points)

Exercise 1.10.3 Repeat the exercise 1.10.2 with a mesh consisting of triangles only. Which elements give better results?

(10 points)

Exercise 1.10.4 L-shape domain problem for elasticity. This is a coding exercise. Code the Osborn exact solution for the L-shape domain. Consult [5] or the literature cite therein for the formulas. With the manufactured solution in place, repeat Exercise 1.10.1.

(20 points)
1.11 Aubin–Nitsche Argument

Cea’s lemma argument establishes convergence in the energy norm. For the $H^1$ energy norm setting, this does not imply an optimal convergence rate in the weaker $L^2$-norm.

Let $u \in U \subset L^2(\Omega)$ be the exact solution and $u_h \in U_h \subset U$ its Bubnov–Galerkin approximation. Consider the dual problem:

\[
\begin{aligned}
\forall w \in U \quad \langle v_g, w \rangle &= g(w) := (w, u - u_h) \\
\end{aligned}
\]

where $(\cdot, \cdot)$ is the $L^2$-product. Assume that the dual problem is well–posed and admits a stability estimate in a Sobolev norm stronger than the energy norm $H^1$.

\[
\|v_g\|_{H^{1+s}(\Omega)} \leq C\|u - u_h\|, \quad s > 0.
\]

Discussion the typical scenario for standard elliptic problems.

Then

\[
\begin{aligned}
\|u - u_h\|^2 &= (u - u_h, u - u_h) \\
&= b(u - u_h, v_g) \quad \text{(definition of the dual problems)} \\
&= b(u - u_h, v_g - v_h) \quad \text{(Galerkin orthogonality)} \\
&\leq M\|u - u_h\|_{U}\|v_g - v_h\|_{U} \\
&\leq CM\|u - u_h\|_{U} h^s\|v_g\|_{H^{1+s}} \quad \text{(best approximation error estimate for $v_g$)} \\
&\leq CMh^s\|u - u_h\|_{U} \|u - u_h\|
\end{aligned}
\]

Dividing both sides by $\|u - u_h\|$, we obtain,

\[
\|u - u_h\| \leq CMh^s\|u - u_h\|_{U}.
\]

Thus, if the solution $u_h$ converges to $u$ with a specific rate $h^r$ in the energy norm, it will converge also to $u$ in the $L^2$-norm with a higher rate $h^{r+s}$. The gain $s$ depends upon the stability properties of the continuous dual problem. For standard second-order elliptic problems and smooth or convex domains, $s = 1$, i.e. the actual $L^2$ error converges with the same rate as the best approximation error.

Exercises

Exercise 1.11.1 $L^2$ convergence. Add the computation of the $L^2$-norm of the FE error to the 2D code (a couple of extra lines of code have to be added to the existing routines computing the $H^1$ error). Set
up two model Poisson problems: a smooth manufactured solution problem in a unit square, and the L-shape domain problem, and study the convergence in the $L^2$-norm comparing it with the convergence in the $H^1$ energy norm. Discuss the results.

(10 points)

---

**Historical Comments**
2

Beyond Coercivity

2.1 Babuška’s Theorem

THEOREM 2.1.1 Babuška - Nečas Theorem

Consider the standard abstract variational problem,

\[
\begin{aligned}
& u \in U \\
& b(u, v) = l(v) \quad \forall v \in V
\end{aligned}
\]  

\tag{2.1}

where \( U, V \) are Hilbert (trial and test) spaces, \( b(u, v) \) is a continuous bilinear (sesquilinear) form, and \( l \in V' \) is a continuous linear (antilinear) form on test space \( V \). Additionally, assume that \( b \) satisfies the inf-sup condition:

\[
\inf_{u \in U, u \neq 0} \sup_{v \in V, v \neq 0} \frac{|b(u, v)|}{\|u\|_U \|v\|_V} \geq \gamma > 0 ,
\]

and \( l \) satisfies the compatibility condition:

\[ l(v) = 0 \quad \forall v \in V_0 := \{ v \in V : b(u, v) = 0 \forall u \in U \} . \]

There exists then a unique solution \( u \) to the variational problem and it satisfies the stability estimate:

\[ \|u\|_U \leq \frac{1}{\gamma} \|l\|_V. \]

PROOF The result is a reinterpretation of Banach Closed Range Theorem, see [28] p. 518, for details.

LEMMA 2.1.1 Del Pasqua, Ljance, Kato [29]

Let \( U, (\cdot, \cdot) \) be a Hilbert space and \( P : U \to U \) a linear projection, i.e. \( P^2 = P \). Then

\[ \|I - P\| = \|P\| \]

PROOF Let \( X = \mathcal{R}(P) \) and \( Y = \mathcal{N}(P) \). It is well known that \( U = X \oplus Y \). Pick an arbitrary unit vector \( u \in U \). Let \( u = x + y \), \( x \in X, y \in Y \) be the unique decomposition of \( u \). By the properties
of a scalar product,

\[ 1 = \|u\|^2 = (x + y, x + y) = \|x\|^2 + \|y\|^2 + 2\text{Re}(x, y). \]

Consider now a “symmetric image” \( w \) of \( u \), see Fig. 2.1,

\[ w = \bar{x} + \bar{y}, \quad \bar{x} = \|y\| \frac{x}{\|x\|}, \quad \bar{y} = \|x\| \frac{y}{\|y\|}. \]

Vector \( w \) has a unit length as well. Indeed,

\[ \|w\|^2 = (\bar{x} + \bar{y}, \bar{x} + \bar{y}) = \|x\|^2 + \|y\|^2 + 2\text{Re}(\bar{x}, \bar{y}) = \|y\|^2 + \|x\|^2 + 2\text{Re} \left( \frac{\|y\| \|x\|}{\|x\| \|y\|} (x, y) \right) = 1. \]

We have now,

\[ \|Pu\| = \|x\| = \|\bar{y}\| = \|(I - P)w\| \leq \|I - P\| \|w\| = \|I - P\| \]

Taking supremum over \( \|u\| = 1 \) finishes the proof. \( \blacksquare \)

**Figure 2.1**
Illustration of the proof of Pasqua-Ljance-Kato Lemma

---

**THEOREM 2.1.2 Babuška Theorem [2]**

Consider the Petrov–Galerkin discretization of variational problem (2.1),

\[
\begin{align*}
    u_h & \in U_h \\
    b(u_h, v_h) & = l(v_h) \quad \forall v_h \in V_h
\end{align*}
\]

(2.2)

where \( U_h \subset U, V_h \subset V \) are discrete trial and test spaces, and \( \dim U_h = \dim V_h < \infty \). Assume that form \( b \) and the discrete spaces \( b \) satisfy the discrete inf-sup condition:

\[
\inf_{u_h \in U_h, u_h \neq 0} \sup_{v_h \in V, v_h \neq 0} \frac{|b(u_h, v_h)|}{\|u_h\|_U \|v_h\|_V} =: \gamma_h > 0.
\]
Beyond Coercivity

There exists then a unique (discrete) solution \( u_h \) to variational problem (2.2) which satisfies the stability estimate:

\[
\|u_h\|_U \leq \frac{1}{\gamma_h} \|l\|_{V_h'}.
\]

Additionally, we have:

\[
\|u - u_h\| \leq \frac{M}{\gamma_h} \inf_{w_h \in U_h} \|u - w_h\|_U.
\]

where \( M \) is the continuity constant for form \( b \).

PROOF The stability result is a direct consequence of Babuška-Nečas Theorem as the discrete variational problem is simply a particular case of the general case. Notice that no compatibility condition is needed on the discrete level, Galerkin stiffness matrix and its transpose have the same rank. The proof of the error estimate (2.3) begins with an observation that the Petrov–Galerkin discretization executes a linear projection \( P_h : U \rightarrow U_h, P_h u = u_h, \)

\[
b(P_h u - u, v_h) = 0 \ \forall v_h \in V_h.
\]

The stability estimate implies an estimate on the norm of the projection,

\[
\|P_h u\|_U = \|u_h\| \leq \frac{1}{\gamma_h} \|l\|_{V_h'} = \frac{1}{\gamma_h} \sup_{\|v_h\|=1} |b(u, v_h)| \leq \frac{M}{\gamma_h} \|u\|_U.
\]

We have then:

\[
\|u - u_h\|_U = \|(I - P_h)u\|_U \quad \text{ (definition of } P_h) \]
\[
= \|(I - P_h)(u - w_h)\|_U \quad \text{ (} P_h w_h = w_h, \ \forall w_h \in U_h \text{)}
\]
\[
\leq \|I - P_h\| \|u - w_h\| \quad \text{(Lemma 2.1.1)}
\]
\[
= \|P_h\| \|u - w_h\| \quad \text{(} \|P_h\| \leq \frac{M}{\gamma_h} \text{)}.
\]

and we conclude the proof by taking infimum over \( w_h \in U_h \).

If \( \gamma_h \) admit a positive lower bound, i.e. a uniform discrete inf–sup condition holds,

\[
\inf_{\gamma_h} \gamma_h =: \gamma_0 > 0,
\]

then,

\[
\|u - u_h\| \leq \frac{M}{\gamma_0} \inf_{w_h \in U_h} \|u - w_h\|_U.
\]

(2.3)

i.e. the actual and the best approximation errors must converge at the same rate. The result has coined the famous phrase:

(Uniform) discrete stability and approximability imply convergence.
It is not an exaggeration to say that the entire numerical analysis for linear problems hinges on the Babuška Theorem. The result, unfortunately, is not constructive. It tells us what we should have to ensure the stability and convergence for the Galerkin method, but it gives no hint how to select the spaces to guarantee the discrete inf–sup condition. However, the result underlines the different criteria for selection of spaces, the trial space choice controls the approximability error, whereas the test space controls the stability. In the case of $U = V$, we may choose $V_h = U_h$ (Bubnov–Galerkin method), but the control of stability becomes then incidental, we may not have it.

**Exercises**

**Exercise 2.1.1** Discrete inf–sup constant. Let $e_i \in U_h$ and $g_j \in V_h$, $i, j = 1, \ldots, \dim U_h = \dim V_h$ be specific basis functions for the discrete trial and test space. Introduce the Galerkin stiffness matrix and the corresponding Gram matrices for the norms,

$$B_{ji} := b(e_i, g_j), \quad U_{ji} := (e_i, e_j)_U, \quad V_{ji} := (g_i, g_j)_V.$$

Derive an explicit formula for the discrete inf–sup constant $\gamma_h$ in terms of matrices $B, U, V$. Can you compute it using standard (iterative) algorithms for computing eigenvalues? (10 points)

**2.2 Asymptotic Stability**

Consider a class of variational problems of the form,

$$\begin{cases} u \in V \\ a(u, v) + c(u, v) = l(v), & v \in V \end{cases}$$

(2.4)

where $V$ is a Hilbert space, sesquilinear form $a(u, v)$ is Hermitian and coercive,

$$a(u, v) = \overline{a(v, u)}, \quad u, v \in V; \quad a(u, u) \geq \alpha \|u\|_V^2, \quad u \in V, \quad \alpha > 0,$$

$l \in V'$, and form $c(u, v)$ is compact. From many possible definitions for a compact form, we choose the one as follows. Form $c(u, v)$ is said to be compact iff

$$\sup_{\|v\| \leq 1} |c(u - u_0, v)| \to 0.$$

(2.5)

**REMARK 2.2.1** Let $c(u, v)$ be compact. Then

$$c(u_n - u_0, u_n - u) \to 0 \quad \text{if} \quad u_n \to u.$$
Indeed, you need to recall only that weak convergence implies boundedness.

Example 2.2.1

Assume that the energy space $V$ is compactly embedded in another Hilbert space $H$,

$$V \overset{c}{\hookrightarrow} H,$$

i.e.

$$u_n \rightharpoonup u \text{ in } V \Rightarrow u_n \rightarrow u \text{ in } H,$$

and,

$$|c(u,v)| \leq C \|u\|_H \|v\|_V,$$

i.e. $c(u,v)$ is continuous on weaker space $H \times V$. It follows then immediately from the definition that $c(u,v)$ is compact.

A specific example of such a scenario will be the Helmholtz problem,

$$\begin{cases} u \in H^1_0(\Omega) \\
(\nabla u, \nabla v) - \omega^2(u,v) = (f,v) \quad v \in H^1_0(\Omega)
\end{cases}$$

where, as usual, $(\cdot,\cdot)$ denotes the $L^2$-inner product.

We shall also make the following assumption.

Density assumption:

$$\forall v \in V \ \exists v_h \in V_h : \|v_h - v\|_V \rightarrow 0 \quad \text{as } h \rightarrow 0. \quad (2.6)$$

In other words, $\bigcup_h V_h$ is dense in $V$.

THEOREM 2.2.1 (Mikhlin, 1959)

Consider problem (2.4) and assume the density assumption (2.6). Assume additionally that operator $B$ corresponding to sesquilinear form $b(u,v)$ is an isomorphism. Then

$$\exists h_0 \ \exists \gamma_0 \ \forall h < h_0 \ (\forall v_h \in V_h) \ \sup_{v_h \in V_h} \frac{b(u_h,v_h)}{\|v_h\|} \geq \gamma_0 \|u_h\|. \quad (2.7)$$

In other words, the problem is asymptotically stable.

PROOF Assume, to the contrary, that

$$\forall h_0 \ \forall \gamma_0 \ \exists h < h_0 \ \exists v_h \in V_h \ \sup_{v_h \in V_h} \frac{b(u_h,v_h)}{\|v_h\|} < \gamma_0 \|u_h\|. $$
Set \( \gamma_n = 1/n, h_0 = 1/n \) to conclude existence of a sequence \( h_n < 1/n \) and the corresponding sequence of unit vectors \( \|u_{h_n}\| = 1 \) such that
\[
\sup_{v_h \in V_h} \frac{|b(u_{h_n}, v_h)|}{\|v_h\|} \leq \frac{1}{n}.
\]

Recall then that, in a Hilbert space, every bounded sequence has a weakly convergence subsequence. Replace the original sequence with the subsequence. We have thus \( u_{h_n} \to u_0 \). We claim that the sequence \( u_{h_n} \) converges to \( u_0 \) actually strongly. Indeed, coercivity of form \( a(u, v) \) implies:
\[
\alpha \|u_0 - u_{h_n}\|^2 \leq a(u_0 - u_{h_n}, u_0 - u_{h_n})
\]
\[
= b(u_0 - u_{h_n}, u_0 - u_{h_n}) - c(u_0 - u_{h_n}, u_0 - u_{h_n})
\]
\[
= b(u_0, u_0 - u_{h_n}) - b(u_{h_n}, u_0 - u_{h_n}) - c(u_0 - u_{h_n}, u_0 - u_{h_n})
\]
The first term converges to zero by definition of weak compactness \( (V, b(\cdot, \cdot) \in V') \), and the third one converges to zero by Remark 2.2.1. It remains to show that the second term converges to zero as well.

By density assumption, we can select a sequence \( w_{h_n} \to u_0 \). We have then,
\[
|b(u_{h_n}, u_0 - u_{h_n})| \leq |b(u_{h_n}, u_0 - w_{h_n}) + b(u_{h_n}, w_{h_n} - u_{h_n})|
\]
\[
\leq M \left( \|u_{h_n}\| \|u_0 - w_{h_n}\| + \sup_{v_{h_n} \to 0} \frac{|b(u_{h_n}, v_{h_n})|}{\|v_{h_n}\|} \right) \|w_{h_n} - u_{h_n}\|
\]
\[
\leq \frac{1}{n} \to 0
\]

Strong convergence of \( u_{h_n} \) to \( u_0 \) implies that \( 1 = \|u_{h_n}\| \to \|u_0\| \), so \( \|u_0\| = 1 \) and, therefore, \( u_0 \neq 0 \).

Consider then arbitrary \( v \) and a sequence \( v_{h_n} \) converging strongly to \( v \). By continuity of form \( b(u, v) \), we have,
\[
b(u, v) = \lim_{n \to \infty} b(u_{h_n}, v_{h_n})
\]
However,
\[
|b(u_{h_n}, v_{h_n})| \leq \sup_{v_{h_n} \in V_h} \frac{|b(u_{h_n}, v_{h_n})|}{\|v_{h_n}\|} \|v_{h_n}\| \leq \frac{1}{n} \to 0.
\]

Consequently,
\[
b(u_0, v) = 0 \quad \forall v \in V
\]

which contradicts the uniqueness of the solution (injectivity of \( B \)).

**Vibrations. A model problem.** Consider an abstract variational problem,
\[
\begin{cases}
  u \in V \\
  a(u, v) - \omega^2 m(u, v) = m(f, v) & v \in V
\end{cases}
\tag{2.8}
\]

where \( V \) is an energy space, hermitian, coercive form \( a(u, v) \) represents the elastic energy, hermitian and positive-definite form \( m(u, v) \) represents the mass, \( \omega \) is the forcing frequency, and \( f \) is a force per unit mass.
Consider the variational eigenproblem,
\[
\begin{align*}
\{ e_i & \in V \\
\ a(e_i, v) & = \omega_i^2 m(e_i, v) \quad v \in V
\end{align*}
\] (2.9)

If mass represents a compact perturbation of energy (case of a bounded domain), there exists an infinite number of eigenpairs \((\omega_i^2, e_i)\) with real and positive eigenvalues \(\omega_i^2 \to \infty\), and eigenvectors \(e_i\) providing an orthogonal (both in terms of mass and energy) basis for \(V\). We equip the energy space with the energy norm,
\[
\|u\|^2 := a(u, u) = \sum_i \omega_i^2 u_i^2
\]
and assume that the eigenvectors have been normalized with mass, i.e.
\[
m(e_i, e_i) = 1, \quad i = 1, 2, \ldots, .
\]

We will compute now explicitly the corresponding inf-sup constant \(\gamma\) and continuity constant \(M\) in terms of the eigenvalues \(\omega_i^2\). Let \(u, v \in V\) and
\[
u_i = \sum_i u_i e_i, \quad v_j = \sum_j v_j e_j
\]
be the corresponding spectral representations. We have,
\[
\sup \frac{|a(u, v) - \omega^2 m(u, v)|}{\|v\|} = \|a(u, \cdot) - \omega^2 m(u, \cdot)\|_{V'} = \|v\| \quad \text{where} \ v = R_v^{-1}(a(u, \cdot) - \omega^2 m(u, \cdot))
\]
with \(R_v\) denoting the Riesz operator. We get
\[
\omega_i^2 v_i = (\omega_i^2 - \omega^2) u_i
\]
and
\[
\|v\|^2 = \sum_{i=1}^{\infty} \omega_i^2 |v_i|^2 = \sum_{i=1}^{\infty} \omega_i^2 \left(\frac{\omega_i^2 - \omega^2}{\omega_i^2}\right)^2 |u_i|^2.
\]
The inf-sup constant \(\gamma\) satisfies:
\[
\sum_{i=1}^{\infty} \omega_i^2 \left(\frac{\omega_i^2 - \omega^2}{\omega_i^2}\right)^2 |u_i|^2 \geq \gamma^2 \sum_{i=1}^{\infty} \omega_i^2 |u_i|^2.
\]
Comparing coefficients, we get,
\[
\gamma = \min_i \frac{\omega_i^2 - \omega^2}{\omega_i^2}.
\]
Notice that, despite the infinite number of spectral components, the minimum is actually attained (explain, why?). Concerning the continuity constant, we have,
\[
|b(u, v)| = \left| \sum_i (\omega_i^2 - \omega^2) u_i \bar{v}_i \right| = \left| \sum_i \frac{\omega_i^2 - \omega^2}{\omega_i^2} \omega_i u_i \bar{v}_i \right| \leq \max_i \left| 1 - \left(\frac{\omega}{\omega_i}\right)^2 \right| \|u\| \|v\|
\]
We see that the continuity constant \(M\) is of order \(\omega^2\) whereas the inf-sup constant \(\gamma = 0\) if the forcing frequency \(\omega\) matches one of the eigenfrequencies \(\omega_i\) (resonance).
The same reasoning can be repeated for the discrete problem using discrete eigenpairs \((\omega_{h,i}^2, \epsilon_{h,i})\). This leads to the analogous formula for the discrete inf-sup constant \(\gamma_h\),

\[
\gamma_h^{-1} = \frac{1}{\min_i |1 - (\frac{\omega_{h,i}}{\omega_i})^2|}
\]

It is well known that the discrete eigenvalues \(\omega_{h,i}^2\) converge monotonically from above to the corresponding exact eigenvalues \(\omega_i^2\). Imagine that the forcing frequency \(\omega\) happens to be in between the exact eigenfrequency \(\omega_i\) and, for some mesh, the corresponding discrete eigenfrequency \(\omega_{h,i}\). As you keep refining (uniformly) the mesh, \(\omega_{h,i}\), marching towards \(\omega_i\), has to migrate over the forcing frequency \(\omega\). It may even hit \(\omega\) and the discrete problem will become then unstable (ill-posed). Or it can get so close to \(\omega\) that the round off error will make the discrete problem effectively singular. The moral of the story is that only once \(\omega_{h,i}\) migrates to the left side of \(\omega\), the danger of resonance (or quasi-resonance) is gone. From now on, the global refinements will lead to stable discrete problems. Of course, the criterion of being on the correct side of the forcing frequency must be satisfied for all eigenfrequencies. In short, the stability is related to the convergence of eigenfrequencies that are close to the forcing frequency.

You can also see that, eventually, the discrete inf-sup constant converges to the exact one. Asymptotically, stability of the discrete problem reflects stability of its continuous counterpart.

For problems with damping, with lose the orthogonality structure and such a characterization becomes impossible although the inf-sup constant can still be represented in terms of the singular values of the operator, and the discrete inf-sup constant still converges to the exact one, see [12].

**Remark 2.2.2** Being in the preasymptotic range is accompanied by a flip in a spectral component of the solution. Spectral components \(u_i\) of the solution to (2.8) are given by the formula:

\[
u_i = \frac{f_i}{\omega_i^2 - \omega^2}
\]

The same formula holds on the discrete level,

\[
u_{h,i} = \frac{f_{h,i}}{\omega_{h,i}^2 - \omega^2}
\]

If the sign of factor \(\omega_{h,i}^2 - \omega^2\) is different form that of \(\omega_i^2 - \omega^2\), the component will be ‘flipped’. If you are in a quasi-resonant mode, this may be the largest component of the solution. Do not look then for a bug in your code as I did 25 years ago loosing several days before I understood the phenomenon.

**Asymptotic optimality of the Galerkin method.** We will make just two assumptions: a/ the method converges in the energy norm implied by the leading Hermitian term:

\[
\|u\|^2_E := a(u, u),
\]
Beyond Coercivity

and b/ the method converges with a faster rate in the weaker norm $\|\cdot\|_H$ controlling the continuity of the compact perturbation term $c(u, v)$,

$$|c(u, v)| \leq C\|u\|_H\|v\|_E.$$  

We have then,

$$\|u - u_h\|_E^2 - C\|u - u_h\|_H\|u - u_h\|_E \leq |b(u - u_h, u - u_h)|$$

$$= |b(u - u_h, u - w_h)|$$

$$= a(u - u_h, u - w_h) + |c(u - u_h, u - w_h)|$$

$$\leq \|u - u_h\|_E\|u - w_h\|_E + C\|u - u_h\|_H\|u - w_h\|_E,$$

or,

$$\|u - u_h\|_E \leq \frac{\|u - u_h\|_E + C\|u - u_h\|_H}{\|u - u_h\|_E - C\|u - u_h\|_H} \inf_{w_h \in U_h} \|u - w_h\|_E.$$  

Due to the faster convergence in the weaker norm, the fraction on the right-hand side converges asymptotically to unity. Hence, asymptotically, the Galerkin error converges to the best approximation error.

Exercises

**Exercise 2.2.1** Convergence of discrete inf-sup constant to the continuous one. Prove that, under the assumptions of the Mikhlin Theory,

$$\gamma_h \to \gamma.$$  

*Hint:* Consult [12], if necessary.

(10 points)

**Exercise 2.2.2** Convergence of eigenvalues. Let $a(u, v)$ be a coercive Hermitian form on a Hilbert space $V$, and $m(u, v)$ a positive-definite form on the same space $V$. Let $V_h$ be a monotone sequence of finite dimensional subspaces of $V$, whose union is dense in $V$. Let $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \to \infty$ be a sequence of generalized eigenvalues with corresponding eigenvectors $e_n$, see (2.9). Let $\lambda_{h,n}$ be the corresponding (finite) sequence of approximate eigenvalues corresponding to space $V_h$. Prove that

$$\lambda_{h,n} \to \lambda_n \quad \text{as} \quad h \to 0 \quad \text{(2.10)}$$

for every $n$.

*Hint:* Use Rayleigh quotient to prove the result for the first eigenvalue, and Riesz-Fischer formula (see [28]) for next eigenvalues.

(10 points)
2.3 Mixed Problems

**Constrained minimization problems.** Consider the standard (potential energy) functional defined on a Hilbert space $V$,

$$J(v) = \frac{1}{2}a(v, v) - \Re f(v), \quad v \in V$$

where $f \in V'$, and $a(u, v)$ is a sesquilinear, hermitian, coercive form on $V \times V$,

$$a(v, v) \geq \alpha \|v\|^2_V, \quad v \in V, \quad \alpha > 0.$$  

Let $Q$ be another Hilbert space, and $b(v, q), \ v \in V, q \in Q$ denote another sesquilinear form. Consider a constrained minimization problem,

$$\inf_{v \in V_g} J(v)$$

where

$$V_g := \{v \in V : b(v, q) = g(q) \quad \forall q \in Q\},$$

with a given $g \in Q'$. Recall that, for complex setting,

$$b(v, q) - g(q) = 0, \quad q \in Q \iff \Re(b(v, q) - g(q)) = 0, \quad q \in Q.$$  

In order to derive necessary conditions for the minimizer, introduce the Lagrangian,

$$L(v, q) := J(v) + \Re(b(v, q) - g(q))$$

and differentiate it with respect to $v$ and $q$ to obtain,

$$\begin{cases} 
  u \in V, \ p \in Q \\
  \Re(a(u, v) + b^*(p, v) - f(v)) = 0 \quad v \in V \\
  \Re(b(u, q) - g(q)) = 0 \quad q \in Q 
\end{cases}$$

or, equivalently,

$$\begin{cases} 
  u \in V, \ p \in Q \\
  a(u, v) + b^*(p, v) = f(v) \quad v \in V \\
  b(u, q) = g(q) \quad q \in Q 
\end{cases} \quad (2.11)$$

with $b^*(p, v) = \overline{b(v, p)}$. Problem (2.11) is identified as a mixed problem to be solved for the minimizer $u$ and the Lagrange multiplier $p$. Eventually, we extend our interest to a larger class of mixed problems where form $a(u, v)$ may be neither hermitian nor coercive.

The mixed problem can be cast into the standard variational setting by introducing the group variables,

$$u := (u, p) \in V \times Q, \quad v := (v, q) \in V \times Q$$
Beyond Coercivity

and a “big” sesquilinear form,

\[ b(u, v) := a(u, v) + b^*(p, v) + b(u, q) = a(u, v) + \overline{b(v, p)} + b(u, q). \]

Mixed problem (2.11) is then equivalent to,

\[
\begin{align*}
\begin{cases}
u \in V \times Q \\
b(u, v) = l(v) \quad v \in V \times Q
\end{cases}
\end{align*}
\]

where \( l(v) := f(v) + g(q). \)

Babuška \Rightarrow Brezzi. Our main tool in deriving the famous Brezzi’s conditions [7, 20] will be the following fundamental property of any sesquilinear (bilinear) continuous form \( c(q, v) \) defined on a pair of Hilbert spaces \( V, Q \),

\[
\inf \sup_{p \in Q} \frac{|c(p, v)|}{\|p\|_Q \|v\|_V} \leq \inf \sup_{[v] \in V/V_0} \frac{|c(p, v)|}{\|p\|_Q \|[v]\|_{V/V_0}}
\]

where \( V_0 := \{ v \in V : c(p, v) = 0 \quad \forall p \in Q \} \) and \( V/V_0 \) is the quotient space whose elements are equivalence classes,

\[
[v] = v + V_0, \quad \|[v]\|_{V/V_0} := \inf_{w \in [v]} \|w\|_V,
\]

see [28] for details.

We shall discuss now how the assumptions of Babuška-Nečas and Babuška Theorems translate into appropriate assumptions on forms \( a(u, v), b(u, q) \). We shall assume that “big” sesquilinear form satisfies the inf-sup condition,

\[
\inf_{u} \sup_{v} \frac{|b(u, v)|}{\|u\| \|v\|} := \gamma > 0.
\]

Setting \( u = (0, p) \) in (2.14), we get,

\[
\sup_{(v, q)} \frac{|b^*(p, v)|}{\|v\|} = \sup_{v} \frac{|b^*(p, v)|}{\|v\|} = \sup_{v} \frac{|b(v, p)|}{\|v\|} \geq \gamma \|p\|.
\]

Condition:

\[
\sup_{v} \frac{|b(v, p)|}{\|v\|} \geq \beta \|p\|, \quad p \in Q, \quad \beta > 0
\]

is the famous BB (Babuška-Brezzi)* or the inf-sup condition relating spaces \( V \) and \( Q \). Note that \( \beta \geq \gamma \).

The inf-sup condition for form \( b(u, v) \) implies uniqueness,

\[
b(u, v) = 0 \quad \forall v \quad \Rightarrow \quad u = 0
\]

Testing with \( v = (0, q) \), we learn that \( u = u_0 \in V_0 \). Condition above implies thus,

\[
a(u_0, v) + b^*(p, v) = 0 \quad \forall v \quad \Rightarrow \quad u_0 = 0 \quad \text{and} \quad p = 0.
\]

*Sometimes also called the LBB (Ladyshenskaya-Babuška-Brezzi) condition.
Assume now that
\[ a(u_0, v_0) = 0 \quad \forall v_0 \in V_0. \tag{2.17} \]

The BB condition (2.15) implies now that there exists a unique \( p \in Q \) such that
\[
\begin{cases}
  p \in Q \\
  b^*(p, v) = -a(u_0, v) \quad v \in V
\end{cases}
\]

Indeed, according to assumption (2.17), the right-hand side in the equation above satisfies the required compatibility condition. The pair \((u_0, p)\) satisfies thus the assumption in the uniqueness condition (2.16) and, therefore \( u_0 = 0 \). In other words, we have the uniqueness in kernel condition:
\[ a(u_0, v_0) = 0 \quad \forall v_0 \in V_0 \quad \Rightarrow \quad u_0 = 0 \tag{2.18} \]
i.e. operator
\[ A_0 : V_0 \to V_0', \quad \langle A_0 u_0, v_0 \rangle = a(u_0, v_0), \quad u_0, v_0 \in V_0 \]
is injective.

Next, restricting ourselves in (2.14) to \( u = (u_0, p) \), \( u_0 \in V_0 \), we have,
\[
\sup_{(v, q)} \frac{|a(u_0, v) + b^*(p, v)|}{\|v\|^2 + \|q\|^2}^{1/2} = \sup_v \frac{|a(u_0, v) + b^*(p, v)|}{\|v\|^2} \geq \gamma (\|u_0\|^2 + \|p\|^2)^{1/2}.
\]

Therefore,
\[
\inf_{u_0 \in V_0, p \in Q} \sup_{v \in V} \frac{|a(u_0, v) + b^*(p, v)|}{\|u_0\|^2 + \|p\|^2}^{1/2} \geq \gamma
\]
where
\[ V_{00} = \{ v \in V : a(u_0, v) + b^*(p, v) = 0 \quad \forall u_0 \in V_0, \forall p \in Q \}
\[
= \{ v_0 \in V : a(u_0, v_0) = 0 \quad \forall u_0 \in V_0 \}.
\]

Consequently,
\[
\inf_{[v_0] \in V_0/V_{00}} \sup_{u_0 \in V_0, p \in Q} \frac{|a(u_0, v_0)|}{\|u_0\|^2 + \|p\|^2}^{1/2} \geq \gamma
\]

Finally, uniqueness in kernel (2.18) implies that
\[
\inf_{u_0 \in V_0} \frac{|a(u_0, v_0)|}{\|u_0\| \|v_0\|} = \inf_{[v_0] \in V_0/V_{00}} \sup_{u_0 \in L_0} \frac{|a(u_0, v_0)|}{\|u_0\| \|v_0\|} \geq \gamma.
\]
i.e. the inf-sup in kernel condition holds:
\[ \sup_{v_0 \in V_0} \frac{|a(u_0, v_0)|}{\|v_0\|} \geq \alpha \|u_0\|, \quad u_0 \in V_0 \tag{2.19} \]

with \( \alpha \geq \gamma \).
On the discrete level, uniqueness implies existence for any right-hand side. Note that on the continuous level the null space of the transpose operator is:

\[
\{ v : b(u, v) = 0 \quad \forall u \} = \{(v, q) \in V \times Q : a(u, v) + b^*(p, v) + b(u, q) = 0 \quad \forall u \in V, \forall p \in Q \}
\]

\[
= \{(v_0, q) \in V_0 \times Q : a(u, v_0) + b(u, q) = 0 \quad \forall u \in V \}
\]

\[
= \{(v_0, q_0) \in V_{00} \times Q \}
\]

where, in the last line, \( q_0 \in Q \) is the unique solution of the problem,

\[
\begin{cases}
q_0 \in Q \\
b(u, q_0) = -a(u, v_0) \quad \forall u \in V, \quad v_0 \in V_{00}.
\end{cases}
\]

In order for the mixed problem to have a solution, the right-hand side must satisfy the compatibility condition:

\[ f(v_0) + g(q_0) = 0 \quad \forall v_0 \in V_{00}. \tag{2.20} \]

**Brezzi \Rightarrow Babuška.** Assume now that Brezzi’s conditions (2.19) and (2.15) hold. We shall demonstrate now that the two “small” inf-sup conditions imply that the “big” condition (2.14) must be satisfied as well. Given \((u, p) \in V \times Q\), define,

\[
f(v) := a(u, v) + b^*(p, v) \quad v \in V
\]

\[
g(v) := b(u, v) \quad q \in Q
\]

We need to demonstrate that we control \(\|u\|, \|p\|\) by norms of \(f, g\).

The BB condition implies that

\[
\inf_{v} \sup_{q} \frac{|b(v, q)|}{\|v\|_{V/V_0} \|q\|} = \inf_{v} \sup_{q} \frac{|b(v, q)|}{\|v\| \|q\|} = \beta > 0,
\]

Consequently,

\[
\|u\|_{V/V_0} = \inf_{w \in V_0} \|u - w\|_V = \|u - u_0\| \leq \frac{1}{\beta} \|g\|_{Q'}
\]

where \(u_0\) is the \(V\)-orthogonal projection of \(u\) onto \(V_0\). Now,

\[
a(u, v_0) = a(u - u_0, v_0) + a(u_0, v_0) = f(v_0) \quad v_0 \in V_0
\]

and the inf-sup in kernel condition gives:

\[
\|u_0\| \leq \frac{1}{\alpha} (\|a\| \|u - u_0\| + \|f\|_{V'}) \leq \frac{1}{\alpha} (\|a\| \|g\|_{Q'} + \|f\|_{V'}). 
\]

Consequently,

\[
\|u\| \leq \|u_0\| + \|u - u_0\| \leq \frac{1}{\alpha} (\|a\| \|u - u_0\| + \|f\|_{V'}) + \frac{1}{\beta} (1 + \frac{\|a\|}{\alpha}) \|g\|_{Q'}.
\tag{2.21}
\]

Finally, we can use the first equation and the BB condition, to control the Lagrange multiplier \(p\).

\[
b^*(p, v) = f(v) - a(u, v)
\]
implies
\[ \|p\| \leq \frac{1}{\beta} (\|f\|_{V'} + \|a\| \|u\|) \]
\[ \leq \frac{1}{\beta} (1 + \frac{\|a\|}{\alpha}) \|f\|_{V'} + \frac{\|a\|}{\beta^2} \|1 + \frac{\|a\|}{\alpha}\|\|g\|_{Q'}. \]

We can formulate now the famous Brezzi Theorem.

**THEOREM 2.3.1 Brezzi, 1973**

Assume that Brezzi’s conditions (2.19) and (2.15) hold on both continuous and discrete level and that discrete inf-sup constants remain uniformly bounded away from zero,
\[ \beta_h \geq \beta_0 > 0 \quad \alpha_h \geq \alpha_0 > 0. \]

Let \( f \in V' \), \( g \in Q' \) satisfy the compatibility condition (2.20). Then both continuous and discrete problems are well posed, i.e. there exist unique solutions \((u, p)\) and \((u_h, p_h)\) and stability constants \( \gamma = \gamma(\alpha, \beta, \|a\|) \) and \( \gamma_0 = \gamma(\alpha_0, \beta_0, \|a\|) \) such that
\[ \|(u, p)\| \leq \frac{1}{\gamma} (\|f\|_{V'}^2 + \|g\|_{Q'}^2)^{1/2} \]
and
\[ \|(u_h, p_h)\| \leq \frac{1}{\gamma_0} (\|f\|_{V'}^2 + \|g\|_{Q'}^2)^{1/2}. \]

Moreover, the following error estimate holds:
\[ \left( \|v - v_h\|_V^2 + \|p - p_h\|_Q^2 \right) \leq \frac{\|b\|}{\gamma_0} \left( \inf_{w_h \in V_h} \|v - w_h\|_V^2 + \inf_{p_h \in Q_h} \|p - p_h\|_Q^2 \right). \]  
(2.23)

**REMARK 2.3.1** Continuity constant \( \|b\| \) can be easily bounded by the continuity constants for the small forms, e.g. \( \|b\| \leq \|a\| + 2\|b\|. \) We have shown that Brezzi’s conditions are not only sufficient but also necessary for the well-posedness of the mixed problem, discrete stability and convergence.

As usual, the inf-sup condition on the continuous level does not imply the corresponding discrete inf-sup condition. However, there is a general tool that helps to relate the two conditions.

**Fortin operator.** Let \( b(v, q) \) be a bilinear or sesquilinear form defined on a pair of Hilbert spaces \( V, Q \). Assume that \( b(v, p) \) satisfies the inf-sup condition,
\[ \sup_{v \in V} \frac{|b(v, p)|}{\|v\|_V} \geq \beta \|p\|_Q, \quad p \in Q, \quad \beta > 0. \]  
(2.24)

Let \( V_h \subset V, Q_h \subset Q \) be a pair of discrete spaces. A continuous operator
\[ \Pi_h : V \to V_h, \quad \|\Pi_h v\|_V \leq \|\Pi_h\| \|v\|_V, \]

is called a Fortin operator, if the following discrete orthogonality condition holds:

\[ b(v - \Pi_h v, p_h) = 0 \quad v \in V, \ p_h \in Q_h. \]  

(2.25)

We have then,

\[ \sup_{v_h \in V_h} \frac{|b(v_h, p_h)|}{\|v\|_V} \geq \sup_{v \in V} \frac{|b(\Pi_h v, P_h)|}{\|\Pi_h v\|_V} \]

\[ = \sup_{v \in V} \frac{|b(v, p_h)|}{\|v\|_V} \frac{\|v\|_V}{\|P_h v\|_V} \]

\[ \geq \frac{\beta}{\|\Pi_h\|} \|p_h\|_Q. \]

In other words, the discrete inf-sup condition holds with a discrete inf-sup constant \( \beta_h \geq \beta / \|\Pi_h\| \).

The concept of Fortin operator provides a general framework for proving discrete stability but a concrete construction of such operator is problem-dependent. Note that the Fortin operator needs to be defined on the whole energy space and, therefore, one cannot use standard interpolation operators that are defined typically only for sufficiently regular functions.

We shall show an example of such construction for the Stokes problem.

**Clément Interpolation** [10, 9]. Let \( \Omega \) be a polygonal (polyhedral) domain partitioned into affine simplicial elements. Consider standard Lagrange elements of order \( p \). For a Lagrangian node \( a_i \), let \( \Omega_i \) denote the support of the corresponding basis function \( e_i \). Let \( u \in L^2(\Omega) \), and let \( u^p_i \in P^p(\Omega_i) \) be the \( L^2 \)-projection of function \( u \) onto polynomials of order \( p \) on the element patch \( \Omega_i \),

\[
\begin{cases}
  u^p_i \in P^p(\Omega_i) \\
  \int_{\Omega_i} (u - u^p_i) \phi = 0 \quad \forall \phi \in P^p(\Omega_i).
\end{cases}
\]

The Clément interpolation is defined similarly to the Lagrange interpolation except that pointwise values \( u(a_i) \) are replaced with values of the corresponding projections \( u^p_i(a_i) \),

\[ \Pi_h u := \sum_i u^p_i(a_i)e_i. \]

Notice that the operator is non-local as, for an element \( K \), \( \Pi u \) depends upon values of \( u \) in the patch of all elements adjacent to \( K \).

Does the non-local operator preserve polynomials \( u \in P^p(\Omega) \)? Does it preserve piecewise (elementwise) polynomials? Can we use the standard scaling arguments to estimate the interpolation error?

Let \( \Omega \subset \mathbb{R}^2 \) be a polygonal domain meshed with affine triangles satisfying the usual shape regularity assumptions. The following interpolation error estimate holds (Exercise 2.3.2):

\[
\sum_K \left( \|u - \Pi_h u\|_{L^2(K)}^2 h_K^2 + |u - \Pi_h u|_{H^1(K)}^2 \right) \leq C^2 \|u\|_{H^1(\Omega)}^2 \quad u \in H^1(\Omega), \ C > 0
\]

(2.26)

Due to the non-locality of Clément Interpolation, the estimate above is formulated globally. It says in a fancy way that the \( L^2 \) norm of interpolation error converges with a linear rate while its \( H^1 \)-seminorm remains bounded.
Example of a stable pair for the Stokes problem. We can give now perhaps the simplest example of a stable of pair of elements for the Stokes problem. Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain with a boundary split into non-zero measure parts of $\Gamma_u$ and $\Gamma_t$. Define now:

$$V := \{ u \in (H^1(\Omega))^2 : u = 0 \text{ on } \Gamma_u \}$$

$$Q := L^2(\Omega)$$

$$a(u,v) := \mu \int_{\Omega} \nabla u : \nabla v \quad u,v \in V, \quad \mu > 0$$

$$b(u,q) := \int_{\Omega} \text{div } u q \quad u \in V, \quad q \in Q$$

$$f(v) := \int_{\Omega} fv + \int_{\Gamma_t} tv$$

$$g(v) := \int_{\Omega} gv$$

where $f,g \in L^2(\Omega)$, $t \in L^2(\Gamma_t)$ are given. All spaces are real.

The essential BC on $\Gamma_u$ and the Poincaré inequality imply that form $a(u,v)$ is coercive on the whole space $V$. It is a very non-trivial result to show that the inf-sup condition holds on the continuous level. The continuous problem is thus well-posed.

Let us discretize now the velocities with quadratic triangular Lagrange elements, and the pressure with piecewise constants defined on the same mesh. We will use now the Fortin “trick” to demonstrate that the pair satisfies the discrete inf-sup condition.

Define now a candidate for the Fortin operator,

$$\Pi_h v = \Pi_1 v + \Pi_2 (u - \Pi_1 u), \quad u \in H^1(\Omega) \quad (2.27)$$

where $\Pi_1$ is the Clément operator and

$$\Pi_2 : (H^1(K))^2 \to (P^2(K))^2$$

is a local operator defined by requesting two conditions:

$$\Pi_2 u = 0 \quad \text{at vertices and } \int_e (u - \Pi_2 u) = 0 \quad \text{for each edge } e.$$

Note that

$$u - \Pi_h u = (u - \Pi_1 u) - \Pi_2 (u - \Pi_1 u) = (I - \Pi_2)(u - \Pi_1 u)$$

It follows now from the construction of $\Pi_2$ that $\Pi_h u$ satisfies the homogeneous BC on $\Gamma_u$ if so does $u$. Indeed, for each edge $e \subset \Gamma_u$,

$$\int_e (u - \Pi_1 u) = \int_e \Pi_2 (u - \Pi_1 u)$$

implies,

$$0 = \int_e u = \int_e (\Pi_1 u + \Pi_2 (u - \Pi_1 u)) = \int_e \Pi_h u$$

which, in turn, implies that $\Pi_h u = 0$ on edge $e$. 
It follows also from the construction of $\Pi_2$ that, for a constant $q$,
\[ \int_K \text{div}(u - \Pi_h u)q = q \sum_e \int_e (u - \Pi_h u) \cdot n = q \sum_e (I - \Pi_2)(u - \Pi_1 u) \cdot n = 0. \]

At the same time a standard scaling argument implies that
\[ \|\Pi_2 v\|_{L^2(K)}^2 + |\Pi_2 v|_{H^1(K)}^2 \lesssim h_K^2 \|\Pi_2 \hat{v}\|_{L^2(K)}^2 + |\Pi_2 \hat{v}|_{H^1(K)}^2 \]
\[ = h_K^2 \|\Pi_2 \hat{v}\|_{L^2(K)}^2 + |\Pi_2 \hat{v}|_{H^1(K)}^2 \]
\[ \lesssim \|\hat{v}\|_{H^1(K)}^2 \]
\[ \lesssim C(h_K^{-2} \|v\|_{L^2(K)}^2 + |v|_{H^1(K)}^2) \]

where, as usual, $A \lesssim B$ means existence of a constant $C$ (independent of element and function $u$) such that $A \leq C B$. The estimate above, combined with estimate (2.26), proves the continuity of operator $\Pi_h$.

**Time-harmonic Maxwell equations as an example of a mixed problem.** Another example of a mixed problem is provided by the Maxwell equations. Recall the stabilized variational formulation for time-harmonic Maxwell equations:
\[
\begin{aligned}
& E \in H_0^1(\Omega), p \in H_0^1(\Omega) \\
& \int_K \frac{1}{\mu} \nabla \times E \cdot \nabla \times F - \omega^2 \int_K \epsilon E \cdot F + \int_K \nabla p \cdot F = -i\omega \int_K J^{\text{imp}} \cdot F \\
& \int_K \epsilon E \cdot \nabla q = 0 \\
& F \in H_0^1(\Omega) \\
& q \in H_0^1(\Omega)
\end{aligned}
\]
(2.28)

where $\text{div} J^{\text{imp}} = 0$ and,
\[ 0 < \mu_0 \leq \mu \leq \mu_\infty < \infty \quad 0 < \epsilon_0 \leq \epsilon \leq \epsilon_\infty < \infty. \]

Using the notation for the mixed problems, we have,
\[ a(E, F) := \int_K \frac{1}{\mu} \nabla \times E \cdot \nabla \times F - \omega^2 \int_K \epsilon E \cdot F \\
\]
\[ b(E, q) := \int_K \epsilon E \cdot \nabla q. \]

Comparing with the Stokes problem, the difficulties are completely reversed. For Stokes, form $a(u, v)$ was $V$-coercive (so the inf-sup in kernel condition was trivially satisfied) but proving the BB condition was a challenge. For Maxwell, the BB condition is simple as it is a direct consequence of the exact sequence property. Indeed, with the homogeeous BCs, the standard norm in $H_0^1(\Omega)$ is equivalent to the $H^1$-seminorm,
\[ \|q\|_{H^1(\Omega)}^2 \sim \int_\Omega |\nabla q|^2. \]

We have now,
\[ \sup_F \frac{|\int_K \epsilon \nabla p \cdot F|}{\|F\|_{H(\text{curl}, \Omega)}} \geq \frac{|\int_\Omega \epsilon \nabla p \cdot \nabla p|}{\|\nabla p\|_{L^2(\Omega)}} \geq \epsilon_0 \|\nabla p\|_{L^2} \sim \epsilon_0 \|p\|_{H^1(\Omega)} \]
since $\nabla H^0_0(\Omega) \subset H(\text{curl}, \Omega)$. Note that the same reasoning applies to the discrete problem.

On the other side, proving the discrete inf-sup in kernel condition is a challenge, as we will see later.

**Exercises**

**Exercise 2.3.1** In the Hilbert space setting, it is elegant to preserve the Hilbert space structure by introducing the Eucklidean norm for the group variable:

$$
\|u\|^2 = \|(u, p)\|^2 := \|u\|^2_V + \|p\|^2_Q.
$$

Revisit the reasoning in the text in an attempt to derive sharper bounds for Babuška’s inf-sup constant $\gamma$ for form $b(u, v)$ in terms of Brezzi’s inf-sup constants $\alpha, \beta$ (and continuity constant $\|a\|$) using the Eucklidean norms for $u, v$. *Hint:* By Pythagoras Theorem we have,

$$
\|u\|^2 = \|u_0\|^2 + \|u - u_0\|^2.
$$

(15 points)

**Exercise 2.3.2** Clément Interpolation. Restrict yourself for simplicity to 2D case and affine linear triangles. Assume $v \in H^1(\Omega)$, and prove the following interpolation error estimate(s),

$$
\sum_K h_K^{2(r-1)} |v - \Pi v|^2_{H^r(K)} \leq C \|v\|^2_{H^1(\Omega)} \quad r = 0, 1
$$

where $C$ is a generic (shape regularity) constant independent of $v$. Consult [10, 9] if necessary.

(15 points)

**Exercise 2.3.3** Clément Interpolation - cont. Is the use of Lagrange d.o.f. in Clément’s definition essential? Attempt to generalize the operator using moments discussed in Exercise 1.8.5.

(5 points)

**Exercise 2.3.4** Stokes problem with pure kinematic boundary conditions. Consider the Stokes problem with kinematic homogeneous BC implied on the whole boundary, i.e.

$$
V = (H^1_0(\Omega))^N, \quad N = 2, 3
$$

Note that the pressure cannot now be unique as, for any constant $p$,

$$
\int_{\Omega} p \text{ div } v = p \int_{\Gamma} v \cdot n = 0 \quad \forall v \in V.
$$

This leads to the modification of space $Q$, the $L^2$-space is replaced with the quotient space $L^2(\Omega)/\mathbb{R}$ that is isomorphic and isometric with the subspace of $L^2(\Omega)$ consisting of functions of zero average,

$$
Q = L^2(\Omega)/\mathbb{R} \sim L^2_0(\Omega) := \{q \in L^2(\Omega) : \int_{\Omega} q = 0\}.
$$
Use the Brezzi Theorem to prove that the problem is well posed. Does the Fortin operator discussed in the text work for this problem as well?

(5 points)

2.4 Non-uniform Meshes

As we have learned, in presence of singularities, the rate of convergence for uniform \( h \)-refinements, is limited not by the polynomial degree but rather by the global regularity of the solution expressed in terms of Sobolev norms. In this section, we will present the fundamental result of Babuška, Kellogg and Pitkäranta [3] demonstrating that, by using properly designed non-uniform meshes graded towards the singular points, one can restore the optimal rate of convergence dictated by the polynomial degree \( p \) alone. In practice, the meshes are obtained by using a-posteriori error estimates and automatic \( h \)-adaptivity. The proof will deal with a simple 2D model problem, triangular elements, and polynomial order \( p = 1 \) only, but the result has been numerically confirmed for elements of all shapes, and 3D elliptic problems [14, 18].

Let \( \Omega \) be a two-dimensional polygonal domain, see Fig. 2.2, with vertices \( x_i \), and corresponding internal angles \( \theta_i \), \( i = 1, \ldots, M \). Let boundary \( \Gamma \) be partitioned into a Dirichlet boundary \( \Gamma_D \) and Neumann boundary \( \Gamma_N \). Note that \( \Gamma_D \) may be terminated inside of an edge in which case, the end point of \( \Gamma_D \) is classified also as a vertex, see vertex \( x_j \) in Fig. 2.2. For each vertex \( x_i \) introduce a parameter \( \alpha_i \),

\[
\alpha_i := \min \{ 1, \frac{\kappa_i \pi}{\theta_i} \}
\]

where \( \kappa_i = 1 \) if both sides of \( x_i \) are contained either in \( \Gamma_D \) or \( \Gamma_N \), and \( \kappa_i = 1/2 \) if vertex \( x_i \) is a transition point between the two parts of the boundary. Let \( \mathcal{M} \) denote the subset of ("singular") vertices for which coefficients \( \alpha_i \) are negative, Let \( \beta := (\beta_1, \ldots, \beta_M) \) be a \( M \)-tuple of exponents associated with the vertices, \( \beta_i \in [0, 1) \). Consider the weight function:

\[
\phi_{\beta}(x) := \prod_{i=1}^{M} |x - x_i|^\beta_i
\]

where \( |x| \) denotes the Euclidean norm of vector \( x \). Let \( H^m(\Omega) \), \( m = 1, 2, \ldots \), denote the regular Sobolev spaces and

\[
H^1_D(\Omega) := \{ u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_D \}.
\]

We will consider the model problem:

\[
\begin{cases}
  u \in H^1_D(\Omega) \\
  \int_{\Omega} \nabla u \cdot \nabla v + uv = \int_{\Omega} fv \quad v \in H^1_D(\Omega).
\end{cases}
\]
In presence of singularities, regularity of the solution can be assessed using the weighted Sobolev norms:

\[ \|u\|_{H^{m,\beta}(\Omega)}^2 := \|u\|_{H^{m-1}(\Omega)}^2 + \int_{\Omega} \phi_\beta^2 \sum_{|\alpha|=m} |D^\alpha u|^2 \varphi_{\alpha}^2, \]

For \( \beta = 0 \), the norm coincides with the standard Sobolev norm. Completion of \( C^\infty(\overline{\Omega}) \) under the weighted norm is identified as the weighted Sobolev space, and denoted by \( H^{m,\beta}(\Omega) \). Note that the weight applies only to the highest order derivatives. One can prove the continuous embedding:

\[ H^{m,\beta}(\Omega) \hookrightarrow C^{m-2}(\overline{\Omega}), \quad m \geq 2. \]

The following regularity result has been established in [3].

**THEOREM 2.4.1**

Assume that

\[ 1 - \alpha_i \leq \beta_i < 1, \quad x_i \in \mathcal{M}, \]

and \( f \in H^{0,\beta}(\Omega) \). The solution \( u \) lives then in \( H^1_D(\Omega) \cap H^{2,\beta}(\Omega) \), and

\[ \|u\|_{H^2(\Omega)} \leq C \|f\|_{H^{0,\beta}(\Omega)} \]

with stability constant \( C \) independent of \( f \).

Note that, for each non-singular vertex \( x_i \), we can select \( \beta_i = 0 \).

The considered model problem is (trivially) \( H^1(\Omega) \)-coercive and the convergence analysis reduces to the interpolation error estimates. Following [3], we will consider a special class of non-uniform meshes whose...
density is controlled by a weight function $\phi_\gamma$. Hereafter $\gamma$ will denote a generic $M$-tuple; $\gamma = \beta$ for the problem of interest.

**Definition.** Let $h, L > 0$. Triangulation $T$ is of type $(h, \gamma, L)$ if the following three conditions are satisfied.

(i) Minimum angle condition:

$$\theta \geq L^{-1} \quad \forall \text{ angle } \theta \in T, \quad \forall \text{ element } T \in T.$$ 

(ii) Control of element size for elements with positive weight. If $\phi_\gamma \neq 0$ on $\bar{T}$, then

$$L^{-1} h \sup_{x \in T} \phi_\gamma(x) \leq d_T \leq L h \inf_{x \in T} \phi_\gamma(x).$$

(ii) Control of element size for elements with vanishing weight. If $\phi_\gamma = 0$ at some point in $\bar{T}$, then

$$L^{-1} h \sup_{x \in T} \phi_\gamma(x) \leq d_T \leq L h \sup_{x \in T} \phi_\gamma(x).$$

As usual, $d_T$ denotes the element diameter,

$$d_T := \sup x, y \in T | x - y |.$$ 

Although several results discussed next will apply to a general $\gamma$, the final interpolation error estimate will be applied to $\gamma = \beta$, with $\beta_i > 0$ only at singular vertices $x_i \in M$. Consequently, case (ii) above applies to elements that are not adjacent to a singular vertex, and case (iii) deals with elements sharing a singular vertex. For the domain illustrated in Fig.2.2, we have only three “singular” vertices: $x_i$ with a reentrant corner $\gamma_i > \pi$, and two transition points (including $x_j$) between Dirichlet and Neumann parts of the boundary.

**LEMMA 2.4.1**

The following inequality holds:

$$\int_0^1 s^{-2} |v(s)|^2 \, ds \leq 4 \int_0^1 |v'(s)|^2 \, ds,$$ 

for every $v \in H^1(0, 1)$ such that $v(0) = 0$.

Notice that, by 1D Poincaré inequality, the $L^2$-norm of $v$ is bounded by the $L^2$-norm of $v'$. The estimate says that we control the stronger, weighted (with singular weight $s^{-2}$) $L^2$-norm of $v$ as well.

**PROOF** The main tool in the proof is the *Integral Minkowski Inequality* (see [28], p. 409),

$$\left( \int_0^1 \left( \int_0^1 f(t, s) \, ds \right)^2 \, dt \right)^{1/2} \leq \int_0^1 \left( \int_0^1 f(t, s) \, dt \right)^{1/2} \, ds.$$
Representing $v(x)$ in terms of its derivative,

$$v(x) = \int_0^x v'(t) \, dt,$$

we have:

$$\int_0^1 s^{-2} \left| \int_0^s v'(t) \, dt \right|^2 ds = \int_0^1 \left| \int_0^s v'(su) \, du \right|^2 ds \leq \left[ \int_0^1 \left( \int_0^1 |v'(su)|^2 ds \right)^{1/2} du \right]^2 \quad \text{(change of variables: } t = su)$$

$$\leq \left[ \int_0^1 \left( \int_0^1 |v'(t)|^2 dt \right)^{1/2} du \right]^2 \quad \text{(Integral Minkowski inequality)}$$

$$\leq \int_0^1 \frac{1}{u^{1/4}} \int_0^1 \frac{1}{u^{1/4}} |v'(t)|^2 dt du \quad \text{(change of variables: } t = su)$$

$$\leq \int_0^1 u^{-1/2} du \int_0^1 |v'(t)|^2 \int_t^1 u^{-1/2} du \, dt \quad \text{(Cauchy-Schwarz inequality)}$$

$$= \int_0^1 u^{-1/2} du \int_0^1 |v'(t)|^2 \int_t^1 u^{-1/2} du \, dt . \quad \text{(Fubini’s Theorem)}$$

Finally,

$$\int_1^t u^{-1/2} du = 2u^{1/2} \big|_1^t = 2(1 - \sqrt{t}) \leq 2, \quad \text{and} \quad \int_0^1 u^{-1/2} du = 2u^{1/2} \big|_0^1 = 2 .$$

\[ \square \]

**Lemma 2.4.2**

The following inequality holds:

$$\int_0^1 t^\alpha - 2 [z(t) - a]^2 dt \leq C(\alpha) \int_0^1 t^\alpha |z'(t)|^2 dt, \quad \alpha \neq 1 \quad (2.32)$$

where

$$a = \begin{cases} z(0) & \text{for } \alpha < 1 \\ z(1) & \text{for } \alpha > 1 . \end{cases}$$

**Proof**

**Case:** $\alpha < 1$.

Using the change of variables:

$$t^{1-\alpha} = s, \quad s^{1-\alpha} = t, \quad (1 - \alpha)t^{-\alpha} dt = ds ,$$

we get,

$$\int_0^1 t^\alpha - 2 [z(t) - z(0)]^2 dt = \int_0^1 t^{2\alpha - 2} [z(t) - z(0)]^2 t^{-\alpha} dt = (1 - \alpha)^{-1} \int_0^1 s^{-2} [z(s^{1/\alpha} - z(0))^2 ds .$$
Using Lemma 2.4.1 and:
\[
\frac{dv}{ds} = z' \frac{1}{1 - \alpha} s^{\frac{\alpha}{\alpha - 1}},
\]
we can bound the last integral by
\[
4(1 - \alpha)^{-1} \int_0^1 \left| \frac{dv}{ds} \right|^2 ds.
\]
Finally, using:
\[
\frac{dv}{ds} = \frac{dz}{dt} \frac{1}{1 - \alpha} s^{\frac{\alpha}{\alpha - 1}}
\]
and returning to the original variable \( t \), we obtain the upper bound:
\[
\frac{1}{(1 - \alpha)^2} \int_0^1 \left| \frac{dz}{dt} \right|^2 s^{\frac{2\alpha}{\alpha - 1}} ds = \frac{1}{1 - \alpha} \int_0^1 \left| \frac{dz}{dt} \right|^2 t^\alpha dt.
\]

**Case:** \( \alpha > 1 \). Use change of variables: \( t = s^{1 - \alpha} \), and proceed along the lines of the first case, see Exercise 2.4.1.

**LEMMA 2.4.3**

Let \( \alpha \neq 0 \) and let \( T \) be the master triangle. There exists a constant \( C > 0 \) such that, for all \( u \) for which
\[
\int_T |x|^{\alpha} |\nabla u|^2 < \infty,
\]
there exists a constant \( a \) such that:
\[
\int_T |x|^{\alpha - 2} |u - a|^2 \leq C \int_T |x|^{\alpha} |\nabla u|^2, \quad \text{and}
\]
\[
|a|^2 \leq C \int_T |x|^{\alpha} |\nabla u|^2.
\]

For \( \alpha < 0 \) and continuous functions \( u, a = u(0) \).

**PROOF**

**Step 1:** We first prove the result for a quadrant of the unit circle:
\[
S := \{(r, \theta) : r < 1, \ 0 < \theta < \pi/2\}.
\]
Consider the average of \( u \) in \( \theta \),
\[
\bar{u}(r) = \frac{2}{\pi} \int_0^{\pi/2} u(r, \theta) d\theta.
\]
Let $0 < r_1 < r_2 < 1$. We have,

\[
|\bar{u}(r_2) - \bar{u}(r_1)| = \frac{2}{\pi} \left| \int_0^{\pi/2} u(r_2, \theta) - u(r_1, \theta) \, d\theta \right|
\]

\[
= \frac{2}{\pi} \left| \int_0^{\pi/2} \int_{r_1}^{r_2} r^{-\alpha+1} \frac{\partial u}{\partial r} \, dr \, d\theta \right|
\]

\[
\leq \frac{2}{\pi} \left( \int_{r_1}^{r_2} r^{-\alpha+1} \, dr \right)^{1/2} \left( \int_{r_1}^{r_2} r^\alpha \left| \frac{\partial u}{\partial r} \right|^2 \, dr \right)^{1/2} d\theta
\]

\[
\leq \left( \frac{2}{\pi} \int_{r_1}^{r_2} r^{-\alpha+1} \, dr \right)^{1/2} \left( \int_S r^\alpha |\nabla u|^2 \, dS \right)^{1/2}
\]

For $\alpha < 0$, integral $\int_0^{1} r^{-(\alpha+1)} \, dr$ is finite which implies that function $\bar{u}(r)$ is uniformly continuous and, therefore, admits a continuous extension to $r = 0$. For $\alpha > 0$, function $\bar{u}(r)$ is uniformly continuous in $[\epsilon, 1]$, for any $\epsilon > 0$.

We have now,

\[
\int_0^{1} r^\alpha \frac{d\bar{u}}{dr} \, dr \leq \frac{4}{\pi^2} \int_0^{1} r^\alpha \frac{\partial u}{\partial r} \, dr \, d\theta
\]

\[
\leq \frac{4}{\pi^2} \int_0^{1} r^{\alpha+1} \left( \frac{\pi}{2} \right)^2 \int_0^{\pi/2} \left| \frac{\partial u}{\partial r} \right|^2 \, dr
\]

\[
= \frac{4}{\pi^2} \int_S r^\alpha \left| \frac{\partial u}{\partial r} \right|^2 \, dS \leq \frac{4}{\pi^2} \int_S r^\alpha |\nabla u|^2 \, dS.
\]

Using Lemma 2.4.2, we obtain,

\[
\int_0^{1} r^{\alpha-1} |\bar{u}(r) - a|^2 \, dr \leq C \int_0^{1} r^{\alpha+1} |\frac{d\bar{u}}{dr}|^2 \, dr \leq C \int_S r^\alpha |\nabla u|^2 \, dS
\]

where $a = \bar{u}(0)$ for $\alpha < 0$, and $a = \bar{u}(1)$ for $\alpha > 0$. In either case, constant $a$ is bounded by $\int_S r^\alpha |\nabla u|^2 \, dS$. In addition, for $\alpha < 0$, if $u(r, \theta)$ is continuous on $\bar{S}$ then $a = u(0)$, comp. Exercise 2.4.3. Integrating in $\theta$, we get,

\[
\int_S r^{\alpha-2} |\bar{u} - a|^2 \, dS \leq C \int_S r^\alpha |\nabla u|^2 \, dS.
\]

The Intermediate Value Theorem implies that there exists an angle $\psi$ such that $\bar{u}(r) = u(r, \psi)$. Consequently,

\[
u(r, \phi) - \bar{u}(r) = u(r, \phi) - u(r, \psi) = \int_{\psi}^{\phi} \frac{\partial u}{\partial \theta}(r, \theta) \, d\theta
\]

\[
\leq C \left[ \int_0^{\pi/2} \left| \frac{\partial u}{\partial \theta}(r, \theta) \right|^2 \, d\theta \right]^{1/2}
\]

Integrating in $\phi$,

\[
\int_0^{\pi/2} |u(r, \phi) - \bar{u}(r)|^2 \, d\phi \leq C \int_0^{\pi/2} \left| \frac{\partial u}{\partial \theta}(r, \theta) \right|^2 \, d\theta.
\]
Finally, multiplying both sides with $r^{\alpha-2}$ and integrating in $r$, we obtain,

$$\int_0^1 r^{\alpha-2} \int_0^{\pi/2} |u(r, \phi) - \bar{u}(r)|^2 d\phi dr \leq C \int_0^1 r^{\alpha-1} \int_0^{\pi/2} \frac{1}{r} \frac{\partial u}{\partial \theta}^2 d\theta dr$$

$$= C \int_0^1 r^{\alpha} \int_0^{\pi/2} \frac{1}{r} \frac{\partial u}{\partial \theta}^2 d\theta rdr$$

$$\leq C \int_S r^\alpha |\nabla u|^2 dS.$$

Using triangle inequality, estimate (2.34) and the estimate above, we get the required result.

Step 2: Consider the map from the master element into the section $S$,

$$\begin{cases} r' = \frac{1}{a(\theta)} r \\ \theta' = \theta \end{cases}$$

where $a(\theta)$ is defined in Fig.2.3, and use change of variables.

![Mapping master element into the quadrant of a circle.](image)

**Figure 2.3**

Mapping master element into the quadrant of a circle.

**Lemma 2.4.4**

Let $\epsilon > 0$, $0 < s < 1$. There exists a constant $C = C(\epsilon, s)$ such that, for every triangle $T$ with vertices $v_0 = 0, v_1, v_2$ and a minimum angle $\geq \epsilon$, the following inequality holds:

$$\int_T |x|^{2s-1}|u-p|^2 + |x|^{2s-2}|d^2_x(u-p)|^2 \leq C \int_T |x|^{2s}|d^2_x u|^2,$$

(2.35)

for every function $u$ such that,

$$\int_T |u|^2 + |d^1_x u|^2 + |x|^{2s}|d^2_x u|^2 < \infty.$$
Case: arbitrary triangle

\[ |d^1_x u|^2 = |\nabla u|^2 = \sum_{|\alpha|=1} |D^\alpha u|^2, \quad |d^2_x u|^2 = \sum_{|\alpha|=2} |D^\alpha u|^2. \]

**Proof** Recall the earlier discussion on regularity of functions from the weighted Sobolev space to realize that functions \( u \) are continuous and, therefore, the vertex interpolant is well defined.

**Case:** \( T \) is the master triangle, \( v_1 = (1,0), v_2 = (0,1) \). Set \( \alpha = 2s \) in Lemma 2.4.3 to claim:

\[
\int_T |x|^{2s-2} \left| \frac{\partial u}{\partial x_i} - a_i \right|^2 \leq C \int_T |x|^{2s} |\nabla (\frac{\partial u}{\partial x_i})|^2 \leq C \int_T |x|^{2s} |d^2_x u|^2.
\]

Replacing now \( u \) with \( v = u - a_1 x_1 - a_2 x_2 \) and use Lemma 2.4.3 with \( \alpha = 2s - 2 \) to obtain:

\[
\int_T |x|^{2s-4} |v - v(0)|^2 \leq C \int_T |x|^{2s-2} |\nabla v|^2.
\]

Combining the two estimates, we get estimate (2.35) but with vertex interpolant \( p \) replaced with polynomial \( q = u(0) + a_1 x_1 + a_2 x_2 \). In order to correct the polynomial, consider function \( u_0 = u - q = v - v(0) \) and polynomial \( p_0 = p - q \). Note that \( p_0 = 0 \) at \( v_0 = 0 \), and

\[
p_0(v_1) = p_0((1,0)) = u(v_1) - (u(0) + a_1) = u_0(v_1).
\]

Similarly, \( p_0(v_2) = u_0(v_2) \). We have now,

\[
\int_T |x|^{2s-4} |p_0|^2 + |x|^{2s-2} |d^1_x p_0|^2 \leq C \left( |p_0(v_2)|^2 + |p_0(v_3)|^2 \right) \quad \text{(finite dimensionality argument)}
\]

\[
= C \left( |u_0(v_2)|^2 + |u_0(v_3)|^2 \right)
\]

\[
\leq C \int_T |u_0|^2 + |d^1_x u_0|^2 + |x|^{2s} |d^2_x u_0|^2 \quad \text{(continuous embedding argument)}
\]

\[
\leq C \int_T |x|^{2s} |d^2_x u|^2 \quad \text{(estimate for function \( u \))}
\]

Use triangle inequality and the estimates for \( u_0 = u - q \) and \( p_0 = p - q \) to arrive at the final estimate for \( u_0 - p_0 = u - p \).

**Case:** arbitrary triangle \( T \). Use linear map:

\[
x = B \xi
\]

with a non-singular matrix \( B \), and standard scaling argument, see Exercise 2.4.2.

**Theorem 2.4.2** Babuška, Kellogg, Pitkaränta, 1979

Let \( T \) be a triangulation of type \((h, \gamma, L)\). The following interpolation error estimate holds:

\[
\|u - \Pi u\|_{H^1(\Omega)} \leq Ch|u|_{H^{2+\gamma}(\Omega)} , \quad u \in H^{2+\gamma}(\Omega) \cap H^1_D(\Omega)
\]
where $C = C(\gamma, L)$, and $\Pi u$ denotes the linear vertex interpolant of $u$.

**PROOF**  We begin by recalling that space $H^{2, \gamma}(\Omega)$ is embedded in $C(\bar{\Omega})$ and, therefore, the vertex interpolant $v := \Pi_h u$ is well defined.

**Case:** element $T$ without a singular vertex.

The standard interpolation error estimate reads:

$$\|u - v\|_{H^1(T)}^2 \leq C d_T^2 |u|_{H^2(T)}^2$$

where constant $C$ depends only upon the minimal angle, element diameter $d_T$ satisfies the condition:

$$d_T \leq L h \inf_{x \in T} \phi_\gamma(x),$$

and, trivially,

$$L^2 h^2 \inf_{x \in T} \phi_\gamma^2(x) \int_T \sum_{|\alpha| = 2} |D^\alpha u|^2 \leq L^2 h^2 \int_T \phi_\gamma^2(x) \sum_{|\alpha| = 2} |D^\alpha u|^2.$$

Consequently,

$$\|u - v\|_{H^1(T)}^2 \leq C h^2 |u|_{H^2, \gamma(T)}^2.$$

**Case:** element $T$ with a singular vertex $x_i$ and weight $0 \leq \gamma_i < 1$. Assume for simplicity that $x_i = 0$.

For $x \in T$ and $s < 1$,

$$|x| \leq d_T \quad \Rightarrow \quad |x|^{2(s-1)} \geq d_T^{2(s-1)}$$

so, by the interpolation estimate (2.35),

$$d_T^{2(s-1)} \int_T |u - v|^2 + |\nabla (u - v)|^2 \leq C \int_T |x|^{2s} \left| d_T^2 u \right|^2.$$

Setting $s = \gamma_i$, we obtain,

$$\int_T |u - v|^2 + |\nabla (u - v)|^2 \leq C d_T^{2(1-\gamma_i)} \int_T |x|^{2\gamma_i} \left| d_T^2 u \right|^2.$$

But, by the mesh design,

$$d_T \leq L h \sup_{x \in T} \phi_\gamma(x) \leq C h \sup_{x \in T} |x - x_i|^{\gamma_i} \leq C h d_T^{1-\gamma_i}$$

so,

$$d_T^{1-\gamma_i} \leq C h$$

which yields the desired estimate. Summing up the element interpolation error estimates over all elements $T$, we obtain the global estimate.

The mesh parameter $h$ can be estimated by the total number of vertices $N$ (degrees-of-freedom).
LEMMA 2.4.5

There exists a constant $C > 0$, dependent upon $\Omega, \gamma, L$ but independent of $h$ such that, for every triangulation $T$ of type $(h, \gamma, L)$,

$$N \leq Ch^{-2}.$$  \hfill (2.37)

PROOF Clearly,

$$N \leq 3 \# \text{ elements}.$$  

As the number of elements adjacent to singular vertices is finite, it is sufficient to estimate the number of elements that are not adjacent to any of the singular vertices. By the mesh design, we have,

$$L^{-1}h\phi_{\gamma}(x) \leq d_T \Rightarrow L^{-2}d_T^{-2} \leq h^{-2}\phi_{\gamma}^{-2}(x),$$

so,

$$L^{-2}d_T^{-2} \int_T 1 \leq h^{-2} \int_T \phi_{\gamma}^{-2}.$$ 

Shape regularity implies that there exists a constant $C$ such that

$$1 \leq Cd_T^{-2}|T| = Cd_T^{-2} \int_T 1 \leq Ch^{-2} \int_T \phi_{\gamma}^{-2}.$$ 

Consequently,

$$\# \text{ elements} \leq Ch^{-2} \int_{\Omega} \phi_{\gamma}^{-2}.$$ 

\[ \blacksquare \]

Lemma 2.4.5 implies that mesh parameter $h$ in estimate (2.36) can be replaced with $N^{-2}$. Note that estimate (2.37) holds trivially for quasiuniform meshes. Use of meshes graded according to the weight function $\phi_{\gamma}$, restores thus the optimal rate of convergence in terms of the total number of degrees-of-freedom $N$.

Exercises

Exercise 2.4.1 Prove the second case of Lemma 2.4.2 using the hint in the text.

(5 points)

Exercise 2.4.2 Provide the scaling arguments in the end of proof of Lemma 2.4.4 to finish the proof for an arbitrary triangle adjacent to the origin.

(5 points)
**Exercise 2.4.3** Let $u(r, \theta)$ be a continuous function in $\bar{S}$ where $S$ is the first quadrant of the unit circle. Let $ar{u}$ denote the average of function $u$ in $\theta$, i.e.

$$
\bar{u}(r) := \frac{2}{\pi} \int_{\pi/2}^{0} u(r, \theta) \, d\theta, \quad r > 0.
$$

Prove that $\bar{u}$ is continuous in $(0, 1]$ and,

$$
\lim_{r \to 0} \bar{u}(r) = u(0).
$$

(5 points)

---

### 2.5 The DPG Method

**Exercises**

**Exercise 2.5.1** Duality pairing. A bilinear (sesquilinear) form $b(u, v)$ defined on two Banach spaces $U, V$, is called a *(generalized)* duality pairing if it is definite, i.e.

$$
b(u, v) = 0 \forall v \in V \implies u = 0
$$

and,

$$
b(u, v) = 0 \forall u \in U \implies v = 0
$$

and, additionally,

$$
\|u\|_U := \sup_{v \neq 0} \frac{|b(u, v)|}{\|v\|_V} \quad \text{and} \quad \|v\|_V := \sup_{u \neq 0} \frac{|b(u, v)|}{\|u\|_U}.
$$

(i) Show that the standard duality pairing between a Banach space and its dual (with the induced norm) satisfies the axioms.

(ii) Let $b(u, v)$ be definite. Replace the original norm in $U$ with an “energy norm”:

$$
\|u\|_E := \sup_{v \neq 0} \frac{|b(u, v)|}{\|v\|_V}.
$$

Prove that the energy norm is indeed a norm on $U$, and that with the energy norm replacing the original norm on $U$, form $b(u, v)$ becomes a duality pairing.

(iii) Repeat the same argument with respect to $v$.

One arrives at non-trivial examples of duality pairings over boundary of a domain studying integration by parts and $L^2$-adjoints.

(5 points)

**Exercise 2.5.2** (10 points)
Historical Comments
References


