Introduction to Error Analysis of Finite Element Methods

Fred Vermolen
Computational Mathematics Group (CMAT)
Department of Mathematics and Statistics
University of Hasselt, Campus Diepenbeek, Agoralaan, Building D, Belgium

Version 2.0: February 12, 2021

1 Introductory Concepts

Error analysis in finite-element methods requires slightly more mathematical concepts than error analysis in finite-difference and finite-volume methods. In the last-mentioned two methods, one typically uses Taylor polynomials to arrive at local discretisation (truncation) errors for the difference formulas to approximate derivatives. Subsequently one uses stability of the finite difference approach through analysis of the norm of the inverse of the discretisation matrix (by the use of eigenvalues of the discretisation matrix, if this matrix is symmetric), which in particular, amounts to

There is a $K > 0$ such that $||A^{-1}|| \leq K$ as the grid size tends to zero (stability), that is as $h \longrightarrow 0$,

to demonstrate that the finite difference approximation converges according to the order of accuracy of the approximation of the derivatives by the difference formulas. This classical result is also known as Lax Equivalence Theorem: A stable, consistent (local truncation error tends to zero if we send the step size to zero) finite difference scheme converges. This concept was analysed in the course Numerical Methods I, see Vuik et al [1].

In this manuscript, some elementary error bounds for finite-element methods are established. Practical implementation of the finite-element method is discussed in van Kan et al [2]. Since this text only aims at providing the idea how apriori finite-element error analysis is carried out, only error estimates for the Poisson equation are considered. Furthermore, the proof of the upper bound for the interpolation error will only be provided for the one-dimensional case. Since finite-element solutions are sought in function spaces, we will estimate finite-element errors in these spaces. Therefore, we start with introducing some mathematical concepts that we will need in the derivations. First, we start with function spaces.

Definition 1 *We introduce the following concepts:*

1. *A function space is a set of functions between two fixed sets, being the domain and codomain (image).*
2. *A Banach space is a complete, normed space.*
3. *A Hilbert space is a Banach space with an inner product induced norm.*
4. *Let $\Omega \subset \mathbf{R}^d$, the space of square integrable functions (L^2–integrable) is defined by*

$$L^2(\Omega) := \{f : \Omega \to \mathbb{R} \ : \ \int_\Omega f^2 d\Omega \text{ is finite}\}.$$

5. *Associated with this space, the L^2–norm is defined by*

$$||f||_{L^2(\Omega)} := \left[\int_\Omega f^2 d\Omega\right]^{1/2}.$$

It will be necessary to understand the notion of an L^2–space and norm. We define the L^2–inner product between two functions as $(f,g)_{L^2(\Omega)} := \int_\Omega fg d\Omega$, which implies that $||f||_{L^2(\Omega)} = (f,f)^{1/2}_{L^2(\Omega)}$, and hence the function space $L^2(\Omega)$ with its norm is a Hilbert space. We use the following recursive convention for Hilbert spaces with (higher-order) partial derivatives of functions:

$$\begin{aligned} &H^0(\Omega) := L^2(\Omega), \\ &H^p(\Omega) := \left\{ f \in H^{p-1}(\Omega) \; : \; D^p f \in L^2(\Omega) \right\}, \text{ for } p \in \mathbb{N} \setminus \{0\}, \end{aligned} \tag{1}$$

where in $\mathbb{R}^d$, we have the multi-indexed derivative

$$D^p u = \left\{ \frac{\partial^p u}{\partial x_1^{q_1} \dots \partial x_d^{q_d}}, \text{ where } (q_1, \dots, q_d) \in \{0, \dots, p\}^d \; : \; \sum_{i=1}^{d} q_i = p \right\}.$$

Note that the above convention differs slightly from the classical convention of multi-indexed derivatives. The above definition implies that the $H^1(\Omega)$ Hilbert space is defined by

$$H^1(\Omega) := \left\{ f \in L^2(\Omega) \; : \; \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \in L^2(\Omega) \right\},$$

where $\Omega \subset \mathbb{R}^2$. This is an important Hilbert space that will be used extensively in this manuscript.

Further, the Hilbert space $H^p(\Omega)$ is associated with the H^p–norm, which is defined by

$$||u||^2_{H^p(\Omega)} := ||u||^2_{H^{p-1}(\Omega)} + \sum_{\left\{(q_1,\dots,q_d) \in \{0,\dots,p\}^d \; : \; \sum_{i=1}^d q_i = p\right\}} || \frac{\partial^p u}{\partial x_1^{q_1} \dots \partial x_d^{q_d}} ||^2_{L^2(\Omega)}, \text{ with } ||u||_{H^0(\Omega)} := ||u||_{L^2(\Omega)}.$$

In finite-element language, these Hilbert spaces are also commonly referred to as Sobolev spaces, which merely represent a generalization of Hilbert spaces regarding different norms, such as p-Hölder norms. At this stage we will not consider this generalization, and use the words Sobolev spaces and Hilbert spaces interchangeably. From the above formalization, it follows that

$$||u||^2_{H^1(\Omega)} = \int_\Omega u^2 + ||\nabla u||^2 \, d\Omega = \int_\Omega u^2 + \sum_{j \in \{1,\dots,d\}} \left(\frac{\partial u}{\partial x_j} \right)^2 \, d\Omega, \text{ in } \mathbb{R}^d.$$

This is an important norm. We also consider semi-norms like

$$|u|^2_{H^p(\Omega)} := \sum_{\left\{(q_1,\dots,q_d) \in \{0,\dots,p\}^d \; : \; \sum_{i=1}^d q_i = p\right\}} || \frac{\partial^p u}{\partial x_1^{q_1} \dots \partial x_d^{q_d}} ||^2_{L^2(\Omega)}, \text{ hence } |u|^2_{H^1(\Omega)} = \int_\Omega \sum_{j \in \{1,\dots,d\}} \left(\frac{\partial u}{\partial x_j} \right)^2 \, d\Omega.$$

Note that norms are characterized by $||.|| \geq 0$, the Triangle Inequality, $||\alpha u|| = |\alpha| ||u||$, and by $||u|| = 0 \iff u = 0$. A semi norm does not satisfy the last requirement, that is: There may exist $u \neq 0$ such that $|u| = 0$.

We note that in order to understand the derivation of the finite-element error in one dimension, it is not necessary to understand the notion of multi-indexed derivatives.

2 The Principle of the Galerkin Finite-Element Method

The finite-element method is most commonly derived by the use of a weak formulation, where the solution is sought in a Hilbert space. To get the idea, let $\Omega \subset \mathbf{R}^d$ be an open domain bounded by piecewise smooth boundary Γ, in which we solve the following problem

$$\begin{aligned} -\Delta u &= f(\mathbf{x}), \quad &\mathbf{x} \in \Omega, \\ u &= g(\mathbf{x}), \quad &\mathbf{x} \in \Gamma. \end{aligned} \tag{2}$$

Here $\Delta u = \nabla \cdot \nabla u =$ div grad u represents the Laplace operator. If $f \in C(\Omega)$ (f is continuous over Ω) and $g \in C(\Gamma)$, then existence and uniqueness of $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ (u has continuous second-order derivatives in Ω, and $\overline{\Omega}$ denotes the closure of Ω, hence $\overline{\Omega} = \Omega \cup \Gamma$) can be demonstrated. The weak form, also known as the finite-element form, is obtained by multiplication of the above partial differential equation (PDE) by a *test function* $\phi(\mathbf{x})$ and integration over the domain of computation Ω, which results into

$$-\int_\Omega \phi \Delta u \, d\Omega = \int_\Omega \phi f \, d\Omega. \tag{3}$$

The above integrals exist if $\phi,\ \Delta u,\ f \in L^2(\Omega)$, or more formally if $u \in H^2(\Omega)$. These continuity requirements for the solution, in fact, are already somewhat weaker than for the original PDE. We will weaken the smoothness requirements further by applying the Product Rule for differentiation on equation (3), to arrive at

$$-\int_\Omega \nabla \cdot (\phi \nabla u) - \nabla \phi \cdot \nabla u \, d\Omega = \int_\Omega \phi f \, d\Omega. \tag{4}$$

The first term in the left-hand side is treated by Gauß' Theorem to arrive at

$$-\int_\Gamma \mathbf{n} \cdot (\phi \nabla u) \, d\Gamma + \int_\Omega \nabla \phi \cdot \nabla u \, d\Omega = \int_\Omega \phi f \, d\Omega. \tag{5}$$

Since we have a Dirichlet boundary condition on Γ, we *choose* $\phi = 0$ on Γ to get rid of the boundary term, to arrive at the following weak formulation:

$$\begin{aligned} &\text{Find } u \in H^1_g(\Omega), \text{ such that for all } \phi \in H^1_0(\Omega), \text{ we have} \int_\Omega \nabla \phi \cdot \nabla u \, d\Omega = \int_\Omega \phi f \, d\Omega, \\ &\text{where } H^1_g(\Omega) := \{u \in H^1(\Omega) \mid u = g(\mathbf{x}) \text{ on } \Gamma\}, \text{ and } H^1_0(\Omega) := \{u \in H^1(\Omega) \mid u = 0 \text{ on } \Gamma\}. \end{aligned} \tag{6}$$

Now the solution is sought in (a subset of) the Hilbert space $H^1(\Omega)$, which requires derivates of u to be square-integrable, rather than requiring C^2-continuity globally. Therefore, equation (6) is referred to as a *weak formulation* of boundary value problem (2). Since we are searching finite-element solution in H^1, which is a Hilbert space, we will estimate the finite-element error in Hilbert spaces such as the L^2–norm. Formally, the left-hand side in the above integral, represents a *bilinear form*, defined by

$$a(u,\phi) := \int_\Omega \nabla u \cdot \nabla \phi \, d\Omega, \tag{7}$$

it is easy to see that $a(.,.)$ is linear in both arguments. Further, we define

$$(\phi, f) := \int_\Omega \phi f \, d\Omega. \tag{8}$$

Then the weak form amounts to:

$$\text{Find } u \in H^1_g(\Omega), \text{ such that for all } \phi \in H^1_0(\Omega), \text{ we have } a(u,\phi) = (\phi, f). \tag{9}$$

The finite-element approximation is constructed by dividing the domain Ω into a mesh, with grid points $\{\mathbf{x}_j\}$, and elements $\{e_k\}$, which are all engaged to its specific set of grid points. In the Lagrangian finite-element frameworks, the basis functions are piecewise smooth or continuous Lagrangian interpolatory polynomials defined over each element, such that

$$\phi_i(\mathbf{x}_j) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

For the sake of simplicity, we set $g(\mathbf{x}) = 0$ on Γ and hence $\phi_i = 0$ on Γ. The general case is dealt with in a straightforward way using a homogenization argument. Hence the finite-element solution is sought in $V_h(\Omega) := \text{Span}\{\phi_i\}_{i=1}^n \subset V(\Omega) := H^1_0(\Omega)$, where we write the finite-element approximation by

$$u(\mathbf{x}) \approx u_h(\mathbf{x}) = \sum_{j=1}^n c_j \phi_j(\mathbf{x}),$$

and for the test function, we choose $\phi_h \in V_h(\Omega)$. In other words, the discrete weak form becomes:

$$\text{Find } u_h \in V_h(\Omega)\text{, such that for all } \phi_h \in V_h(\Omega)\text{, we have } a(u_h, \phi_h) = (\phi_h, f). \tag{10}$$

Since $V_h(\Omega) \subset V(\Omega)$, we can naturally 'test' the continuous weak form by all functions in $V_h(\Omega)$:

$$\text{Find } u \in V(\Omega)\text{, such that for all } \phi_h \in V_h(\Omega)\text{, we have } a(u, \phi_h) = (\phi_h, f). \tag{11}$$

Subtraction of the last two forms, yields the following orthogonality relation for the *finite-element error* $u - u_h$:

$$\boxed{a(u_h - u, \phi_h) = 0, \text{ for all } \phi_h \in V_h(\Omega). \qquad (12)}$$

This relation states that the difference between the finite-element approximation, u_h and the (exact) solution, u, is in a sense orthogonal to the test space $V_h(\Omega)$. This orthogonality relation is crucially important in the further estimates that we will make, and this result is valid for general bilinear forms. Suppose that a given bilinear form satisfies the following requirements in $V(\Omega)$:

1. $a(.,,)$ is continuous (or bounded) in $V(\Omega)$, that is:

   There is a $K > 0$ such that $|a(u,\phi)| \leq K||u||_{V(\Omega)}||\phi||_{V(\Omega)}$ for all u, $\phi \in V(\Omega)$;

2. $a(.,.)$ is coercive (or strongly elliptic) in $V(\Omega)$, that is:

   There is a $c > 0$ such that $a(u,u) \geq c||u||^2_{V(\Omega)}$, for all $u \in V(\Omega)$.

Going back to our bilinear form, then one can demonstrate Poincaré's (Lax-Friedrichs) Inequality, which says that there is a $\beta > 0$ such that

$$\int_\Omega ||\nabla u||^2 \, d\Omega = ||\nabla u||^2_{L^2(\Omega)} \geq \beta ||u||^2_{L^2(\Omega)} = \int_\Omega u^2 \, d\Omega.$$

This implies that

$$||u||^2_{H^1(\Omega)} = \int_\Omega u^2 + ||\nabla u||^2 \, d\Omega \leq \left(\frac{1}{\beta} + 1\right) \int_\Omega ||\nabla u||^2 \, d\Omega = \left(\frac{1}{\beta} + 1\right) a(u,u)$$

$$\Longleftrightarrow \qquad a(u,u) \geq \frac{\beta}{1+\beta} ||u||^2_{H^1(\Omega)} = c||u||^2_{H^1(\Omega)}, \text{ where } c := \frac{\beta}{1+\beta}.$$

Since $\beta > 0$, we have $c > 0$ since and remember that we had $a(u,u) = \int_\Omega ||\nabla u||^2 \, d\Omega$. The inequality by Cauchy-Schwartz immediately gives boundedness (or continuity) of this symmetric (that is $a(u,\phi) = a(\phi,u)$) bilinear form $a(u,\phi) = \int_\Omega \nabla u \cdot \nabla \phi \, d\Omega$. It can be proved that all symmetric bilinear forms are bounded.

Hence, the two properties are satisfied by the current choice of bilinear form on the function space $H^1_0(\Omega)$. For the sake of finite-element error, one often considers the *energy norm* of the solution. The energy norm is defined by

Definition 2 *Given $u \in H^1(\Omega)$, then the* energy norm *is defined by*

$$||u||_{E(\Omega)} := \left[\int_\Omega ||\nabla u||^2 d\Omega\right]^{1/2}.$$

Note that if $u = 0$ on boundary Γ, then the energy norm defines a proper norm. Next, we consider the energy-norm of the error for the chosen problem, and choose any $\tilde{u}_h \in V_h(\Omega)$, then for the energy norm of the finite-element error, we get

$$\begin{aligned} &||u - u_h||^2_{E(\Omega)} = \int_\Omega ||\nabla(u - u_h)||^2 \, d\Omega = a(u - u_h, u - u_h) = a(u - u_h, u - \tilde{u}_h + \tilde{u}_h - u_h) = \\ &a(u - u_h, u - \tilde{u}_h) + a(u - u_h, \tilde{u}_h - u_h) = a(u - u_h, u - \tilde{u}_h) \leq ||u - u_h||_{E(\Omega)} \cdot ||u - \tilde{u}_h||_{E(\Omega)}. \end{aligned} \tag{13}$$

Note that the boundary condition $u = 0$ on Γ implies that the energy norm is a proper norm. The above relation implies that

$$||u - u_h||_{E(\Omega)} \leq ||u - \tilde{u}_h||_{E(\Omega)}, \text{ for all } \tilde{u}_h \in V_h(\Omega). \tag{14}$$

This can be formalized in Céa's Lemma:

Lemma 1 *Céa's Lemma: Let $u \in H_0^1(\Omega)$ satisfy*

$$\int_\Omega \nabla u \cdot \nabla \phi d\Omega = \int_\Omega \phi f d\Omega, \textit{ for all } \phi \in H_0^1(\Omega),$$

let $V_h(\Omega)$ be a finite dimensional subset of $H_0^1(\Omega)$, and let $u_h \in V_h(\Omega)$ satisfy

$$\int_\Omega \nabla u_h \cdot \nabla \phi_h d\Omega = \int_\Omega \phi_h f d\Omega, \textit{ for all } \phi_h \in V_h(\Omega),$$

and let $||u||_{E(\Omega)}^2 := \int_\Omega ||\nabla u||^2 d\Omega$, then

$$||u - u_h||_{E(\Omega)} \leq ||u - \tilde{u}_h||_{E(\Omega)}, \textit{ for all } \tilde{u}_h \in V_h(\Omega).$$

This lemma implies that the energy norm of the finite element solution u_h is estimated from above by the energy norm of the interpolatory approximation of the exact solution u. Hence, in order to prove convergence of the finite element solution, it is sufficient to prove convergence of the interpolatory approximation of the exact solution. Further, the conditions in this lemma (as well as continuity of the right-hand side, which is satisfied if $f \in L^2(\Omega)$, because then one gets $|\int_\Omega \phi f d\Omega| \leq ||\phi||_{L^2(\Omega)}||f||_{L^2(\Omega)} \leq \frac{1}{\sqrt{\alpha}}||\phi||_{H^1(\Omega)}||f||_{L^2(\Omega)}$) are also satisfied in Lax-Milgram's Lemma (or the Riesz Representation Theorem) for existence and uniqueness of solutions to variational problems.

3 Preliminaries from Functional Analysis: Riesz' Representation Lemma and Lax Milgram's Lemma

This section may be skipped if one is only interested in how finite-element errors are estimated. However for readers with more mathematical interest, this section will shed some light on the existence of solutions to weak formulations in finite-element spaces. Further, we present some generic existence and uniqueness results that may be used for the analysis of linear boundary value problems. Before we state the Riesz' Representation Theorem, which already is an important preliminary and fundamental tool for proving existence and uniqueness to certain boundary value problems, we introduce the definition of a bounded linear functional:

Definition 3 *Let H be a Hilbert space. A linear functional $f : H \longrightarrow \mathbb{R}$ is bounded in H if there exists a $K > 0$ such that*

$$|f(v)| \leq K||v||_H, \qquad \forall v \in H.$$

The set of linear, bounded functionals on H is denoted by H^{-1}, and $||f||_{H^{-1}} := \sup_{0 \neq v \in H} \frac{|f(v)|}{||v||_H}$. This norm is often referred to as a dual norm or a negative norm. Further H^{-1} is often referred to as the dual space of H.

Here $||v||_H$ denotes the H–norm of v. A simple example could be $H = L^2(\Omega)$, then we would be referring to the $L^2(\Omega)$–norm. We denote the inner product on Hilbert space H by $(u, v)_H$ for $u, v \in H$. As examples, we consider

$$H = L^2(\Omega), \text{ where } (u,v)_{L^2(\Omega)} = \int_\Omega uv \; d\Omega,$$

or for

$$H = H^1(\Omega), \text{ we have } (u,v)_{H^1(\Omega)} = \int_\Omega uv \; + \; \nabla u \cdot \nabla v \; d\Omega.$$

Next we introduce the fundamental Riesz' Representation Theorem:

Theorem 1 *Let H be a Hilbert space, and let $f \in H^{-1}$ (i.e. a bounded, linear functional on H), then there exists exactly one $u \in H$ such that*

$$(u,v)_H = f(v), \textit{ for all } v \in H.$$

Further, we have $||f||_{H^{-1}} = ||u||_H$.

Although one can find the proof (in multiple ways) of this assertion in standard textbooks in Functional Analysis, the proof is given for the sake of completeness.

Proof: First we prove that there exists at most one $u \in H$ that satisfies the variational form $(u,v)_H = f(v)$. Suppose that multiple u, say u_1 and u_2 satisfy the variational form, then we obtain $(u_1 - u_2, v)_H = 0$, for all $v \in H$. We choose $v = u_1 - u_2$, then we obtain $(u_1 - u_2, u_1 - u_2)_H = ||u_1 - u_2||_H^2 = 0$. Hence, since this is proper norm, we arrive at $u_1 - u_2 = 0$. Hence there is at most one $u \in H$ for which $(u,v)_H = f(v)$ for all $v \in H$.

Next, we consider existence. Consider the functional

$$J(v) = \frac{1}{2}||v||_H^2 - f(v),$$

then it is not hard to prove that this functional is convex and that it has a (unique (which makes the first uniqueness part unnecessary in the proof)) minimiser. The minimiser of this functional over H, which hence exists, is denoted by u and is determined by setting the Gateaux differential to zero,

$$\frac{d}{dt} J(u + tv)_{t=0} = 0,$$

From this relation, it follows that the minimiser of J over H satisfies

$$u \in H \text{ such that } (u,v)_H = f(v), \quad \forall v \in H.$$

This is exactly the variational problem. Hence existence of a solution to the minimisation problem implies the existence of a solution to the variational problem. Hence existence and uniqueness have been demonstrated.

Next we demonstrate that $||f||_{H^{-1}} = ||u||_H$. From the definition of the negative norm, it follows that

$$||f||_{H^{-1}} = \sup_{0 \neq v \in H} \frac{|f(v)|}{||v||_H} \geq \frac{|f(u)|}{||u||_H} = \frac{(u,u)_H}{||u||_H} = ||u||_H.$$

Hence we have

$$||f||_{H^{-1}} \geq ||u||_H. \tag{15}$$

Combining Cauchy-Schwartz' Inequality with the existence result, we arrive at

$$|f(v)| = |(u,v)_H| \leq ||u||_H \; ||v||_H, \qquad u, v \in H.$$

From this inequality, and choosing $v \neq 0$, we obtain

$$\frac{|f(v)|}{||v||_H} \leq ||u||_H, \text{ for all } u, v \in H.$$

Since this inequality holds for all $0 \neq v \in H$, it follows that it should also hold for its supremum over H, hence we arrive at

$$||f||_{H^{-1}} = \sup_{0 \neq v \in H} \frac{|f(v)|}{||v||_H} \leq ||u||_H. \tag{16}$$

Combining inequalities (15) with (16), concludes that $||f||_{H^{-1}} = ||u||_H$. □

This pivotal result is used to demonstate the generalisation Lax-Milgram's Lemma for existence and uniqueness of non-symmetric problems. Before we state the Lemma, we first introduce the concepts of continuity and coerciveness of bilinear forms:

1. $a(.,,)$ is continuous (or bounded) in H, that is:

   There is a $K > 0$ such that $|a(u,v)| \leq K||u||_H \; ||v||_H$ for all u, $v \in H$;

2. $a(.,.)$ is coercive (or strongly elliptic) in H, that is:

   There is a $c > 0$ such that $a(u,u) \geq c||u||_H^2$, for all $u \in H$.

These concepts are used in Lax-Milgram's Lemma:

Lemma 2 *Let H be a Hilbert space, let $a(.,.) : H \times H \longrightarrow \mathbb{R}$ be a continuous (bounded), coercive bilinear form, and let $f(v)$ be a bounded linear functional in H (that is $f \in H^{-1}$), then there exists exactly one $u \in H$ such that*

$$a(u,v) = f(v), \qquad \forall v \in H.$$

Furthermore $||u||_H \leq \frac{1}{c}||f||_{H^{-1}}$, where $a(v,v) \geq c||v||_H^2$, $\forall v \in H$.

Proof: We use a combination of Banach's Contraction Theorem and Riesz' Representation Theorem. We consider the following variational problem derived from $a(u,v) = f(v)$:

$$\text{Find } u \in H \text{ such that } (u,v)_H = (u,v)_H - s(a(u,v) - f(v)), \qquad \forall v \in H, \qquad s \in \mathbb{R}.$$

We will use Banach's Contraction Theorem to demonstrate existence and uniqueness of $u \in H$. This form can be written as $(u,v)_H = (P_s(u),v)_H$, $\forall v \in H$. This step is motivated by the fact that $f(v)$ is bounded in H, which allows us to write $f(v) = (u_0, v)$ for a $u_0 \in H$, for all $v \in H$. It also follows that $P_s : H \longrightarrow H$. We will show that $P_s(u)$ is a contractive mapping for some $s \in \mathbb{R}$. Let

$$(w_1,v)_H = (P_s(u_1),v)_H = (u_1,v)_H - s(a(u_1,v) - f(v)),$$

$$(w_2,v)_H = (P_s(u_2),v)_H = (u_2,v)_H - s(a(u_2,v) - f(v)).$$

We will demonstrate that $||w_1 - w_2||_H < ||u_1 - u_2||_H$ for some $s \in \mathbb{R}$, which makes $P_s(u)$ a contractive mapping. From the above equations, it follows that

$$(w_1 - w_2, v)_H = (u_1 - u_2, v)_H - sa(u_1 - u_2, v) = (u_1 - u_2, v)_H - s\langle A(u_1 - u_2), v\rangle.$$

In the last step, we used the fact that $a(u,v)$ can be seen as a duality pairing of linear bounded operator $Au \in H^{-1}$ with $v \in H$, which makes $\langle Au, v\rangle$ a linear, bounded functional in v, hence on H. Due to Reisz' Representation Theorem, there is a unique $JAu \in H$ such that $\langle Au, v\rangle = (JAu, v)_H$ for all $v \in H$. The operator $J : H^{-1} \longrightarrow H$ must be an isometric isomorphism (linear bijection and $||Jf||_H = ||f||_{H^{-1}}$). This implies that the above formula gives

$$(w_1 - w_2, v)_H = (u_1 - u_2, v)_H - s(JA(u_1 - u_2), v).$$

This implies that $w_1 - w_2 = (I - sJA)(u_1 - u_2)$. This gives

$$||w_1 - w_2||_H^2 = (w_1 - w_2, w_1 - w_2)_H = ||u_1 - u_2||_H^2 - 2s(JA(u_1 - u_2), u_1 - u_2) + s^2||JA(u_1 - u_2)||_H^2 \leq$$

$$\leq (1 - 2sc + s^2K^2) \cdot ||u_1 - u_2||_H^2.$$

Here we used coerciveness and boundedness of the bilinear form. From the quadratic form, it can be seen that for $s \in (0, \frac{2c}{K^2})$, we have $||w_1 - w_2||_H < ||u_1 - u_2||_H$, for which a contraction is obtained on H. Hence

$$\exists! u \in H \ : \ a(u,v) = f(v), \ \forall v \in H.$$

Next, we consider the second part of the Lemma:

$$c||u||_H^2 \leq a(u,u) = f(u) = |f(u)| = \frac{|f(u)|}{||u||_H}||u||_H \leq \sup_{0 \neq u \in H} \frac{|f(u)|}{||u||_H}||u||_H = ||f||_{H^{-1}}||u||_H.$$

Division by $||u||_H$ and $c > 0$, gives

$$||u||_H \leq \frac{1}{c}||f||_{H^{-1}}.$$

This concludes the proof of the lemma. □

Next, we apply the Lax-Milgram Lemma to an n dimensional space, and let $H^n = \text{Span}\{v_1, \dots, v_n\}$, where $v_j \in H$, hence $H^n \subset H$ is an n dimensional subspace of H. Then one obtains the following claim

Lemma 3 *Let $a(.,.) : H \times H \longrightarrow \mathbb{R}$ be coercive and continuous on H, and let $f(v)$ be bounded in H. Suppose $H^n = Span\{v_1, \dots, v_n\}$, where $v_j \in H$, then*

There is exactly one $u^n \in H^n$ such that $a(u^n, v_i) = f(v_i)$, for all $v_i \in H^n$.

Furthermore $||u^n||_H \leq \frac{1}{c}||f||_{H^{-1}}$.

Proof: Since H^n is a subspace of a Hilbert space, which is a Hilbert space itself, it follows that $a(.,.)$ is bounded and coercive on H^n. Further, $f(v)$ is also bounded on H^n. This proves the existence and uniqueness part of the theorem. Further, we have

$$c||u^n||_H^2 \leq a(u^n,u^n) = f(u^n) = |f(u^n)| = \frac{|f(u^n)|}{||u^n||_H}||u^n||_H \leq \sup_{0 \neq u^n \in H^n} \frac{|f(u^n)|}{||u^n||_H}||u^n||_H \leq \sup_{0 \neq u \in H} \frac{|f(u)|}{||u||_H}||u^n||_H = ||f||_{H^{-1}}||u^n||_H.$$

Division by $||u^n||_H$ and $c > 0$, gives

$$||u^n||_H \leq \frac{1}{c}||f||_{H^{-1}}.$$

This concludes the proof of the lemma. □

This lemma proves the existence of a finite-element solution. This also implies that the system matrix (stiffness matrix), defined by $S_{ij} = a(v_i, v_j)$ is non-singular and the solution to $S\underline{x} = \underline{b}$ is uniquely defined.

4 Convergence of the Galerkin Approximation to the Exact Solution

This section is not necessary for the understanding of the derivation of elementary error estimation of the finite-element method. This section is meant for the reader with a larger mathematical interest and opens the road to more general error analysis for Galerkin-based finite elements, where the function spaces for the test functions (i.e. the test space) and the function space of the solution (i.e. the solution space) are the same. We consider the convergence of the Galerkin approximation to the solution of the weak form as the number of basis functions, that is, the dimensionality, is sent to infinity. First, we generalise Céa's Lemma from the previous section.

Theorem 2 *Let H be a Hilbert space, and let $a(.,.)$ be a bounded, coercive (with respective constants $K > 0$ and $c > 0$) bilinear form on H. Suppose that $H^n = Span\{v_1, \dots, v_n\}$, where $v_j \in H$, and let*

$$\begin{aligned} &u \in H \textit{ such that } a(u,v) = f(v), \textit{ for all } v \in H, \\ &u^n \in H^n \textit{ such that } a(u^n, v_j) = f(v_j), \textit{ for all } v_j \in H^n. \end{aligned} \tag{17}$$

Then $||u - u_n||_H \leq \frac{K}{c}||u - v_n||_H$ for all $v^n \in H^n$.

This is a generalisation of Lemma 1, which we will prove using similar principles.

Proof: See Assignment 3. □

The above theorem is known as a generalisation of Lemma 1, and we will use this result to conclude convergence of the Galerkin method. Before we state the result, we review the definition of *density*.

Definition 4 *Let H be a Hilbert space, and let H^n be a n dimensional subset, for which $H^n = Span\{v_1, \dots, v_n\}$, with $v_j \in H$ as basis functions. Then H^n is* dense *in H if for all $v \in H$ we have that either $v \in H^n$ or v is a limit point of H^n, that is*

$\forall v \in H, \forall \varepsilon > 0 :$ $\exists N > 0$ *such that there is a $v^n \in H^n$ for which $||v^n - v||_H < \varepsilon$ (i.e. $dist(H^n, v) < \varepsilon$),$\forall n > N$.*

The above definition says that we can approximate each $v \in H$ arbitrarily well in H^n if we take enough basis functions. This also means that the solution to the weak form, $u \in H$ can be approximated arbitrarily well in H^n. Then the generalisation of Céa's Lemma will warrant convergence. We formulate this in the following convergence theorem:

Theorem 3 *Let H be a Hilbert space, and let $a(.,.)$ be a bounded, coercive (with respective constants $K > 0$ and $c > 0$) bilinear form on H. Suppose that $H^n = Span\{v_1, \dots, v_n\}$, where $v_j \in H$, and let H^n be dense in H, and let*

$$\begin{aligned} &u \in H \textit{ such that } a(u,v) = f(v), \textit{ for all } v \in H, \\ &u^n \in H^n \textit{ such that } a(u^n, v_j) = f(v_j), \textit{ for all } v_j \in H^n. \end{aligned} \tag{18}$$

Then $||u - u_n||_H \longrightarrow 0$ as $n \longrightarrow \infty$. In other words, the Galerkin method converges.

Proof: Since H^n is dense in H, and since $u \in H$, the definition of density implies that

$$\forall \varepsilon > 0 : \quad \exists N > 0 \text{ such that there is a } \hat{u}^n \in H^n \text{ for which } 0 \leq ||\hat{u}^n - u||_H < \varepsilon,\ \forall n > N.$$

Combining this relation with Céa's Lemma, gives

$$\forall \varepsilon > 0 : \quad \exists N > 0 \text{ such that } n > N \Longrightarrow 0 \leq ||u^n - u||_H \leq \frac{K}{c}||\hat{u}^n - u||_H < \frac{K}{c}\varepsilon.$$

This implies convergence of the Galerkin method. □

5 Convergence of the Interpolatory Approximation in the Energy Norm

In this section, we will prove convergence in the energy norm of the one-dimensional interpolatory approximation of the exact solution. To this extent, we first introduce and prove Taylor's Theorem with the integral representation of the error:

Theorem 4 *Let $f \in H^p(I)$ where $I = (\alpha, \beta)$ $\alpha < \beta$, let $a \in I$, then for all $x \in I$ we have*

$$f(x) = f(a) + (x-a)f'(a) + \ldots + \frac{(x-a)^p}{p!} f^{(p)}(a) + \frac{1}{p!}\int_a^x (x-s)^p f^{(p+1)}(s)ds. \tag{19}$$

Proof: We proceed by Mathematical Induction. For $p = 0$, the above equation (19) gives

$$f(x) = f(a) + \int_a^x f'(s)ds = f(a) + f(x) - f(a) = f(x),$$

hence the equation works for $p = 0$. Next we use equation (19), which we have to use to prove the following Induction Hypothesis:

$$f(x) = f(a) + (x-a)f'(a) + \ldots + \frac{(x-a)^{p+1}}{(p+1)!} f^{(p+1)}(a) + \frac{1}{(p+1)!}\int_a^x (x-s)^{p+1} f^{(p+2)}(s)ds. \tag{20}$$

Taking the integral error term in the Induction Hypothesis (20), gives using Integration by Parts

$$\frac{1}{(p+1)!}\int_a^x (x-s)^{p+1} f^{(p+2)}(s)ds = \frac{1}{(p+1)!}\Big[(x-s)^{p+1} f^{(p+1)}(s)\Big]_a^x + \int_x^a \frac{1}{p!}(x-s)^p f^{(p)}(s)ds =$$

$$-\frac{1}{(p+1)!}(x-a)^{p+1} f^{(p+1)}(a) + \int_x^a \frac{1}{p!}(x-s)^p f^{(p)}(s)ds.$$

Using equation (19) for the integral term in the above equation, shows, after some rearrangement, that the Induction Hypothesis (20) is satisfied. This proves the assertion. □

The above theorem is pivotal in expressing the interpolatory error in terms of the L_2–norm of the second derivative of the function that is to be interpolated. Consider $x_{k-1} < x < x_k$, $x_{k-1} < x_k$, then the linear interpolant of $f(x)$ on $[x_{k-1}, x_k]$ is given by

$$f_h(s) = f(x_{k-1})\frac{s - x_k}{x_{k-1} - x_k} + f(x_k)\frac{s - x_{k-1}}{x_k - x_{k-1}}, \qquad s \in [x_{k-1}, x_k],$$

using Taylor's Theorem for $p = 1$ (linearization around x), gives

$$\begin{aligned} f_h(s) = \quad & \frac{s - x_k}{x_{k-1} - x_k}\Big(f(x) + (x_{k-1} - x)f'(x) + \int_x^{x_{k-1}} (x_{k-1} - t)f''(t)dt\Big) + \\ & \frac{s - x_{k-1}}{x_k - x_{k-1}}\Big(f(x) + (x_k - x)f'(x) + \int_x^{x_k} (x_k - t)f''(t)dt\Big). \end{aligned} \tag{21}$$

Chosing $s = x$, gives, see Assignment,

$$f_h(x) = f(x) + \frac{x - x_k}{x_{k-1} - x_k}\int_x^{x_{k-1}} (x_{k-1} - t)f''(t)dt + \frac{x - x_{k-1}}{x_k - x_{k-1}}\int_x^{x_k} (x_k - t)f''(t)dt. \tag{22}$$

Assignment 1 *Show that upon setting $s = x$, equation (21) can be written as equation (22).*

This relation can be used to demonstrate that the error becomes of second order, however, due to Céa's Lemma, we are only interested in the interpolatory error of the L^2–norm of the derivative. Differentiation of equation (21) with respect to s gives

$$f_h'(s) = f'(x) + \int_x^{x_{k-1}} \frac{x_{k-1} - t}{x_{k-1} - x_k} f''(t)dt + \int_x^{x_k} \frac{x_k - t}{x_k - x_{k-1}} f''(t)dt. \tag{23}$$

Assignment 2 *Show that differentiation of equation (21) with respect to s implies equation (23).*

Since $|\frac{x_{k-1}-t}{x_{k-1}-x_k}| \leq 1$ and $|\frac{x_k-t}{x_k-x_{k-1}}| \leq 1$ for $t \in [x_{k-1}, x_k]$, the above equation implies

$$0 \leq |f_h'(s) - f'(x)| \leq \int_{x_{k-1}}^{x_k} |f''(t)|dt \leq \sqrt{(x_k - x_{k-1})} \left[\int_{x_{k-1}}^{x_k} (f''(t))^2 dt \right]^{1/2}.$$

The latest inequality results from application of Cauchy-Schwartz' Inequality. Hence, integration over the interval (x_{k-1}, x_k) gives

$$0 \leq \int_{x_{k-1}}^{x_k} |f_h'(s) - f'(x)|^2 dx \leq (x_k - x_{k-1})^2 \int_{x_{k-1}}^{x_k} (f''(t))^2 dt.$$

We summarize the result in the following theorem:

Theorem 5 *Given $x_{k-1} < x_k$, and let $f \in H^2(x_{k-1}, x_k)$, and let $x \in [x_{k-1}, x_k]$ and*

$$f_h(x) = f(x_{k-1}) \frac{x - x_k}{x_{k-1} - x_k} + f(x_k) \frac{x - x_{k-1}}{x_k - x_{k-1}},$$

then we have

$$0 \leq \int_{x_{k-1}}^{x_k} |f_h'(s) - f'(x)|^2 dx \leq (x_k - x_{k-1})^2 \int_{x_{k-1}}^{x_k} (f''(t))^2 dt.$$

Suppose we are dealing with an interval (a, b) that is divided into subintervals $a = x_0 < x_1 < \ldots < x_{k-1} < x_k < \ldots < x_n = b$. Then, assuming that $f \in H^2(a, b)$, Theorem 5 implies that

$$0 \leq \int_a^b |f_h'(s) - f'(x)|^2 dx = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} |f_h'(s) - f'(x)|^2 dx \leq$$

$$\sum_{k=1}^{n} (x_k - x_{k-1})^2 \int_{x_{k-1}}^{x_k} (f''(t))^2 dt \leq h^2 \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} (f''(t))^2 dt = h^2 \int_a^b (f''(t))^2 dt,$$

where $h := \max_{k \in \{1,\ldots,n\}} (x_k - x_{k-1})$. This implies that (note that the square root should be taken for the norm)

$$0 \leq ||f_h' - f'||_{L^2(a,b)} \leq h||f''||_{L^2(a,b)}.$$

We formalize this result in the following theorem:

Theorem 6 *Let $f \in H^2(a,b)$, given $a = x_0 < x_1 < \ldots < x_{k-1} < x_k < \ldots < x_n = b$, and let f_h be the piecewise linear interpolation of f, such that*

$$f_h(x) = f(x_{k-1})\frac{x - x_k}{x_{k-1} - x_k} + f(x_k)\frac{x - x_{k-1}}{x_k - x_{k-1}}, \textit{ for } x_{k-1} \leq x \leq x_k,$$

and let $h := \max_{k \in \{1,\ldots,n\}} (x_k - x_{k-1})$, then we have

$$0 \leq ||f'_h - f'||_{L^2(a,b)} \leq h||f''||_{L^2(a,b)}.$$

This result can be similarly extended in a straightforward way to higher-order (p–th order) interpolation and to higher dimensionality. In general, one can write

Theorem 7 *Let $p \in \{1,2,\ldots\}$ and $f \in H^{p+1}(\Omega)$, and let f_h be the Lagrangian interpolation of f over a discrete representation of the bounded region $\Omega \subset \mathbb{R}^d$ with nodal points $\{\mathbf{x}_k\}$, such that f_h represents a p–th order interpolation of f using n_e elements and let h be the maximum diameter of all elements in Ω, then we have*

$$0 \leq ||\nabla(f_h - f)||_{L^2(\Omega)} \leq h^p |f|_{H^{p+1}(\Omega)}.$$

Proof: Proof will be given in future version. □

Theorems 6 and 7 are special cases of the *Bramble–Hilbert Lemma*, and state that the energy norm of the interpolant is bounded and converges to zero linearly as $h \longrightarrow 0$ for linear interpolation (Theorem 6) and with order p for p–order interpolation (Theorem 7). We saw from Céa's Lemma that the energy norm of the finite-element error is bounded from above by the energynorm of the Interpolatory error. Hence, the finite-element solution converges in the energy norm, this is summarised in the following theorem:

Theorem 8 *Let $p \in \{1,2,\ldots\}$ and $f \in H^{p+1}(\Omega)$ and let $\Omega \subset \mathbb{R}^d$ be bounded by (piecewise) smooth boundary Γ, and suppose that*

$$u \in H^1_0(\Omega), \textit{ such that for all } \phi \in H^1_0(\Omega), \textit{ we have} \int_\Omega \nabla\phi \cdot \nabla u \; d\Omega = \int_\Omega \phi f \; d\Omega, \tag{24}$$

and let u_h be the finite element solution with piecewise p–th order interpolation Lagrangian basis functions and let h be the maximum diameter of all elements in Ω, then we have

$$0 \leq ||\nabla(u_h - u)||_{L^2(\Omega)} \leq h^p |f|_{H^{p+1}(\Omega)}.$$

Note that with the given boundary condition, this implies with Poincaré's Inequality that $||u - u_h||_{L^2(\Omega)} \longrightarrow 0$ as $h \longrightarrow 0$, and in principle, Poincaré's Inequality suggests that the finite-element method converges in the L^2–norm with at least an order p if p–th order basis functions are used. In the next section, we will analyse the actual convergence behaviour in the L^2–norm and we will see that convergence will be of order $p+1$.

6 The L^2–Norm of the Error

In the previous section, we demonstrated convergence in the energy norm of the finite element error. Next, we demonstrate convergence of the error in the L^2–norm. The analysis is done by the use of *Nitsche's trick*. We consider finite element solution of the problem

$$\begin{aligned} -\Delta u = f, &\quad \text{in } \Omega,\\ u = 0, &\quad \text{on } \Gamma. \end{aligned}$$

The idea of Nitsche's trick is to consider w, which solves

$$\begin{aligned} -\Delta w = u - u_h, &\quad \text{in } \Omega,\\ w = 0, &\quad \text{on } \Gamma. \end{aligned}$$

Note that the right-hand side of the above partial differential equation represents the finite-element error. The weak form for w is given by

$$\text{Find } w \in H_0^1(\Omega) \text{ such that} \int_\Omega \nabla w \cdot \nabla \phi d\Omega = \int_\Omega \phi (u - u_h) d\Omega, \text{ for all } \phi \in H_0^1(\Omega).$$

Let $w_h \in V_h(\Omega)$ where $V_h(\Omega) \subset H_0^1(\Omega)$ is a finite dimensional subset, then

$$\text{Find } w_h \in V_h(\Omega) \text{ such that} \int_\Omega \nabla w_h \cdot \nabla \phi_h d\Omega = \int_\Omega \phi_h (u - u_h) d\Omega, \text{ for all } \phi_h \in V_h(\Omega).$$

Choose $\phi_h = u - u_h$ in the weak form for w, then one obtains

$$0 \leq \int_\Omega (u-u_h)^2 d\Omega = \int_\Omega \nabla w \cdot \nabla (u-u_h) d\Omega = \int_\Omega \nabla (w - w_h) \cdot \nabla (u-u_h) d\Omega. \tag{25}$$

The last step follows because $w_h \in V_h(\Omega)$ and because of orthogonality relation equation (12). For the sake of illustration, we consider linear elements. Higher order elements can be dealt with analogously. For linear elements, we have the following errors (see Theorem 5 with $p = 1$) for the energy norms of w and u, respectively:

There is a $C > 0$ such that

$$\begin{aligned} &||w - w_h||_{E(\Omega)} = ||\nabla (w - w_h)||_{L^2(\Omega)} \leq Ch||u - u_h||_{L^2(\Omega)}, \text{ and}\\ &||u - u_h||_{E(\Omega)} = ||\nabla (u - u_h)||_{L^2(\Omega)} \leq Ch||f||_{L^2(\Omega)}. \end{aligned} \tag{26}$$

Using Cauchy-Schwartz on equation (25) and subsequent combination with inequality (26), gives

$$\begin{aligned} &0 \leq \int_\Omega (u-u_h)^2 d\Omega = ||u-u_h||^2_{L^2(\Omega)} = \int_\Omega \nabla (w-w_h) \cdot \nabla (u-u_h) d\Omega \leq\\ &||\nabla (w-w_h)||_{L^2(\Omega)} ||\nabla (u-u_h)||_{L^2(\Omega)} \leq Ch||u-u_h||_{L^2(\Omega)} Ch||f||_{L^2(\Omega)}. \end{aligned} \tag{27}$$

Division by $||u - u_h||_{L^2(\Omega)}$ immediately implies that there is a $K > 0$ such that

$$0 \leq ||u - u_h||_{L^2(\Omega)} \leq Kh^2 ||f||_{L^2(\Omega)}. \tag{28}$$

We summarize this result in the following theorem:

Theorem 9 *Let $f \in L^2(\Omega)$ and let $\Omega \subset \mathbb{R}^d$ be bounded by polygonal boundary Γ, and suppose that*

$$u \in H_0^1(\Omega), \text{ such that for all } \phi \in H_0^1(\Omega), \text{ we have} \int_\Omega \nabla\phi \cdot \nabla u \, d\Omega = \int_\Omega \phi f \, d\Omega, \tag{29}$$

and let u_h be the finite element solution with piecewise linear elements, then there is a $K > 0$ such that

$$0 \leq ||u - u_h||_{L^2(\Omega)} \leq Kh^2 ||f||_{L^2(\Omega)}. \tag{30}$$

We can also extend this theorem to general degree of finite element approximations:

Theorem 10 *Let $u \in H^{p+1}(\Omega)$ and let $\Omega \subset \mathbb{R}^d$ be bounded by piecewise smooth boundary Γ, and suppose that*

$$u \in H_0^1(\Omega), \text{ such that for all } \phi \in H_0^1(\Omega), \text{ we have} \int_\Omega \nabla\phi \cdot \nabla u \, d\Omega = \int_\Omega \phi f \, d\Omega, \tag{31}$$

and let u_h be the finite element solution with piecewise p-th order basis functions, then there is a $K > 0$ such that

$$0 \leq ||u - u_h||_{L^2(\Omega)} \leq Kh^{p+1} ||D^{p+1}u||_{L^2(\Omega)}. \tag{32}$$

We note that this text has not treated the errors that may appear if one approximates a (piecewise) smooth boundary by polygons (for $p = 1$) or by Lagrangian interpolation functions of order p, where p also represents the order of the basis functions by which the solution is spanned. This contribution will give an additional error of the same order as the error that we get from approximating the solution by interpolation functions. The proof of this fact is very technical.

Further, in order for the error bound to be valid, it is needed that the right-hand side function satisfies $f \in L^2(\Omega)$. If this is not satisfied, then the current error bound analysis breaks down. Sometimes, convergence with a lower error bound can be demonstrated, or even convergence in the norm of a different Sobolev space (so not in a Hilbert space), for instance in $W^{1,1}(\Omega)$, which is given by

$$W^{1,1}(\Omega) := \left\{ f \in L^1(\Omega) \; : \; \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \in L^1(\Omega) \right\},$$

where

$$L^1(\Omega) = \left\{ f : \Omega \to \mathbb{R} \; : \; \int_\Omega |f| d\Omega \text{ is finite.} \right\}$$

This very interesting, but complicated, convergence analysis is beyond the scope of this version of this manuscript. More theory on the error analysis of finite-element methods with Dirac Delta functions can be found in the studies by Scott [3] and Bertoluzza [4] among many other studies.

For cases in which the differential operator is symmetric and positive definite, such as for

$$-\Delta u + \lambda u = f(\mathbf{x}), \quad \mathbf{x} \in \Omega,$$

$$u = g(\mathbf{x}), \qquad \mathbf{x} \in \Gamma,$$

where $\lambda \geq 0$, the analysis is entirely analogous except for some straightforward technicalities. In the case of nonsymmetric differential operators, such as in the convection-diffusion equation

$$\begin{aligned} -\Delta u + \mathbf{q} \cdot \nabla u + \lambda u = f(\mathbf{x}), \quad & \mathbf{x} \in \Omega, \\ u = g(\mathbf{x}), \quad & \mathbf{x} \in \Gamma_1, \\ \mathbf{n} \cdot \nabla u = 0, \quad & \mathbf{x} \in \Gamma_2, \end{aligned}$$

the convergence analysis becomes somewhat more complicated. Here one uses the adjoint operator in Nitsche's trick for estimation of the L^2–norm of the finite-element error. Existence and uniqueness of the weak solution in H^1 for the case that $\mathbf{q} \cdot \mathbf{n} \geq 0$ (that is an outflow boundary) is demonstrated by the use of Lax-Milgram's Lemma. Some of the finite-element convergence analysis can be done in the assignments.

In general, more information regarding finite-element error analysis can be found in the books by, among many others, Brenner & Scott [5] and in Braess [6] for solid mechanics and saddle point problems. We note that there are many more references that treat convergence theory of finite element methods.

Assignment 3 *Given any continuous, coercive nonsymmetric bilinear form $a(u,v)$ for $u,v \in V(\Omega)$. Let u be solution to*

Find $u \in V(\Omega)$, such that for all $\phi \in V(\Omega)$, we have $a(u,\phi) = (\phi, f)$,

and let $V_h(\Omega) \subset V(\Omega)$ be a finite dimensional (that is, the finite-element space) subspace, for which we have

Find $u_h \in V_h(\Omega)$, such that for all $\phi_h \in V_h(\Omega)$, we have $a(u_h, \phi_h) = (\phi_h, f)$.

Use

There is a $K > 0$ such that $|a(u,v)| \leq K||u||_V \; ||v||_V$ for all $u,v \in V$ (continuity);

There is a $c > 0$ such that $a(u,u) \geq c||u||_V^2$ for all $u \in V$ (coerciveness).

to prove Céa's Lemma:

$$||u - u_h||_V \leq \frac{K}{c} ||u - v_h||_V.$$

Remark: *Note that $a(u,v) \neq a(v,u)$ for a nonsymmetric bilinear form.*

Assignment 4 *Consider the linear diffusion-reaction equation, take for simplicity*

$$\begin{aligned} -\Delta u + \lambda u = f(\mathbf{x}), \quad & \mathbf{x} \in \Omega, \\ u = 0, \quad & \mathbf{x} \in \Gamma, \end{aligned}$$

where $\lambda \geq 0$,
(a) use Nitsche's trick that for linear elements, to arrive at

$$||u - u_h||_{L^2(\Omega)} \leq Kh^2.$$

(b) do the same for quadratic elements, to arrive at

$$||u - u_h||_{L^2(\Omega)} \leq Kh^3.$$

Hint: *Make use of the results from Assignment 3 and Theorem 7.*

Assignment 5 *Consider the linear convection-diffusion-reaction equation, take for simplicity*

$$\begin{aligned} -\Delta u + \mathbf{q} \cdot \nabla u + \lambda u = f(\mathbf{x}), \quad & \mathbf{x} \in \Omega, \\ u = 0, \quad & \mathbf{x} \in \Gamma_1, \\ \mathbf{n} \cdot \nabla u = 0, \quad & \mathbf{x} \in \Gamma_2, \end{aligned}$$

where $\lambda \geq 0$, use Nitsche's trick with the adjoint differential operator (if $a(u,\phi)$ is the associated bilinear corresponding to the linear differential operator, then $a(\phi,u)$ is its adjoint) that for linear elements, to arrive at

$$||u-u_h||_{L^2(\Omega)} \leq Kh^2.$$

Hint: *Make use of the results from Assignment 3 and Theorem 7.*

References

[1] C. Vuik, F.J. Vermolen, M.B. van Gijzen, M.J. Vuik (2015). Numerical methods for ordinary differential equations. Delft Academic Press (VSSD), second edition

[2] J. van Kan, A. Segal, F.J. Vermolen (2014). Numerical methods in scientific computing. Delft Academic Press (VSSD), second edition

[3] R. Scott (1973). Finite element convergence for singular data. Numerische Mathematik 21 (4): 317–327.

[4] S. Bertoluzza, A. Decoene, L. Lacouture, S. Martin (2018). Local error estimates of the finite element method for an elliptic problem with a Dirac source term. Numerical Methods for Partial Differential Equations 34: 99–120.

[5] S.C. Brenner, L.R. Scott (2008). The mathematical theory of finite element methods. Springer *Texts in applied mathematics*, 15, third edition.

[6] D. Braess (2001). Finite elements: theory, fast solvers, and applications in solid mechanics, Cambridge University Press, third edition.