NUMERICAL PARTIAL DIFFERENTIAL EQUATIONS

Lecture Notes, Winter 2011/12

Christian Clason

January 31, 2012

Institute for Mathematics and Scientific Computing
Karl-Franzens-Universität Graz

CONTENTS

I BACKGROUND

II CONFORMING FINITE ELEMENTS

III NONCONFORMING FINITE ELEMENTS

IV TIME-DEPENDENT PROBLEMS

Part I

BACKGROUND

INTRODUCTION

Partial differential equations appear in many mathematical models of physical, biological and economic phenomena, such as elasticity, electromagnetics, fluid dynamics, quantum mechanics, pattern formation or derivative valuation. However, closed-form or analytic solutions of these equations are only available in very specific cases (e.g., for simple geometries or constant coefficients), and so one has to resort to numerical approximations of these solutions.

In these notes, we will consider *finite element methods*, which have developed into one of the most flexible and powerful frameworks for the numerical (approximate) solution of partial differential equations. They were first proposed by Richard Courant [Courant 1943]; but the method did not catch on until engineers started applying similar ideas in the early 1950s. Their mathematical analysis began later, with the works of Miloš Zlámal [Zlámal 1968].

Knowledge of real analysis (in particular, Lebesgue integration theory) and functional analysis (especially Hilbert space theory) as well as some familiarity of the weak theory of partial differential equations is assumed, although the fundamental results of the latter (Sobolev spaces and the variational formulation of elliptic equations) are recalled in Chapter 2.

These notes are mostly based on the following works:

[1] E. Süli (2011). "Finite Element Methods for Partial Differential Equations". Lecture notes. URL: http://people.maths.ox.ac.uk/suli/fem.pdf

[2] R. Rannacher (2008). "Numerische Mathematik 2". Lecture notes. URL: http://numerik.iwr.uni-heidelberg.de/~lehre/notes/num2/numerik2.pdf

[3] S. C. Brenner and L. R. Scott (2008). *The Mathematical Theory of Finite Element Methods*. 3rd ed. Vol. 15. Texts in Applied Mathematics. Springer, New York

[4] D. Braess (2007). *Finite Elements*. 3rd ed. Cambridge University Press, Cambridge

[5] A. Ern and J.-L. Guermond (2004). *Theory and Practice of Finite Elements*. Vol. 159. Applied Mathematical Sciences. Springer-Verlag, New York

[6] V. Thomée (2006). *Galerkin Finite Element Methods for Parabolic Problems*. 2nd ed. Vol. 25. Springer Series in Computational Mathematics. Springer-Verlag, Berlin

1 OVERVIEW OF THE FINITE ELEMENT METHOD

We begin with a “bird’s-eye view” of the finite element method by considering a simple one-dimensional example. Since the goal here is to give the flavor of the results and techniques used in the construction and analysis of finite element methods, not all arguments will be completely rigorous (especially those involving derivatives and function spaces). These gaps will be filled by the more general theory in the following chapters.

1.1 VARIATIONAL FORM OF PDES

Consider for a given function f the two-point boundary value problem

$$\begin{cases} -u''(x) = f(x) & \text{for } x \in (0,1), \\ \quad u(0) = 0, & \quad u'(1) = 1. \end{cases} \tag{BVP}$$

Similar to Krylov methods, the idea is to pass from this differential equation to a system of linear equations which can be solved on a computer by projection onto a finite-dimensional subspace. Any projection requires a kind of inner product, which we introduce now. We begin by multiplying this equation with any sufficiently regular function v with $v(0) = 0$, integrating over $x \in [0,1]$ and integrating by parts. Then any solution u of (BVP) satisfies

$$\begin{aligned} (f,v) := \int_0^1 f(x)v(x)\,dx &= -\int_0^1 u''(x)v(x)\,dx \\ &= \int_0^1 u'(x)v'(x)\,dx \\ &=: a(u,v), \end{aligned}$$

where we have used that $u'(1) = 0$ and $v(0) = 0$. Let us (formally for now) define the space

$$V := \left\{ v \in L^2(0,1) : a(v,v) < \infty,\ v(0) = 0 \right\}.$$

Then we can pose the following problem: Find $u \in V$ such that

$$\text{(W)} \qquad a(u, v) = (f, v) \qquad \text{for all } v \in V$$

holds. This is called the *weak* or *variational* form of (BVP) (since v varies over all V). If the solution u of (W) is twice continuously differentiable and f is continuous, one can prove (by taking suitable *test functions* v) that u satisfies (BVP). On the other hand, there are solutions of (W) even for discontinuous $f \in L^2(0, 1)$. Since then the second derivative of u is discontinuous, u cannot be a solution of (BVP). For this reason, $u \in V$ satisfying (W) is called a *weak solution* of (BVP).

Note that the Dirichlet boundary condition $u(0) = 0$ appears explicitly in the definition of V, while the Neumann condition $u'(1) = 0$ is implicitly incorporated in the variational formulation. In the context of finite element methods, Dirichlet conditions are therefore frequently called *essential conditions*, while Neumann conditions are referred to as *natural conditions*.

1.2 RITZ–GALERKIN APPROXIMATION

The fundamental idea is now to approximate u by considering (W) on a *finite-dimensional* subspace $S \subset V$. We are thus looking for $u_S \in S$ satisfying

$$(\text{W}_S) \qquad a(u_S, v) = (f, v) \qquad \text{for all } v \in S.$$

Note that this is still the same equation; only the function spaces have changed. This is a crucial point in (conforming) finite element methods. (Nonconforming methods, for which $S \not\subseteq V$ or $v \notin V$, will be treated in Part C.)

Since S is finite-dimensional, there exists a basis $\varphi_1, \dots, \varphi_n$ of S. Due to the bilinearity of $a(\cdot, \cdot)$, it suffices to require $u_S = \sum_{i=1}^n U_i \varphi_i \in S$ to satisfy

$$a(u_S, \varphi_j) = (f, \varphi_j) \qquad \text{for all } 1 \leqslant j \leqslant n.$$

If we define

$$\begin{aligned} \mathbf{U} &= (U_1, \dots, U_n)^T \in \mathbb{R}^n, \\ \mathbf{F} &= (F_1, \dots, F_n)^T \in \mathbb{R}^n, \quad && F_i = (f, \varphi_i), \\ \mathbf{K} &= (K_{ij}) \in \mathbb{R}^{n \times n}, && K_{ij} = a(\varphi_i, \varphi_j), \end{aligned}$$

we have that u_S satisfies (W_S) if and only if (“iff”) $\mathbf{KU} = \mathbf{F}$. This linear system has a unique solution iff $\mathbf{KV} = \mathbf{o}$ implies $\mathbf{V} = \mathbf{o}$. To show this, we set $v := \sum_{i=1}^n V_i \varphi_i \in S$. Then,

$$\mathbf{o} = \mathbf{KV} = (a(v, \varphi_1), \dots, a(v, \varphi_n))^T$$

implies that

$$0 = \sum_{i=1}^{n} V_i a(v, \varphi_i) = a(v, v) = \int_0^1 v'(x)^2 \, dx.$$

This means that v' must vanish almost everywhere and thus v is constant. (This argument will be made rigorous in the next chapter.) Since $v(0) = 0$, we deduce $v \equiv 0$, and hence, by the linear independence of the φ, $V_i = 0$ for all $1 \leqslant i \leqslant n$.

There are two remarks to made here. First, we have argued unique solvability of the finite-dimensional system by appealing to the properties of the variational problem to be approximated. This is a standard argument in finite element methods, and the fact that the approximation "inherits" the well-posedness of the variational problem is one of the strengths of the Galerkin approach. Second, this argument shows that the *stiffness matrix* $\mathbf{K}$ is (symmetric and) positive definite, since $\mathbf{V}^T\mathbf{K}\mathbf{V} = a(v, v) > 0$ for all $\mathbf{V} \neq 0$.

Now that we have an approximate solution $u_S \in S$, we are interested in estimating the *discretization error* $\|u_S - u\|$, which of course depends on the choice of S. The fundamental observation is that by subtracting (W) and (W_S) for the same test function $w \in S$, we obtain

$$a(u - u_S, w) = 0 \quad \text{for all } w \in S.$$

This key property is called *Galerkin orthogonality*, and expresses that the discretization error is (in some sense) orthogonal to S. This can be exploited to derive error estimates in the *energy norm*

$$\|v\|_E^2 = a(v, v) \quad \text{for } v \in V.$$

It is straightforward to verify that this indeed defines a norm, which satisfies the *Cauchy–Schwarz* inequality

$$a(v, w) \leqslant \|v\|_E \, \|w\|_E \quad \text{for all } v, w \in V.$$

We can thus show that for any $v \in S$,

$$\begin{aligned}\|u - u_S\|_E^2 &= a(u - u_S, u - v) + a(u - u_S, v - u_S) \\ &= a(u - u_S, u - v) \\ &\leqslant \|u - u_S\|_E \, \|u - v\|_E\end{aligned}$$

due to the Galerkin orthogonality for $v - u_S \in S$. Taking the infimum over all v, we obtain

$$\|u - u_S\|_E \leqslant \inf_{v \in S} \|u - v\|_E \, ,$$

and equality holds – and hence this infimum is attained – for $u_S \in S$ solving (W_S). The discretization error is thus completely determined by the approximation error of the solution u of (W) by functions in S:

$$\|u - u_S\|_E = \min_{v \in S} \|u - v\|_E \, . \tag{1.1}$$

To derive error estimates in the $L^2(0,1)$ norm

$$\|v\|_{L^2}^2 = (v, v) = \int_0^1 v(x)^2 \, dx,$$

we apply a *duality argument* (also called *Aubin–Nitsche trick*). Let w be the solution of the *dual* (or *adjoint*) *problem*

$$\begin{cases} -w''(x) = u(x) - u_S(x) & \text{for } x \in (0,1), \\ \quad w(0) = 0, \qquad w'(1) = 1. \end{cases} \tag{1.2}$$

Inserting this into the error and integrating by parts (using $(u-u_S)(0) = w'(1) = 0$), we obtain for all $v \in S$ the estimate

$$\begin{aligned} \|u - u_S\|_{L^2}^2 &= (u - u_S, u - u_S) = (u - u_S, -w'') \\ &= (u', w') - (u_S', w') \\ &= a(u - u_S, w) - a(u - u_S, v) \\ &= a(u - u_S, w - v) \\ &\leqslant \|u - u_S\|_E \, \|w - v\|_E \,. \end{aligned}$$

Dividing by $\|u - u_S\|_{L^2} = \|w''\|_{L^2}$, inserting (1.2) and taking the infimum over all $v \in S$ yields

$$\|u - u_S\|_{L^2} \leqslant \inf_{v \in S} \|w - v\|_E \, \|u - u_S\|_E \, \|w''\|_{L^2}^{-1} \,.$$

To continue, we require an *approximation property* for S: There exists $\varepsilon > 0$ such that

$$\inf_{v \in S} \|f - v\|_E \leqslant \varepsilon \, \|f''\|_{L^2} \tag{1.3}$$

holds for sufficiently smooth $f \in V$. If we can apply this estimate to w and u, we obtain

$$\begin{aligned} \|u - u_S\|_{L^2} &\leqslant \varepsilon \, \|u - u_S\|_E = \varepsilon \min_{v \in S} \|u - v\|_E \\ &\leqslant \varepsilon^2 \, \|u''\|_{L^2} = \varepsilon^2 \, \|f\|_{L^2} \,. \end{aligned}$$

This is another key observation: The error estimate depends on the regularity of the weak solution u, and hence on the data f. The smoother u, the better the approximation. The finite element method is characterized by a special class of subspaces – of piecewise polynomials – which have these approximation properties.

1.3 APPROXIMATION BY PIECEWISE POLYNOMIALS

Given a set of *nodes*

$$0 = x_0 < x_1 < \cdots < x_n = 1,$$

set

$$S = \left\{v \in C^0(0,1) : v|_{[x_{i-1},x_i]} \in P_1 \text{ and } v(0) = 0\right\},$$

where P_1 is the space of all linear polynomials. (The fact that $S \subset V$ is not obvious, and will be proved later.) This is a subspace of the space of linear splines. A basis of S, which is especially convenient for the implementation, is formed by the linear B-splines (*hat functions*)

$$\varphi_i(x) = \begin{cases} \frac{x-x_{i-1}}{x_i-x_{i-1}} & \text{if } x \in [x_{i-1}, x_i], \\ \frac{x_{i+1}-x}{x_{i+1}-x_i} & \text{if } x \in [x_i, x_{i+1}], \\ 0 & \text{else,} \end{cases}$$

for $1 \leqslant i \leqslant n$, which satisfy

$$\varphi_i(x_j) = \delta_{ij} := \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

This *nodal basis property* immediately yields linear independence of the φ_i. To show that the φ_i span S, we consider the *interpolant* of $v \in V$, defined as

$$v_I := \sum_{i=1}^{n} v(x_i)\varphi_i(x) \in S.$$

For $v \in S$, the interpolation error $v - v_I$ is piecewise linear as well, and since $(v - v_I)(0) = 0$, this implies that $v - v_I \equiv 0$. Any $v \in S$ can thus be written as a linear combination of φ_i (given by its interpolant), and hence the φ_i form a basis of S. We also note that this implies that the *interpolation operator* $\mathcal{I} : V \to S, v \mapsto v_I$ is a projection.

We are now in a position to prove the approximation property of S. Let

$$h := \max_{1\leqslant i\leqslant n} h_i, \qquad h_i := (x_i - x_{i-1})$$

denote the *mesh size*. We wish to show that there exists a constant $C > 0$ such that for all sufficiently smooth $u \in V$,

$$\|u - u_I\|_E \leqslant Ch \|u''\|_{L^2}.$$

It suffices to consider this error separately on each *element* $[x_{i-1}, x_i]$, i.e., to show

$$\int_{x_{i-1}}^{x_i} (u - u_I)'(x)^2 \, dx \leqslant C^2 h_i^2 \int_{x_{i-1}}^{x_i} u''(x)^2 \, dx.$$

Furthermore, since u_I is piecewise linear, the error $e := u - u_I$ satisfies $(e|_{[x_{i-1},x_i]})'' = (u|_{[x_{i-1},x_i]})''$. Using the affine transformation $\tilde{e}(t) := e(x(t))$ with $x(t) = x_{i-1} + t(x_i - x_{i-1})$ (a *scaling argument*), we just need to prove

$$\int_0^1 \tilde{e}'(t)^2 \, dt \leqslant c \int_0^1 \tilde{e}''(t)^2 \, dt. \tag{1.4}$$

(This is an elementary version of *Poincaré's inequality*). Since u_I is the nodal interpolant of u, the error satisfies $\tilde{e}(x_{i-1}) = \tilde{e}(x_i) = 0$. In addition, u_I is linear and u continuously differentiable on $[x_{i-1}, x_i]$. By Rolle's theorem, there hence exists a $\xi \in (0,1)$ with $\tilde{e}'(\xi) = 0$. Thus, for all $y \in [0,1]$ we have (with $\int_a^b f(t)\,dt = -\int_b^a f(t)\,dt$ for $a > b$)

$$\tilde{e}'(y) = \int_\xi^y w''(t)\,dt.$$

We can now use the Cauchy–Schwarz inequality to estimate

$$\begin{aligned}|\tilde{e}'(y)|^2 = \left|\int_\xi^y \tilde{e}''(t)\right|^2 &\leqslant \left|\int_\xi^y 1\,dt\right| \cdot \left|\int_\xi^y \tilde{e}''(t)^2\,dt\right| \\ &\leqslant |y-\xi| \int_0^1 \tilde{e}''(t)^2\,dt.\end{aligned}$$

Integrating both sides with respect to y and taking the supremum over all $\xi \in (0,1)$ yields (1.4) with

$$c := \sup_{\xi \in (0,1)} \int_0^1 |y-\xi|\,dy = \frac{1}{2}.$$

Summing over all elements and estimating h_i by h shows the approximation property (1.3) for S with $\varepsilon := ch$. For this choice of S, the solution u_S of (W_S) satisfies

$$\|u - u_S\|_{L^2} \leqslant c^2 h^2 \|u''\|_{L^2}. \tag{1.5}$$

This is called an *a priori estimate*, since it only requires knowledge of the given data $f = u''$, but not of the solution u_S. It tells us that if we can make the mesh size h arbitrarily small, we can approximate the solution u of (W) arbitrarily well. Note that the power of h is one order higher for the $L^2(0,1)$ norm compared to the energy norm.

1.4 IMPLEMENTATION

As seen in section 1.2, the numerical computation of $u_S \in S$ boils down to solving the linear system $\mathbf{KU} = \mathbf{F}$ for the vector of coefficients $\mathbf{U}$, e.g., by the method of conjugate gradients (since $\mathbf{K}$ is symmetric and positive definite). The missing step is the computation of the elements $K_{ij} = a(\varphi_i, \varphi_j)$ of $\mathbf{K}$ and the entries $F_j = (f, \varphi_j)$ of $\mathbf{F}$. (This procedure is called *assembly*.) In principle, this can be performed by computing the integrals for each pair (i,j) in a nested loop (*node-based assembly*). A more efficient approach (especially in higher dimensions) is *element-based assembly*: The integrals are split into sums of contributions from each element, e.g.,

$$a(\varphi_i, \varphi_j) = \int_0^1 \varphi_i'(x)\varphi_j'(x)\,dx = \sum_{k=1}^n \int_{x_{k-1}}^{x_k} \varphi_i'(x)\varphi_j'(x)\,dx,$$

and the contributions from a single element for all (i,j) are computed simultaneously. Here we can exploit that by the definition, φ_i is non-zero only on the two elements $[x_{i-1},x_i]$ and $[x_i,x_{i+1}]$. Hence, for each element $[x_{k-1},x_k]$, the integrals are non-zero only for pairs (i,j) with $k-1 \leqslant i,j \leqslant k$. Note that this implies that $\mathbf{K}$ is tridiagonal and therefore *sparse* (meaning that the number of non-zero elements grows as n, not n^2), which allows efficient solution of the linear system even for large n.

Another useful observation is that except for an affine transformation, the basis functions are the same on each element. We can thus use the substitution rule to transform the integrals over $[x_{k-1},x_k]$ to the *reference element* $[0,1]$. Setting $\xi(x) = \frac{x-x_{k-1}}{x_k-x_{k-1}}$ and

$$\hat\varphi_1(\xi) = 1-\xi, \qquad \hat\varphi_2(\xi) = \xi,$$

we have that $\varphi_{k-1}(x) = \hat\varphi_1(\xi(x))$ and $\varphi_k(x) = \hat\varphi_2(\xi(x))$ and thus that

$$\int_{x_{k-1}}^{x_k} \varphi_i'(x)\varphi_j'(x)\,dx = (x_k - x_{k-1})^{-1}\int_0^1 \hat\varphi_{\tau(i)}'(\xi)\hat\varphi_{\tau(j)}'(\xi)\,d\xi,$$

where

$$\tau(i) = \begin{cases} 1 & \text{if } i = k-1 \\ 2 & \text{if } i = k, \end{cases}$$

is the so-called *global-to-local index*. (Correspondingly, the inverse mapping τ^{-1} is called the *local-to-global index*.) The contribution from the element $[x_{k-1},x_k]$ to $a(\varphi_i,\varphi_j)$ is thus

$$a_k(\varphi_i,\varphi_j) = \begin{cases} h_k^{-1} & \text{if } i = j, \\ -h_k^{-1} & \text{if } i \neq j. \end{cases}$$

The right-hand side (f,φ_j) can be computed in a similar way, using numerical quadrature if necessary. Alternatively, one can replace f by its nodal interpolant $f_I = \sum_{i=0}^n f(x_i)\varphi_i$ and use

$$(f,\varphi_j) \approx (f_I,\varphi_j) = \sum_{i=0}^n f(x_i)\,(\varphi_i,\varphi_j) =: \mathbf{Mf}.$$

The elements $M_{ij} := (\varphi_i,\varphi_j)$ of the *mass matrix* $\mathbf{M}$ are again computed elementwise using transformation to the reference element:

$$\int_{x_{k-1}}^{x_k} \varphi_i(x)\varphi_j(x)\,dx = h_k\int_0^1 \hat\varphi_{\tau(i)}(\xi)\hat\varphi_{\tau(j)}(\xi)\,d\xi = \begin{cases} \frac{h_k}{3} & \text{if } i = j, \\ \frac{h_k}{6} & \text{if } i \neq j. \end{cases}$$

This can be done at the same time as assembling $\mathbf{K}$.

Finally, the Dirichlet condition $u(0) = 0$ is enforced by replacing the first equation in the linear system by $U_0 = 0$, i.e., replacing the first row of $\mathbf{K}$ by $(1,0,\dots)^T$ and the first element of $\mathbf{Mf}$ by 0. The main advantage is that this procedure can easily be extended to non-homogeneous Dirichlet conditions. The full algorithm (in MATLAB-like notation) for our boundary value problem is given in Algorithm 1.1.

Algorithmus 1.1 Finite element method in 1d

```
Input: $0 = x_0 < \cdots < x_n = 1$, $F := (f(x_0), \dots, f(x_n))^T$
Set $K_{ij} = M_{ij} = 0$
for $k = 1, \dots, n$ do
    Set $h_k = x_k - x_{k-1}$
    Set $K_{k-1:k,k-1:k} \leftarrow K_{k-1:k,k-1:k} + \frac{1}{h_k}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$
    Set $M_{k-1:k,k-1:k} \leftarrow M_{k-1:k,k-1:k} + \frac{h_k}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$
end for
$K_{0,1:n} = 0$, $K_{0,0} = 1$, $M_{0,0:n} = 0$
Solve $KU = MF$
Output: $U$
```

1.5 A POSTERIORI ERROR ESTIMATES AND ADAPTIVITY

The a priori estimate (1.5) is important for proving convergence as the mesh size $h \to 0$, but often pessimistic in practice since it depends on the global regularity of u''. If $u''(x)$ is large only in some parts of the domain, it would be preferable to reduce the mesh size locally. For this, *a posteriori estimates* are useful, which involve the computed solution u_S but are able to give information on which elements should be refined (i.e., replaced by a larger number of smaller elements).

We consider again the space S of piecewise linear finite elements on the nodes $x_0, \dots, x_N$ with mesh size h, as defined in section 1.3. We once more apply a duality trick: Let w be the solution of

$$\begin{cases} -w''(x) = u(x) - u_S(x) & \text{for } x \in (0,1), \\ \quad w(0) = 0, \qquad w'(1) = 1, \end{cases}$$

and proceed as before, yielding

$$\|u - u_S\|_{L^2}^2 = a(u - u_S, w - v)$$

for all $v \in S$. We now choose $v = w_I \in S$, the interpolant of w. Then we have

$$\begin{aligned} \|u - u_S\|_{L^2}^2 &= a(u - u_S, w - w_I) = a(u, w - w_I) - a(u_S, w - w_I) \\ &= (f, w - w_I) - a(u_S, w - w_I). \end{aligned}$$

Note that the unknown solution u of (W) no longer appears on the right hand side. We now use the specific choice of v to localize the error inside each element $[x_{i-1}, x_i]$: Writing the integrals over $[0,1]$ as sums of integrals over the elements, we can integrate by parts on each element and use the fact that $(w - w_I)(x_i) = 0$ to obtain

$$\begin{aligned}
\|u - u_S\|_{L^2}^2 &= \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)(w - w_I)(x)\,dx - \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} u_S'(x)(w - w_I)'(x)\,dx \\
&= \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} (f + u_S'')(x)(w - w_I)(x)\,dx \\
&\leqslant \sum_{i=1}^{n} \left(\int_{x_{i-1}}^{x_i} (f + u_S'')(x)^2\,dx\right)^{\frac{1}{2}} \left(\int_{x_{i-1}}^{x_i} (w - w_I)(x)^2\,dx\right)^{\frac{1}{2}}
\end{aligned}$$

by the Cauchy–Schwarz inequality. The first term contains the *finite element residual*

$$R_h := f + u_S'',$$

which we can evaluate after computing u_S. For the second term, one can show (similarly as in the proof of the inequality (1.5)) that

$$\left(\int_{x_{i-1}}^{x_i} (w - w_I)(x)^2\,dx\right)^{\frac{1}{2}} \leqslant \|w - w_I\|_{L^2} \leqslant \frac{h_i^2}{4}\,\|w''\|_{L^2}$$

holds, from which we deduce

$$\begin{aligned}
\|u - u_S\|_{L^2}^2 &\leqslant \frac{1}{4}\,\|w''\|_{L^2} \sum_{i=1}^{n} h_i^2\,\|R_h\|_{L^2(x_{i-1},x_i)} \\
&= \frac{1}{4}\,\|u - u_S\|_{L^2} \sum_{i=1}^{n} h_i^2\,\|R_h\|_{L^2(x_{i-1},x_i)}
\end{aligned}$$

by the definition of w. This yields the *a posteriori estimate*

$$\|u - u_S\|_{L^2} \leqslant \frac{1}{4} \sum_{i=1}^{n} h_i^2\,\|R_h\|_{L^2(x_{i-1},x_i)}\,.$$

This estimate can be used for an adaptive procedure: Given a tolerance $\tau > 0$,

```
Choose initial mesh $0 < x_0^{(0)} < \dots x_{N^{(0)}}^{(0)} = 1$, compute corresponding solution $u_{S^{(0)}}$,
evaluate $R_{h^{(0)}}$, set $m = 0$
while $\sum_{i=1}^{n} (h_i^{(m)})^2 \|R_{h^{(m)}}\|_{L^2(x_{i-1}^{(m)},x_i^{(m)})} < \tau$ do
    Choose new mesh $0 < x_0^{(m+1)} < \dots x_{N^{(m+1)}}^{(m+1)} = 1$
    compute corresponding solution $u_{S^{(m+1)}}$
    evaluate $R_{h^{(m+1)}}$
    set $m \leftarrow m + 1$
end while
```

There are different strategies to choose the new mesh. They should be *reliable*, meaning that the error on the new mesh in a certain norm can be guaranteed to be less than a given tolerance, and *efficient*, meaning that the number of new nodes should not be larger than necessary. One (simple) possibility is to refine those elements where $\|R_h\|$ is largest (or larger than a given threshold) by replacing them with two elements of half size.

2 VARIATIONAL THEORY OF PDES

In this chapter, we collect – for the most part without proof – some necessary results from functional analysis and the weak theory of (elliptic) partial differential equations. Details and proofs can be found in, e.g., [Adams and Fournier 2003], [Evans 2010] and [Zeidler 1995a].

2.1 FUNCTION SPACES

As we have seen, the regularity of the solution of partial differential equations plays a crucial role in how well it can be approximated numerically. This regularity can be described by the two properties of (Lebesgue-)*integrability* and *differentiability*.

LEBESGUE SPACES Let Ω be an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}_0$. We recall that for $1 \leqslant p \leqslant \infty$,

$$L^p(\Omega) := \left\{ f \text{ measurable} : \|f\|_{L^p(\Omega)} < \infty \right\}$$

with

$$\begin{aligned} \|f\|_{L^p(\Omega)} &= \left(\int_\Omega |f(x)|^p \, dx \right)^{\frac{1}{p}} \qquad \text{for } 1 \leqslant p < \infty, \\ \|f\|_{L^\infty(\Omega)} &= \operatorname*{ess\,sup}_{x \in \Omega} |f(x)| \end{aligned}$$

are Banach spaces of (equivalence classes up to equality apart from a set of zero measure of) Lebesgue-integrable functions. The corresponding norms satisfy *Hölder's inequality*

$$\|fg\|_{L^1(\Omega)} \leqslant \|f\|_{L^p(\Omega)} \|f\|_{L^q(\Omega)}$$

if $p^{-1} + q^{-1} = 1$ (with $\infty^{-1} := 0$). For bounded Ω, this implies (by using $g \equiv 1$) that $L^p(\Omega) \subset L^q(\Omega)$ for $p \geqslant q$. We will also use the space

$$L^1_{\text{loc}}(\Omega) := \left\{ f : f|_K \in L^1(K) \text{ for all compact } K \subset \Omega \right\}.$$

For $p = 2$, $L^p(\Omega)$ is a Hilbert space with inner product

$$(f, g) := \langle f, g\rangle_{L^2(\Omega)} = \int_\Omega f(x)g(x)\,dx,$$

and Hölder's inequality for $p = q = 2$ reduces to the *Cauchy–Schwarz inequality*.

HÖLDER SPACES We now consider functions which are continuously differentiable. It will be convenient to use a *multi-index*

$$\alpha := (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n,$$

for which we define its *length* $|\alpha| := \sum_{i=1}^n \alpha_i$, to describe the (partial) *derivative of order* $|\alpha|$

$$D^\alpha f(x_1, \dots, x_n) := \frac{\partial^{|\alpha|} f(x_1, \dots, x_n)}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}$$

For brevity, we will often write $\partial_i := \frac{\partial}{\partial x_i}$. We denote by $C^k(\Omega)$ the set of all continuous functions f for which $D^\alpha f$ is continuous for all $|\alpha| \leqslant k$. If Ω is bounded, $C^k(\overline{\Omega})$ is the set of all functions in $C^k(\Omega)$ for which all $D^\alpha f$ can be extended to a continous function on $\overline{\Omega}$, the closure of Ω. These spaces are Banach spaces if equipped with the norm

$$\|f\|_{C^k(\overline{\Omega})} = \sum_{|\alpha|\leqslant k} \sup_{x\in\overline{\Omega}} |D^\alpha f(x)|.$$

Finally, we define $C_0^k(\overline{\Omega})$ as the space of all $f \in C^k(\overline{\Omega})$ whose support (the closure of $\{x \in \Omega : f(x) \neq 0\}$) is a bounded subset of Ω, as well as

$$C_0^\infty(\overline{\Omega}) = \bigcap_{k\geqslant 0} C_0^k(\overline{\Omega})$$

(and similarly $C^\infty(\overline{\Omega})$).

SOBOLEV SPACES If we are interested in weak solutions, it is clear that the Hölder spaces entail a too strong notion of (pointwise) differentiability. All we required is that the derivative is integrable, and that an integration by parts is meaningful. This motivates the following definition: A function $f \in L^1_{\text{loc}}(\Omega)$ has a *weak derivative* if there exists $g \in L^1_{\text{loc}}(\Omega)$ such that

$$\int_\Omega g(x)\varphi(x)\,dx = (-1)^{|\alpha|} \int_\Omega f(x) D^\alpha \varphi(x)\,dx \tag{2.1}$$

for all $\varphi \in C_0^\infty(\overline{\Omega})$. In this case, the weak derivative is (uniquely) defined as $D^\alpha f := g$. For $f \in C^k(\Omega)$, the weak derivative coincides with the usual (pointwise) derivative (justifying the abuse of notation), but the weak derivative exists for a larger class of functions such as continuous and piecewise smooth functions. For example, $f(x) = |x|$, $x \in \Omega = (-1, 1)$, has the weak derivative $Df(x) = \mathrm{sign}(x)$, while $Df(x)$ itself does not have any weak derivative.

We can now define the *Sobolev spaces* $W^{k,p}(\Omega)$ for $k \in \mathbb{N}_0$ and $1 \leqslant p \leqslant \infty$:

$$W^{k,p}(\Omega) = \{f \in L^p(\Omega) : D^\alpha f \in L^p(\Omega) \text{ for all } |\alpha| \leqslant k\},$$

which are Banach spaces when endowed with the norm

$$\begin{aligned} \|f\|_{W^{k,p}(\Omega)} &= \left(\sum_{|\alpha|\leqslant k} \|D^\alpha f\|^p_{L^p(\Omega)}\right)^{\frac{1}{p}} \qquad \text{for } 1 \leqslant p < \infty, \\ \|f\|_{W^{k,\infty}(\Omega)} &= \sum_{|\alpha|\leqslant k} \|D^\alpha f\|_{L^\infty(\Omega)}. \end{aligned}$$

We shall also use the corresponding semi-norms

$$\begin{aligned} |f|_{W^{k,p}(\Omega)} &= \left(\sum_{|\alpha|= k} \|D^\alpha f\|^p_{L^p(\Omega)}\right)^{\frac{1}{p}} \qquad \text{for } 1 \leqslant p < \infty, \\ |f|_{W^{k,\infty}(\Omega)} &= \sum_{|\alpha|= k} \|D^\alpha f\|_{L^\infty(\Omega)}. \end{aligned}$$

We are now concerned with the relation between the different norms introduced so far. For many of these results to hold, we require that the boundary $\partial\Omega$ of Ω is sufficiently smooth. We shall henceforth assume that $\Omega \subset \mathbb{R}^n$ has a *Lipschitz boundary*, meaning that $\partial\Omega$ can be parametrized by a finite set of functions which are uniformly Lipschitz continuous. (This condition is satisfied, for example, by polygons for $n = 2$ and polyhedra for $n = 3$.) A fundamental result is then the following approximation property (which does not hold for arbitrary domains).

Theorem 2.1 (Density[1]). *For* $1 \leqslant p < \infty$ *and any* $k \in \mathbb{N}_0$, $C^\infty(\overline{\Omega})$ *is dense in* $W^{k,p}(\Omega)$.

This theorem allows us to prove results for Sobolev spaces – such as chain rules – by showing them for smooth functions (in effect, transferring results for usual derivatives to their weak counterparts). This is called a *density argument*.

The next theorem states that, within limits determined by the spatial dimension, we can trade differentiability for integrability for Sobolev space functions.

[1] Originally shown in a paper by Meyers and Serrin rightfully celebrated both for its content and the brevity of its title, “$H = W$”. For the proof, see, e.g., [Evans 2010, § 5.3.3, Th. 3], [Adams and Fournier 2003, Th. 3.17]

Theorem 2.2 (Sobolev[2], Rellich–Kondrachov[3] embedding). *Let $1 \leqslant p, q < \infty$ and $\Omega \subset \mathbb{R}^n$ be a bounded open set with Lipschitz boundary. Then, the following embeddings are continuous:*

$$W^{k,p}(\Omega) \subset \begin{cases} L^q(\Omega) & \textit{if } p < \frac{n}{k} \textit{ and } p \leqslant q \leqslant \frac{np}{n-p}, \\ L^q(\Omega) & \textit{if } p = \frac{n}{k} \textit{ and } p \leqslant q < \infty, \\ C^0(\overline{\Omega}) & \textit{if } p > \frac{n}{k}. \end{cases}$$

Moreover, the following embeddings are compact:

$$W^{k,p}(\Omega) \subset \begin{cases} L^q(\Omega) & \textit{if } p \leqslant \frac{n}{k} \textit{ and } 1 \leqslant q < \frac{n-pk}{np}, \\ C^0(\overline{\Omega}) & \textit{if } p > \frac{n}{k}. \end{cases}$$

In particular, the embedding $W^{k,p}(\Omega) \subset W^{k-1,p}(\Omega)$ is compact for all k and $1 \leqslant p \leqslant \infty$.

We now prove a generalization of the observation that piecewise linear and continuous functions have a weak derivative.

Theorem 2.3. *Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain which can be partitioned into $N \in \mathbb{N}$ Lipschitz subdomains Ω_j, i.e., $\overline{\Omega} = \bigcup_{j=1}^N \overline{\Omega}_j$. Then, for every $k \geqslant 1$ and $1 \leqslant p \leqslant \infty$,*

$$\left\{v \in C^{k-1}(\overline{\Omega}) : v|_{\Omega_j} \in C^k(\overline{\Omega}_j), 1 \leqslant j \leqslant N\right\} \subset W^{k,p}(\Omega).$$

Proof. It suffices to show the inclusion for $k = 1$. Let $v \in C^0(\overline{\Omega})$ such that $v|_{\Omega_j} \in C^1(\overline{\Omega}_j)$ for all $1 \leqslant j \leqslant N$. We need to show that $\partial_i v$ exists as a weak derivative for all $1 \leqslant i \leqslant n$ and that $\partial_i v \in L^p(\Omega)$. An obvious candidate is

$$w_i := \begin{cases} \partial_i v|_{\Omega_j}(x) & \text{if } x \in \Omega_j, 1 \leqslant j \leqslant N, \\ c & \text{else} \end{cases}$$

for arbitrary $c \in \mathbb{R}$. By the embedding $C^0(\overline{\Omega}_j) \subset L^\infty(\Omega_j)$ and the boundedness of Ω, $w_i \in L^p(\Omega)$ for any $1 \leqslant p \leqslant \infty$. It remains to verify (2.1). By splitting the integration into a sum over the Ω_j and integrating by parts on each subdomain (where v is continuously differentiable), we obtain for any $\varphi \in C_0^\infty(\overline{\Omega})$

$$\begin{aligned} \int_\Omega w_i \varphi \, dx &= \sum_{j=1}^N \int_{\Omega_j} \partial_i v|_{\Omega_i} \varphi \, dx \\ &= \sum_{j=1}^N \int_{\partial\Omega_j} v|_{\Omega_j} \varphi \, (\nu_j)_i \, dx - \sum_{j=1}^N \int_{\Omega_j} v|_{\Omega_j} \partial_i \varphi \, dx \\ &= \sum_{j=1}^N \int_{\partial\Omega_j} v|_{\Omega_j} \varphi \, (\nu_j)_i \, dx - \int_\Omega v \partial_i \varphi \, dx, \end{aligned}$$

[2] e.g., [Evans 2010, § 5.6], [Adams and Fournier 2003, Th. 4.12]

[3] e.g., [Evans 2010, § 5.7], [Adams and Fournier 2003, Th. 6.3]

where $\nu_j = ((\nu_j)_1, \dots, (\nu_j)_n)$ is the outer normal vector on Ω_j, which exists almost everywhere since Ω_j is a Lipschitz domain. Now the sum over the boundary integrals vanishes since either $\varphi(x) = 0$ if $x \in \partial\Omega_j \subset \partial\Omega$ or $v|_{\Omega_j}(x)\varphi(x)(\nu_j)_i(x) = -v|_{\Omega_k}(x)\varphi(x)(\nu_k)_i(x)$ if $x \in \partial\Omega_j \cap \partial\Omega_k$ due to the continuity of v. This implies $\partial_i v = w_i$ by definition. □

Next, we would like to see how Dirichlet boundary conditions make sense for weak solutions. For this, we define a *trace operator* T (via limits of continous functions) which maps a function f on a bounded domain $\Omega \subset \mathbb{R}^n$ to a function Tf on $\partial\Omega$.

Theorem 2.4 (Trace theorem[4]). *Let* $kp < n$ *and* $q \leqslant (n-1)p/(n-kp)$, *and* $\Omega \subset \mathbb{R}^n$ *be a bounded open set with Lipschitz boundary. Then,* $T : W^{k,p}(\Omega) \to L^q(\partial\Omega)$ *is a bounded linear operator, and there exists a constant* $C > 0$ *depending only on* p *and* Ω *such that for all* $f \in W^{k,p}(\Omega)$,

$$\|Tf\|_{L^q(\partial\Omega)} \leqslant C \|f\|_{W^{k,p}(\Omega)} .$$

If $kp = n$, *this holds for any* $p \leqslant q < \infty$.

This implies (although it is not obvious[5]) that

$$W_0^{k,p}(\Omega) := \left\{ f \in W^{k,p}(\Omega) : T(D^\alpha f) = 0 \in L^p(\partial\Omega) \text{ for all } |\alpha| < k \right\}$$

is well-defined, and that $W^{k,p}(\Omega) \cap C_0^\infty(\overline{\Omega})$ is dense in $W_0^{k,p}(\Omega)$.

For functions in $W_0^{1,p}(\Omega)$, the semi-norm $|\cdot|_{W^{1,p}(\Omega)}$ is equivalent to the full norm $\|\cdot\|_{W^{1,p}(\Omega)}$.

Theorem 2.5 (Poincaré's inequality[6]). *Let* $1 \leqslant p < \infty$ *and let* Ω *be a bounded open set. Then, there exists a constant* $C > 0$ *depending only on* p *and* Ω *such that for all* $f \in W_0^{1,p}(\Omega)$,

$$\|f\|_{W^{1,p}(\Omega)} \leqslant C|f|_{W^{1,p}(\Omega)}$$

holds.

The proof is very similar to the argumentation in Chapter 1, using the density of $C_0^\infty(\overline{\Omega})$ in $W_0^{1,p}(\Omega)$; in particular, it is sufficient that Tf is zero on a part of the boundary $\partial\Omega$ of non-zero measure. In general, we have that any $f \in W^{1,p}(\Omega)$, $1 \leqslant p \leqslant \infty$, for which $D^\alpha f = 0$ almost everywhere in Ω for all $|\alpha| = 1$ must be constant.

Again, $W^{k,p}(\Omega)$ is a Hilbert space for $p = 2$, with inner product

$$\langle f, g \rangle_{W^{k,2}(\Omega)} = \sum_{|\alpha| \leqslant k} (D^\alpha f, D^\alpha g) .$$

[4] e.g., [Evans 2010, § 5.5], [Adams and Fournier 2003, Th. 5.36]

[5] e.g., [Evans 2010, § 5.5, Th. 2], [Adams and Fournier 2003, Th. 5.37]

[6] e.g, [Adams and Fournier 2003, Cor. 6.31]

For this reason, one usually writes $H^k(\Omega) := W^{k,2}(\Omega)$. In particular, we will often consider $H^1(\Omega) := W^{1,2}(\Omega)$ and $H_0^1(\Omega) := W_0^{1,2}(\Omega)$. With the usual notation $\nabla f := (\partial_1 f, \dots, \partial_n f)$ for the gradient of f, we can write

$$|f|_{H^1(\Omega)} = \|\nabla f\|_{L^2(\Omega)}$$

for the semi-norm on $H^1(\Omega)$ (which, by the Poincaré inequality 2.5, is a full norm on $H_0^1(\Omega)$) and

$$\langle f, g\rangle_{H^1(\Omega)} = (f, g) + (\nabla f, \nabla g)$$

for the inner product on $H^1(\Omega)$. Finally, we denote the topological dual of $H_0^1(\Omega)$ (i.e., the space of all continuous linear functionals on $H_0^1(\Omega)$) by $H^{-1}(\Omega) := (H_0^1(\Omega))^*$, which is endowed with the operator norm

$$\|f\|_{H^{-1}(\Omega)} = \sup_{\varphi \in H_0^1(\Omega), \varphi \neq 0} \frac{\langle f, \varphi\rangle_{H^{-1}(\Omega), H_0^1(\Omega)}}{\|\varphi\|_{H_0^1(\Omega)}},$$

where $\langle f, \varphi\rangle_{V^*,V} := f(\varphi)$ denotes the *duality pairing* between a Banach space V and its dual V^*.

We can now tie together some loose ends from Chapter 1. The space V can now be rigorously defined as

$$V := \left\{v \in H^1(0,1) : v(0) = 0\right\},$$

which makes sense due to the embedding (for $n = 1$) of $H^1(0,1)$ in $C([0,1])$. Due to Poincaré's inequality, $|v|_{H^1(\Omega)}^2 = a(v,v) = 0$ implies $\|v\|_{H^1(\Omega)} = 0$ and hence $v = 0$. Similarly, the existence of a unique weak solution $u \in V$ follows from the Riesz representation theorem. Finally, Theorem 2.3 guarantees that $S \subset V$.

2.2 WEAK SOLUTIONS OF ELLIPTIC PDES

In most of these notes, we consider *boundary value problems* of the form

$$-\sum_{j,k=1}^{n} \partial_j(a_{jk}(x)\partial_k u) + \sum_{j=1}^{n} b_j(x)\partial_j u + c(x)u = f \tag{2.2}$$

on a bounded open set $\Omega \subset \mathbb{R}^n$, where a_{jk}, b_j, c and f are given functions on Ω. We do not fix boundary conditions at this time. This problem is called *elliptic* if there exists a constant $\alpha > 0$ such that

$$\sum_{j,k=1}^{n} a_{jk}(x)\xi_j\xi_k \geqslant \alpha \sum_{j=1}^{n} \xi_j^2 \quad \text{for all } \xi \in \mathbb{R}^n, x \in \Omega. \tag{2.3}$$

Assuming all functions and the domain are sufficiently smooth, we can multiply by a smooth function v, integrate over $x \in \Omega$ and integrate by parts to obtain

$$\sum_{j,k=1}^{n} (a_{jk}\partial_j u, \partial_k v) + \sum_{j=1}^{n} (b_j \partial_j u, v) + (cu, v) + \sum_{j,k=1}^{n} (a_{jk}\partial_k u \nu_j, v)_{\partial\Omega} = (f, v), \tag{2.4}$$

where $\nu := (\nu_1, \dots, \nu_n)^T$ is the outward unit normal on $\partial\Omega$ and

$$(f, g)_{\partial\Omega} := \int_{\partial\Omega} f(x)g(x)\,dx.$$

Note that this formulation only requires $a_{jk}, b_j, c \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$ in order to be well-defined. We then search for $u \in V$ satisfying (2.4) for all $v \in V$ including boundary conditions which we will discuss next. We will consider the following three conditions:

DIRICHLET CONDITIONS We require $u = g$ on $\partial\Omega$ (in the sense of traces) for given $g \in L^2(\partial\Omega)$. If $g = 0$ (*homogeneous* Dirichlet conditions), we take $V = H_0^1(\Omega)$, in which case the boundary integrals in (2.4) vanish since $v = 0$ on $\partial\Omega$. The weak formulation is thus: Find $u \in H_0^1(\Omega)$ satisfying

$$a(u, v) := \sum_{j,k=1}^{n} (a_{jk}\partial_j u, \partial_k v) + \sum_{j=1}^{n} (b_j \partial_j u, v) + (cu, v) = (f, v)$$

for all $v \in H_0^1(\Omega)$.

If $g \neq 0$, and g and $\partial\Omega$ are sufficiently smooth (e.g., $g \in H^1(\partial\Omega)$ with $\partial\Omega$ of class C^1)[7], we can find a function $u_g \in H^1(\Omega)$ such that $Tu_g = g$. We then set $u = \tilde{u} + u_g$, where $\tilde{u} \in H_0^1(\Omega)$ satisfies

$$a(u, v) = (f, v) - a(u_g, v)$$

for all $v \in H_0^1(\Omega)$.

NEUMANN CONDITIONS We require $\sum_{j,k=1}^{n} a_{jk}\partial_k u \nu_j = g$ on $\partial\Omega$ for given $g \in L^2(\partial\Omega)$. In this case, we can substitute this equality in the boundary integral in (2.4) and take $V = H^1(\Omega)$. We then look for $u \in H^1(\Omega)$ satisfying

$$a(u, v) = (f, v) + (g, v)_{\partial\Omega}$$

for all $v \in H^1(\Omega)$, where v in the last inner product should be understood in the sense of traces, i.e., as Tv.

[7] [Renardy and Rogers 2004, Th. 7.40]

ROBIN CONDITIONS We require $du + \sum_{j,k=1}^{n} a_{jk}\partial_k u \nu_j = g$ on $\partial\Omega$ for given $g \in L^2(\partial\Omega)$ and $d \in L^\infty(\partial\Omega)$. Again we can substitute this in the boundary integral and take $V = H^1(\Omega)$. The weak form is then: Find $u \in H^1(\Omega)$ satisfying

$$a_R(u, v) := a(u, v) + (du, v)_{\partial\Omega} = (f, v) + (g, v)_{\partial\Omega}$$

for all $v \in H^1(\Omega)$.

These problems have a common form: For a given Hilbert space V, a bilinear form $a : V \times V \to \mathbb{R}$ and a linear functional $F : V \to \mathbb{R}$ (e.g., $F : v \mapsto (f, v)$ in the case of Dirichlet conditions), find $u \in V$ such that

$$a(u, v) = F(v), \qquad \text{for all } v \in V. \tag{2.5}$$

The existence and uniqueness of a solution can be guaranteed by the Lax–Milgram theorem, which is a generalization of the Riesz representation theorem (note that a is in general not symmetric).

Theorem 2.6 (Lax–Milgram theorem)**.** *Let a Hilbert space* V*, a bilinear form* $a : V \times V \to \mathbb{R}$ *and a linear functional* $F : V \to \mathbb{R}$ *be given satisfying the following conditions:*

(i) Coercivity*: There exists a* $c_1 > 0$ *such that*

$$a(v, v) \geqslant c_1 \|v\|_V^2$$

for all $v \in V$.

(ii) Continuity*: There exists* $c_2, c_3 > 0$ *such that*

$$\begin{aligned} a(v, w) &\leqslant c_2 \|v\|_V \|w\|_V, \\ F(v) &\leqslant c_3 \|v\|_V \end{aligned}$$

for all $v, w \in V$.

Then, there exists a unique solution $u \in V$ *to problem* (2.5) *satisfying*

$$\|u\|_V \leqslant \frac{1}{c_1} \|F\|_{V^*}. \tag{2.6}$$

Proof. For every fixed $u \in V$, the mapping $v \mapsto a(u, v)$ is a linear functional on V, which is continuous by assumption (ii), and so is F. By the Riesz–Fréchet representation theorem,[8] there exist unique $\varphi_u, \varphi_F \in V$ such that

$$\langle \varphi_u, v\rangle_V = a(u, v) \quad \text{and} \quad \langle \varphi_F, v\rangle_V = F(v)$$

[8] e.g., [Zeidler 1995a, Th. 2.E]

for all $v \in V$. We recall that $w \mapsto \varphi_w$ is a continuous linear mapping from V^* to V with operator norm 1. Thus,

$$0 = a(u, v) - F(v) = \langle \varphi_u - \varphi_F, v \rangle_V = 0$$

for all $v \in V$, which holds if and only if $\varphi_u = \varphi_F$ in V.

We now wish to solve this equation using the Banach fixed point theorem.[9] For $\delta > 0$, consider the mapping $T : V \to V$,

$$T(v) = v - \delta(\varphi_v - \varphi_F).$$

If T is a contraction, then there exists a unique fixed point u such that $T(u) = u$ and hence $\varphi_u - \varphi_F = 0$. It remains to show that there exists a $\delta > 0$ such that T is a contraction, i.e., there exists $0 < L < 1$ with $\|Tv_1 - Tv_2\|_V \leqslant L \|v_1 - v_2\|_V$. Let $v_1, v_2 \in V$ be arbitrary and set $v = v_1 - v_2$. Then we have

$$\begin{aligned}
\|Tv_1 - Tv_2\|_V^2 &= \|v_1 - v_2 - \delta(\varphi_{v_1} - \varphi_{v_2})\|_V^2 \\
&= \|v - \varphi_v\|_V^2 \\
&= \|v\|_V^2 - 2\delta \langle v, \varphi_v \rangle_V + \delta^2 \|\varphi_v\|_V^2 \\
&= \|v\|_V^2 - 2\delta a(v, v) + \delta^2 a(v, \varphi_v) \\
&\leqslant \|v\|_V^2 - 2\delta c_1 \|v\|_V^2 + \delta^2 c_2 \|v\|_V \|\varphi_v\|_V \\
&\leqslant (1 - 2\delta c_1 + \delta^2 c_2) \|v_1 - v_2\|_V^2 .
\end{aligned}$$

We can thus choose $\delta > 0$ such that $L^2 := (1 - 2\delta c_1 + \delta^2 c_2) < 1$, and the Banach fixed point theorem yields existence and uniqueness of the solution $u \in V$.

To show the estimate (2.6), assume $u \neq 0$ (otherwise the inequality holds trivially). Note that F is a bounded linear functional by assumption (ii), hence $F \in V^*$. We can then apply the coercivity of a and divide by $\|u\|_V \neq 0$ to obtain

$$c_1 \|u\|_V \leqslant \frac{a(u, u)}{\|u\|_V} \leqslant \sup_{v \in V} \frac{a(u, v)}{\|v\|_V} = \sup_{v \in V} \frac{F(v)}{\|v\|_V} = \|F\|_{V^*} .$$

□

We can now give sufficient conditions on the coefficients a_{jk}, b_j, c and d such that the boundary value problems defined above have a unique solution.

Theorem 2.7 (Well-posedness). *Let $a_{jk} \in L^\infty(\Omega)$ satisfying the ellipticity condition (2.3) with constant $\alpha > 0$, $b_j, c \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ be given. Set $\beta = \alpha^{-1} \sum_{j=1}^n \|b_j\|_{L^\infty(\Omega)}^2$.*

[9] e.g., [Zeidler 1995a, Th. 1.A]

a) *The homogeneous Dirichlet problem has a unique solution* $u \in H^1_0(\Omega)$ *if*

$$c(x) - \frac{\beta}{2} \geqslant 0 \quad \textit{for almost all } x \in \Omega.$$

In this case, there exists a $C > 0$ *such that*

$$\|u\|_{H^1(\Omega)} \leqslant C \, \|f\|_{L^2(\Omega)} .$$

Consequently, the inhomogeneous Dirichlet problem for $g \in H^1(\partial\Omega)$ *has a unique solution satisfying*

$$\|u\|_{H^1(\Omega)} \leqslant C(\|f\|_{L^2(\Omega)} + \|g\|_{H^1(\partial\Omega)}).$$

b) *The Neumann problem for* $g \in L^2(\partial\Omega)$ *has a unique solution* $u \in H^1(\Omega)$ *if*

$$c(x) - \frac{\beta}{2} > 0 \quad \textit{for almost all } x \in \Omega.$$

In this case, there exists a $C > 0$ *such that*

$$\|u\|_{H^1(\Omega)} \leqslant C(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}).$$

c) *The Robin problem for* $g \in L^2(\partial\Omega)$ *and* $d \in L^\infty(\partial\Omega)$ *has a unique solution if*

$$c(x) - \frac{\beta}{2} \geqslant 0 \quad \textit{for almost all } x \in \Omega. d(x) \qquad \geqslant 0 \quad \textit{for almost all } x \in \partial\Omega,$$

and at least one inequality is strict. In this case, there exists a $C > 0$ *such that*

$$\|u\|_{H^1(\Omega)} \leqslant C(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}).$$

Proof. We wish to apply the Lax–Milgram theorem. Continuity of a and F follow by the Hölder inequality and the boundedness of the coefficients. It thus remains to verify the coercivity of a, which we only do for the case of homogeneous Dirichlet conditions (the others being similar). Let $v \in H^1_0(\Omega)$ be given. First, the ellipticity of a_{jk} implies that

$$\int_\Omega \sum_{j,k=1}^n a_{jk}\partial_j v(x)\partial_k v(x)\,dx \geqslant \alpha \int_\Omega \sum_{j=1}^n \partial_j v(x)^2\,dx = \alpha \sum_{j=1}^n \|\partial_j v\|^2_{L^2(\Omega)} = \alpha|v|^2_{H^1(\Omega)}.$$

We then have by repeated application of Hölder's inequality

$$\begin{aligned} a(v,v) &\geqslant \alpha|v|^2_{H^1(\Omega)} - \left(\sum_{j=1}^n \|b_j\|^2_{L^\infty(\Omega)}\right)^{\frac{1}{2}} |v|_{H^1(\Omega)}\,\|v\|_{L^2(\Omega)} + \int_\Omega c(x)v(x)^2\,dx \\ &\geqslant \frac{\alpha}{2}|v|^2_{H^1(\Omega)} + \int_\Omega \left(c(x) - \frac{1}{2\alpha}\sum_{j=1}^n \|b_j\|^2_{L^\infty(\Omega)}\right)|v|^2\,dx, \end{aligned}$$

where we have used Young's inequality $ab \leqslant \frac{\alpha}{2}a^2 + \frac{1}{2\alpha}b^2$ for $a, b \in \mathbb{R}$ and any $\alpha > 0$. Under the assumption that $c - \frac{\beta}{2} \geqslant 0$, the second term is non-negative and we deduce using Poincaré's inequality that

$$a(v,v) \geqslant \frac{\alpha}{2}|v|^2_{H^1(\Omega)} \geqslant \frac{\alpha}{4}|v|^2_{H^1(\Omega)} + \frac{\alpha}{4c_\Omega^2}\|v\|^2_{L^2(\Omega)} \geqslant C\,\|v\|^2_{H^1(\Omega)}$$

holds for $C := \alpha/(4+4c_\Omega^2)$. □

Note that these conditions are not sharp; different ways of estimating the first-order terms in a give different conditions. For example, if $b_j \in W^{1,\infty}(\Omega)$, we can take $\beta = \sum_{j=1}^n \|\partial_j b_j\|_{L^\infty(\Omega)}$.

Naturally, if the data has higher regularity, we can expect more regularity of the solution as well. The corresponding theory is quite involved, and we give only two useful results.

Theorem 2.8 (Higher regularity[10]). *Let $\Omega \subset \mathbb{R}^n$ be bounded domain with C^{k+1} boundary, $k \geqslant 0$, $a_{jk} \in C^k(\overline{\Omega})$ and $b_j, c \in W^{k,\infty}(\Omega)$. Then for any $f \in H^k(\Omega)$, the solution of the homogeneous Dirichlet problem is in $H^{k+2}(\Omega) \cap H^1_0(\Omega)$, and there exists a $C > 0$ such that*

$$\|u\|_{H^{k+2}(\Omega)} \leqslant C(\|f\|_{H^k(\Omega)} + \|u\|_{H^1(\Omega)}).$$

Theorem 2.9 (Higher regularity[11]). *Let Ω be a convex polygon in $\mathbb{R}^2$ or a parallelepiped in $\mathbb{R}^3$, $a_{jk} \in C^1(\overline{\Omega})$ and $b_j, c \in C^0(\overline{\Omega})$. Then the solution of the homogeneous Dirichlet problem is in $H^2(\Omega)$, and there exists a $C > 0$ such that*

$$\|u\|_{H^2(\Omega)} \leqslant C\,\|f\|_{L^2(\Omega)}.$$

For non-convex polygons, $u \in H^2(\Omega)$ is not possible. This is due to the presence of so-called *corner singularities* at reentrant corners, which severely limits the accuracy of finite element approximations. This requires special treatment, and is a topic of extensive current research.

[10] [Troianiello 1987, Th. 2.24]

[11] [Grisvard 1985, Th. 5.2.2], [Ladyzhenskaya and Ural'tseva 1968, pp. 169–189]

Part II

CONFORMING FINITE ELEMENT APPROXIMATION OF ELLIPTIC PDES

GALERKIN APPROACH FOR VARIATIONAL PROBLEMS

3

We have seen that elliptic partial differential equations can be cast into the following form: Given a Hilbert space V, a bilinear form $a : V \times V \to \mathbb{R}$ and a continuous linear functional $F : V \to \mathbb{R}$, find $u \in V$ satisfying

$$\text{(W)} \qquad a(u, v) = F(v) \quad \text{for all } v \in V.$$

According to the Lax–Milgram theorem, this problem has a unique solution if there exist $c_1, c_2 > 0$ such that

$$\text{(3.1)} \qquad a(v, v) \geqslant c_1 \|v\|_V^2,$$

$$\text{(3.2)} \qquad a(u, v) \leqslant c_2 \|u\|_V \|v\|_V$$

hold for all $u, v \in V$ (which we will assume from here on).

The *conforming Galerkin approach* consists in choosing a (finite-dimensional) closed subspace $V_h \subset V$ and looking for $u_h \in V_h$ satisfying[1]

$$\text{(W}_h\text{)} \qquad a(u_h, v_h) = F(v_h) \quad \text{for all } v_h \in V_h.$$

Since we have chosen a closed $V_h \subset V$, the subspace V_h is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_V$. Furthermore, the conditions (3.1) and (3.2) are satisfied for all $u_h, v_h \in V_h$ as well. The Lax–Milgram theorem thus immediately yields the well-posedness of (W_h).

Theorem 3.1. *Under the assumptions of Theorem 2.6, for any closed subspace $V_h \subset V$, there exists a unique solution $u_h \in V_h$ of (W_h) satisfying*

$$\|u_h\|_V \leqslant \frac{1}{c_1} \|F\|_{V^*}.$$

[1] The subscript h stands for a *discretization parameter*, and indicates that we expect convergence of u_h to the solution of (W) as $h \to 0$.

The following result is essential for all error estimates of Galerkin approximations.

Lemma 3.2 (Céa's lemma)**.** *Let u_h be the solution of* (W_h) *for given $V_h \subset V$ and u be the solution of* (W)*. Then,*

$$\|u - u_h\|_V \leqslant \frac{c_2}{c_1} \inf_{v_h \in V_h} \|u - v_h\|_V ,$$

where c_1 and c_2 are the constants from (3.1) *and* (3.2)*.*

Proof. Since $V_h \subset V$, we deduce (by subtracting (W) and (W_h) with the same $v \in V_h$) the *Galerkin orthogonality*

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

Hence, for arbitrary $v_h \in V_h$, we have $v_h - u_h \in V_h$ and therefore $a(u - u_h, v_h - u_h) = 0$. Using (3.1) and (3.2), we obtain

$$\begin{aligned} c_1 \|u - u_h\|_V^2 &\leqslant a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\ &\leqslant c_2 \|u - u_h\|_V \|u - v_h\|_V . \end{aligned}$$

Dividing by $\|u - u_h\|_V$, rearranging, and taking the infimum over all $v_h \in V_h$ yields the desired estimate. □

This implies that the error of any (conforming) Galerkin approach is determined by the approximation error of the exact solution in V_h. The derivation of such error estimates will be the topic of the next chapters.

THE SYMMETRIC CASE The estimate in Céa's lemma is weaker than the corresponding estimate (1.1) for the model problem in Chapter 1. This is due to the symmetry of the bilinear form in the latter case, which allows characterizing solutions of (W) as minimizers of a functional.

Theorem 3.3. *If a is symmetric, $u \in V$ satisfies* (W) *if and only if u is the minimizer of*

$$J(v) := \tfrac{1}{2} a(v, v) - F(v)$$

over all $v \in V$.

Proof. For any $u, v \in V$ and $t \in \mathbb{R}$,

$$J(u + tv) = J(u) + t(a(u, v) - F(v)) + \frac{t^2}{2} a(v, v)$$

due to the symmetry of a. Assume that u satisfies $a(u, v) - F(v) = 0$ for all $v \in V$. Then, setting $t = 1$, we deduce that for all $v \neq 0$,

$$J(u + v) = J(u) + \tfrac{1}{2}a(v, v) > J(u)$$

holds. Hence, u is the unique minimizer of J. Conversely, if u is the (unique) minimizer of J, every directional derivative of J at u must vanish, which implies

$$0 = \frac{d}{dt}J(u + tv)|_{t=0} = a(u, v) - F(v)$$

for all $v \in V$. □

Together with coercivity and continuity, the symmetry of a implies that $a(u, v)$ is an inner product on V that induces an *energy norm* $\|u\|_a := a(u, u)^{\frac{1}{2}}$. (In fact, in many applications, the functional J represents an energy which is minimized in a physical system. For example in continuum mechanics, $\frac{1}{2}\|u\|_a^2 = \frac{1}{2}a(u, u)$ represents the elastic deformation energy of a body, and $-F(v)$ its potential energy under external load.)

Arguing as in Chapter 1.2, we see that the solution $u_h \in V_h$ of (W_h) – which is called *Ritz–Galerkin approximation* in this context – satisfies

$$\|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a,$$

i.e., u_h is the best approximation of u in V_h in the energy norm. Equivalently, one can say that the error $u - u_h$ is orthogonal to V_h in the inner product defined by a.

Often it is more useful to estimate the error in a weaker norm. This requires a *duality argument*. Let H be a Hilbert space with inner product $(\cdot, \cdot)$ and V a closed subspace satisfying the conditions of the Lax–Milgram theorem theorem such that the embedding $V \hookrightarrow H$ is continuous (e.g., $V = H^1 \subset L^2 = H$). Then we have the following estimate.

Lemma 3.4 (Aubin–Nitsche lemma). *Let* u_h *be the solution of* (W_h) *for given* $V_h \subset V$ *and* u *be the solution of* (W). *Then, there exists a* $C > 0$ *such that*

$$\|u - u_h\|_H \leqslant C\,\|u - u_h\|_V \sup_{g \in H}\left(\frac{1}{\|g\|_H}\inf_{v_h \in V_h}\|\varphi_g - v_h\|_V\right)$$

holds, where for given $g \in H$, φ_g *is the unique solution of the* adjoint problem

$$a(w, \varphi_g) = (g, w) \quad \textit{for all } w \in V.$$

Since a is symmetric, the existence of a unique solution of the adjoint problem is guaranteed by the Lax–Milgram theorem.

Proof. We make use of the dual representation of the norm in any Hilbert space,

$$\|w\|_H = \sup_{g\in H} \frac{(g,w)}{\|g\|_H}, \tag{3.3}$$

where the supremum is taken over all $g \neq 0$.

Now, inserting $w = u - u_h$ in the adjoint problem, we obtain for any $v_h \in V_h$ using the Galerkin orthogonality and continuity of a that

$$\begin{aligned}(g, u-u_h) &= a(u-u_h, \varphi_g)\\ &= a(u-u_h, \varphi_g - v_h)\\ &\leqslant C\,\|u-u_h\|_V\,\|\varphi_g - v_h\|_V\,.\end{aligned}$$

Inserting $w = u - u_h$ into (3.3), we thus obtain

$$\begin{aligned}\|u-u_h\|_H &= \sup_{g\in H}\frac{(g,u-u_h)}{\|g\|_H}\\ &\leqslant C\,\|u-u_h\|_V \sup_{g\in H}\frac{\|\varphi_g - v_h\|_V}{\|g\|_H}\end{aligned}$$

for arbitrary $v_h \in V_h$, and taking the infimum over all v_h yields the desired estimate. □

The Aubin–Nitsche lemma also holds for nonsymmetric a, provided both the original and the adjoint problem satisfy the conditions of the Lax–Milgram theorem (e.g., for constant coefficients b_j).

FINITE ELEMENT SPACES

4

Finite element methods are a special case of Galerkin methods, where the finite-dimensional subspace consists of piecewise polynomials. To construct these subspaces, we proceed in two steps:

1. We define a *reference element* and study polynomial interpolation on this element.
2. We use suitably modified copies of the reference element to partition the given domain and discuss how to construct a global interpolant using local interpolants on each element.

We then follow the same steps in proving interpolation error estimates for functions in Sobolev spaces.

4.1 CONSTRUCTION OF FINITE ELEMENT SPACES

To allow a unified study of the zoo of finite elements proposed in the literature,[1] we define a finite element in an abstract way.

Definition 4.1. Let

(i) $K \subset \mathbb{R}^n$ be a simply connected bounded open set with piecewise smooth boundary (the *element domain*, or simply *element* if there is no possibility of confusion),

(ii) $\mathcal{P}$ be a finite-dimensional space of functions defined on K (the *space of shape functions*),

(iii) $\mathcal{N} = \{N_1, \dots, N_d\}$ be a basis of $\mathcal{P}^*$ (the *set of nodal variables* or *degrees of freedom*).

Then $(K, \mathcal{P}, \mathcal{N})$ is a *finite element*.

[1] For a – far from complete – list of elements, see, e.g., [Brenner and Scott 2008, Chapter 3], [Ciarlet 2002, Section 2.2]

As we will see, condition (iii) guarantees that the interpolation problem on K using functions in $\mathcal{P}$ – and hence the Galerkin approximation – is well-posed. The nodal variables will play the role of interpolation conditions. This is a somewhat backwards definition compared to our introduction in Chapter 1 (where we have directly specified a basis for the shape functions). However, it leads to an equivalent characterization that allows much greater freedom in defining finite elements. The connection is given in the next definition.

Definition 4.2. Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element. The basis $\{\psi_1, \dots, \psi_d\}$ of $\mathcal{P}$ dual to $\mathcal{N}$, i.e., satisfying $N_i(\psi_j) = \delta_{ij}$, is called *nodal basis* of $\mathcal{P}$.

For example, for the linear finite elements in one dimension, $K = (0, 1)$, $\mathcal{P}$ is the space of linear polynomials, and $\mathcal{N} = \{N_1, N_2\}$ are the *point evaluations* $\mathcal{N}_1(v) = v(0)$, $\mathcal{N}_2(v) = v(1)$ for every $v \in \mathcal{P}$. The nodal basis is given by $\psi_1(x) = 1 - x$ and $\psi_2(x) = x$.

Condition (iii) is the only one that is difficult to verify. The following Lemma simplifies this task.

Lemma 4.3. *Let $\mathcal{P}$ be a d-dimensional vector space and let $\{N_1, \dots, N_d\}$ be a subset of $\mathcal{P}^*$. Then, the following statements are equivalent:*

a) $\{N_1, \dots, N_d\}$ *is a basis of* $\mathcal{P}^*$,

b) If $v \in \mathcal{P}$ satisfies $N_i(v) = 0$ for all $1 \leqslant i \leqslant d$, then $v = 0$.

Proof. Let $\{\psi_1, \dots, \psi_d\}$ be a basis of $\mathcal{P}$. Then, $\{N_1, \dots, N_d\}$ is a basis of $\mathcal{P}^*$ if and only if for any $L \in \mathcal{P}^*$, there exist (unique) α_i, $1 \leqslant i \leqslant d$ such that

$$L = \sum_{j=1}^{d} \alpha_j N_j.$$

Using the basis of $\mathcal{P}$, this is equivalent to $L(\psi_i) = \sum_{j=1}^{d} \alpha_j N_j(\psi_i)$ for all $1 \leqslant i \leqslant d$. Let us define the matrix $\mathbf{B} = (N_j(\psi_i))_{i,j=1}^{d}$ and the vectors

$$\mathbf{L} = (L(\psi_1), \dots, L(\psi_d))^T, \qquad \mathbf{a} = (\alpha_1, \dots, \alpha_d)^T.$$

Then, (a) is equivalent to $\mathbf{Ba} = \mathbf{L}$ being uniquely solvable, i.e., $\mathbf{B}$ being invertible.

On the other hand, given any $v \in \mathcal{P}$, we can write $v = \sum_{j=1}^{d} \beta_j \psi_j$. The condition (b) can be expressed as

$$\sum_{j=1}^{n} \beta_j N_i(\psi_j) = N_i(v) = 0 \quad \text{for all } 1 \leqslant i \leqslant d$$

implies $v = 0$, or, in matrix form, that $\mathbf{B}^T\mathbf{b} = 0$ implies $0 = \mathbf{b} := (\beta_1, \dots, \beta_d)^T$. But this too is equivalent to the fact that $\mathbf{B}$ is invertible. □

Note that (b) in particular implies that the corresponding interpolation problem is uniquely solvable.

Another useful tool is the following Lemma, which is a multidimensional form of polynomial division.

Lemma 4.4. *Let* $L \neq 0$ *be a linear function on* $\mathbb{R}^n$ *and* P *be a polynomial of degree* $d \geqslant 1$ *with* $P(x) = 0$ *for all* x *with* $L(x) = 0$*. Then, there exists a polynomial* Q *of degree* $d - 1$ *such that* $P = LQ$.

Proof. First, we note that affine transformations map the space of polynomials of degree d to itself. Thus, we can assume without loss of generality that P vanishes on the hyperplane orthogonal to the x_n axis, i.e. $L(x) = x_n$ and $P(\hat{x}, 0) = 0$, where $\hat{x} = (x_1, \dots, x_{n-1})$. Since the degree of P is d, we can write

$$P(\hat{x}, x_n) = \sum_{j=0}^{d} \sum_{|\alpha| \leqslant d-j} c_{\alpha,j} \hat{x}^{\alpha} x_n^j$$

for a multi-index $\alpha \in \mathbb{N}^{n-1}$ and $\hat{x}^{\alpha} = x_1^{\alpha_1} \cdots x_{n-1}^{\alpha_{n-1}}$. For $x_n = 0$, this implies

$$0 = P(\hat{x}, 0) = \sum_{|\alpha| \leqslant d} c_{\alpha,0} \hat{x}^{\alpha},$$

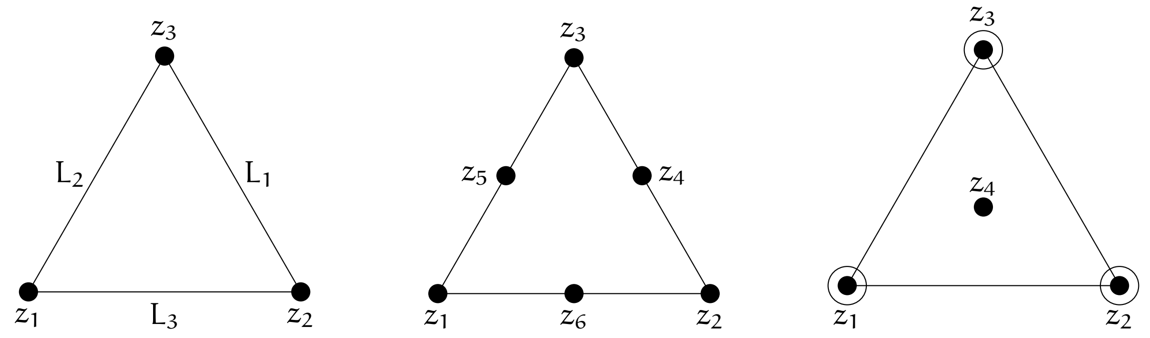
and therefore $c_{\alpha 0} = 0$ for all $|\alpha| \leqslant d$. Hence,

$$\begin{aligned} P(\hat{x}, x^n) &= \sum_{j=1}^{d} \sum_{|\alpha| \leqslant d-j} c_{\alpha,j} \hat{x}^{\alpha} x_n^j \\ &= x_n \sum_{j=1}^{d} \sum_{|\alpha| \leqslant d-j} c_{\alpha,j} \hat{x}^{\alpha} x_n^{j-1} \\ &=: x_n Q = LQ, \end{aligned}$$

where Q is of degree $d - 1$. □

4.2 EXAMPLES OF FINITE ELEMENTS

We restrict ourselves to the case $n = 2$ (higher dimensions being similar) and the most common examples.



(a) Linear Lagrange element (b) Quadratic Lagrange element (c) Cubic Hermite element

Figure 4.1: Triangular finite elements. Filled circles denote point evaluation, open circles gradient evaluations.

TRIANGULAR ELEMENTS Let K be a triangle and

$$P_k = \left\{ \sum_{|\alpha| \leqslant k} c_\alpha x^\alpha : c_\alpha \in \mathbb{R} \right\}$$

denote the space of all bivariate polynomials of total degree less than or equal k, e.g., $P_2 = \operatorname{span}\{1, x_1, x_2, x_1^2, x_2^2, x_1x_2\}$. It is straightforward to verify that P_k (and hence P_k^*) is a vector space of dimension $\frac{1}{2}(k+1)(k+2)$. We consider two types of interpolation conditions: function values (*Lagrange interpolation*) and gradient values (*Hermite interpolation*). The following examples define valid finite elements. Note that the argumentation is essentially the same as for the well-posedness of the corresponding one-dimensional polynomial interpolation problems.

- *Linear Lagrange elements*. Let $k = 1$ and take $\mathcal{P} = P_1$ (hence the dimension of $\mathcal{P}$ and $\mathcal{P}^*$ is 3) and $\mathcal{N} = \{N_1, N_2, N_3\}$ with $N_i(v) = v(z_i)$, where z_1, z_2, z_3 are the vertices of K (see Figure 4.1a). We need to show that condition (iii) holds, which we will do by way of Lemma 4.3. Suppose that $v \in P_1$ satisfies $v(z_1) = v(z_2) = v(z_3) = 0$. Since v is linear, it must also vanish on each line connecting the vertices, which can be defined as the zero-sets of the linear functions L_1, L_2, L_3. Hence, by Lemma 4.4, there exists a constant (i.e., polynomial of degree 0) c such that $v = cL_1$. Now let z_1 be the vertex not on the edge defined by L_1. Then,

  $$0 = v(z_1) = cL_1(z_1),$$

  and therefore $c = 0$ and so $v = 0$ (since $L_1(z_1) \neq 0$).

- *Quadratic Lagrange elements*: Let $k = 2$ and take $\mathcal{P} = P_2$ (hence the dimension of $\mathcal{P}$ and $\mathcal{P}^*$ is 6). Set $\mathcal{N} = \{N_1, N_2, N_3, N_4, N_5, N_6\}$ with $N_i(v) = v(z_i)$, where z_1, z_2, z_3 are again the vertices of K and z_4, z_5, z_6 are the midpoints of the edges described by the linear functions L_1, L_2, L_3, respectively (see Figure 4.1b). To show that condition (iii) holds, we argue as above. Let $v \in P_2$ vanish at z_i, $1 \leqslant i \leqslant 6$. On each edge, v

is a quadratic function that vanishes at three points (say, z_2, z_3, z_4) and thus must be identically zero. If L_1 is the function vanishing on the edge containing z_2, z_3, z_4, then by Lemma 4.4, there exists a linear polynomial Q_1 such that $P = L_1 Q_1$. By an analogous argument, $P = L_1 Q_1$ vanishes on the remaining edges as well. By definition, $L_1 = 0$ only on the first edge, and thus Q_1 must vanish. Now consider one of the remaining edges and let L_2 be the linear function which vanishes on it. Then, we can apply Lemma 4.4 to Q_1 to obtain a constant c such that $v = L_1 Q_1 = cL_1 L_2$. Taking the midpoint of the remaining edge, z_6, we have

$$0 = v(z_6) = cL_1(z_6)L_2(z_6),$$

and since neither L_1 nor L_2 vanish on z_6, we deduce $c = 0$ and hence $v = 0$.

- *Cubic Hermite elements*: Let $k = 3$ and take $\mathcal{P} = P_3$ (hence the dimension of $\mathcal{P}$ and $\mathcal{P}^*$ is 10). Instead of taking $\mathcal{N}$ as function evaluations at 10 suitable points, we take N_i, $1 \leqslant i \leqslant 4$ as the point evaluation at the vertices z_1, z_2, z_3 and the barycenter $z_4 = \frac{1}{3}(z_1 + z_2 + z_3)$ (see Figure 4.1c) and take the remaining nodal variables as gradient evaluations:

  $$N_{i+4}(v) = \partial_1 v(z_i), \qquad N_{i+7} = \partial_2 v(z_i), \quad 1 \leqslant i \leqslant 3.$$

  Now we again consider $v \in P_3$ with $N_i(v) = 0$ for all $1 \leqslant i \leqslant 10$. On each edge, v is a cubic polynomial with double roots at each vertex, and hence must vanish. Arguing as above, we can write $v = cL_1 L_2 L_3$ which implies

  $$0 = v(z_4) = cL_1(z_4)L_2(z_4)L_3(z_4)$$

  and hence $c = 0$ since the barycenter z_4 lies on neither of the edges. Therefore, $v = 0$.

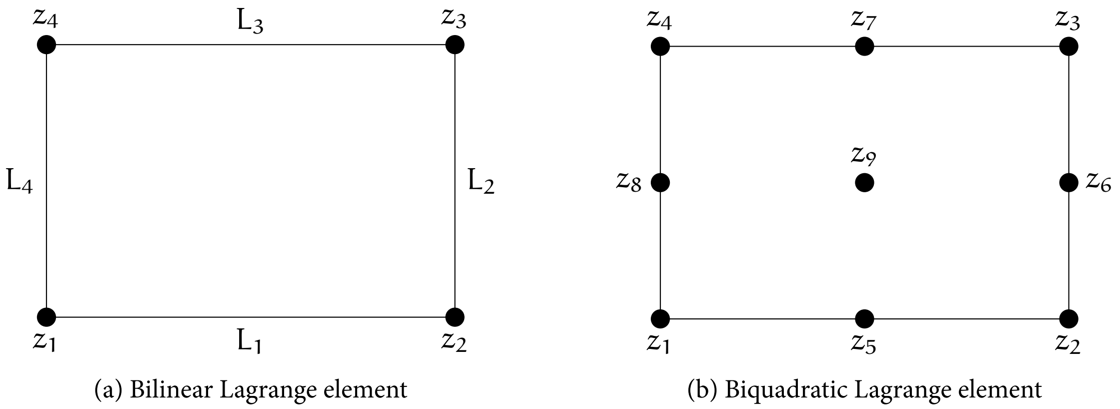
Both types of elements can be defined for arbitrary degree k. It should be clear from this that this definition of finite elements gives us a blueprint for constructing elements with desired properties. This should be contrasted with, e.g., the choice of finite difference stencils.

RECTANGULAR ELEMENTS For rectangular elements, we can follow a tensor-product approach. We consider the vector space

$$Q_k = \left\{ \sum_j c_j p_j(x_1) q_j(x_2) : c_j \in \mathbb{R}, p_j, q_j \in P_k \right\}$$

of products of univariate polynomials of degree up to k, which has dimension $(k+1)^2$. By analogous arguments as in the triangular case, we can show that the following examples are finite elements:

- *Bilinear Lagrange elements*: Let $k = 1$ and take $\mathcal{P} = Q_1$ (hence the dimension of $\mathcal{P}$ and $\mathcal{P}^*$ is 4) and $\mathcal{N} = \{N_1, N_2, N_3, N_4\}$ with $N_i(v) = v(z_i)$, where z_1, z_2, z_3, z_4 are the vertices of K (see Figure 4.2a).



(a) Bilinear Lagrange element

(b) Biquadratic Lagrange element

Figure 4.2: Rectangular finite elements. Filled circles denote point evaluation.

- *Biquadratic Lagrange elements*: Let $k = 2$ and take $\mathcal{P} = Q_2$ (hence the dimension of $\mathcal{P}$ and $\mathcal{P}^*$ is 9) and $\mathcal{N} = \{N_1, \ldots, N_9\}$ with $N_i(v) = v(z_i)$, where z_1, z_2, z_3, z_4 are the vertices of K, z_5, z_6, z_7, z_8 are the edge midpoints and z_9 is the centroid of K (see Figure 4.2b).

4.3 THE INTERPOLANT

We wish to estimate the error of the best approximation of a function in a finite element space. An upper bound for this approximation is given by stitching together interpolating polynomials on each element.

Definition 4.5. Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element and let $\{\psi_1, \ldots, \psi_d\}$ be the corresponding nodal basis of $\mathcal{P}$. For a given function v such that $N_i(v)$ is defined for all $1 \leqslant i \leqslant d$, the *local interpolant* of v is defined as

$$\mathcal{I}_K v = \sum_{i=1}^{d} N_i(v)\psi_i.$$

The nodal interpolant can be explicitly constructed once the dual basis is known. This can be simplified significantly if the reference element domain is chosen as, e.g., the unit simplex.

Useful properties of the local interpolant are given next.

Lemma 4.6. *Let* $(K, \mathcal{P}, \mathcal{N})$ *be a finite element and* $\mathcal{I}_K$ *the local interpolant. Then,*

a) The mapping $v \mapsto \mathcal{I}_K$ *is linear,*

b) $N_i(\mathcal{I}_K v)) = N_i(v)$, $1 \leqslant i \leqslant d$, *c)* $\mathcal{I}_K(v) = v$ *for all* $v \in \mathcal{P}$*, i.e.,* $\mathcal{I}_K$ *is a projection.*

Proof. The claim (a) follows directly from the linearity of the N_i. For (b), we use the definition of $\mathcal{I}_K$ and ψ_i to obtain

$$\begin{aligned}
N_i(\mathcal{I}_K v) &= N_i\left(\sum_{j=1}^{d} N_j(v)\psi_j\right) \\
&= \sum_{j=1}^{d} N_j(v)N_i(\psi_j) = \sum_{j=1}^{d} N_j(v)\delta_{ij} \\
&= N_i(v)
\end{aligned}$$

for all $1 \leqslant i \leqslant d$ and arbitrary v. This implies that $N_i(v - \mathcal{I}_K v) = 0$ for all $1 \leqslant i \leqslant d$, and hence by Lemma 4.3 that $\mathcal{I}_K v = v$. □

We now use the local interpolant on each element to define a global interpolant on a union of elements.

Definition 4.7. A *subdivision* $\mathcal{T}$ of a bounded open set $\Omega \subset \mathbb{R}^n$ is a finite collection of open sets K_i such that

(i) $\operatorname{int} K_i \cap \operatorname{int} K_j = \emptyset$ if $i \neq j$ and

(ii) $\bigcup_i \overline{K}_i = \overline{\Omega}$

Definition 4.8. Given a subdivision $\mathcal{T}$ of Ω such that for each K_i there is a finite element $(K_i, \mathcal{P}_i, \mathcal{N}_i)$ with local interpolant $\mathcal{I}_{K_i}$. Let m be the order of the highest partial derivative appearing in any nodal variable. Then, the *global interpolant* $\mathcal{I}_\mathcal{T} f$ of $f \in C^m(\overline{\Omega})$ on $\mathcal{T}$ is defined by

$$(\mathcal{I}_\mathcal{T} f)|_{K_i} = \mathcal{I}_{K_i} f \qquad \text{for all } K_i \in \mathcal{T}.$$

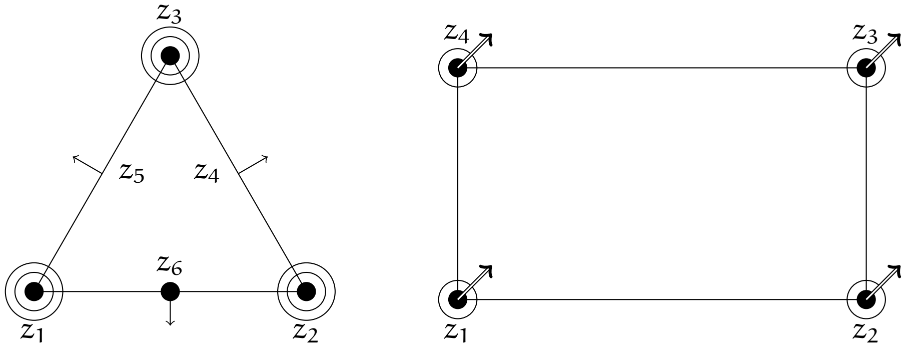
To obtain some regularity of the global interpolant, we need additional assumptions on the subdivision. Roughly speaking, where two elements meet, the corresponding nodal variables have to match as well. For triangular elements, this can be expressed concisely.

Definition 4.9. A *triangulation* of a bounded open set $\Omega \subset \mathbb{R}^2$ is a subdivision $\mathcal{T}$ of Ω such that

(i) every $K_i \in \mathcal{T}$ is a triangle and

(ii) no vertex of any triangle lies in the interior or on an edge of another triangle.

Similar conditions can be given for $n \geqslant 3$ (tetrahedra, simplices), in which case we will also speak of triangulations. Note that this supposes that Ω is polyhedral itself. For non-polyhedral domains, it is possible to use suitable geometric transformations on the elements at the boundary to obtain curved elements which faithfully represent it.



(a) Argyris triangle (b) Bogner–Fox–Schmit rectangle

Figure 4.3: C^1 elements. Filled circles denote point evaluation, double circles evaluation of gradients up to total order 2, and arrows evaluation of normal derivatives. The double arrow stands for evaluation of the second mixed derivative ∂^2_{12}.

Definition 4.10. A global interpolant $\mathcal{I}_\mathcal{T}$ has *continuity order* m (in short, "is C^m") if $\mathcal{I}_\mathcal{T} f \in C^m(\overline{\Omega})$ for all $f \in C^m(\overline{\Omega})$. The space

$$V_\mathcal{T} = \left\{\mathcal{I}_\mathcal{T} f : f \in C^m(\overline{\Omega})\right\}$$

is called a C^m *finite element space*.

To obtain global continuity of the interpolant, we need to make sure that the local interpolants coincide where two element domains meet. This requires that the corresponding nodal variables are compatible. For Lagrange and Hermite elements, where each nodal variable is taken as the evaluation of a function or its derivative at a point z_i, this reduces to a geometric condition on the placement of nodes on edges. The z_i are called *nodes* (not to be confused with the *vertices* defining the element domain).

Theorem 4.11. *The triangular Lagrange and Hermite elements are all* C^0 *elements. More precisely, given a triangulation* $\mathcal{T}$ *of* Ω*, it is possible to choose edge nodes for the corresponding elements* $(K_i, \mathcal{P}_i, \mathcal{N}_i)$, $K_i \in \mathcal{T}$, *such that* $\mathcal{I}_\mathcal{T} v \in C^0(\overline{\Omega})$ *for all* $v \in C^m(\overline{\Omega})$, *where* $m = 0$ *for Lagrange and* $m = 1$ *for Hermite elements.*

Proof. It suffices to show that the global interpolant is continuous across each edge. Let K_1 and K_2 be two triangles sharing an edge e. Assume that the nodes on this edge are placed symmetrically with respect to rotation (i.e., the placement of the nodes should "look the same" from K_1 and K_2), and that $\mathcal{P}_1$ and $\mathcal{P}_2$ consist of polynomials of degree k.

Let $v \in C^m(\overline{\Omega})$ be given and set $w := \mathcal{I}_{K_1} v - \mathcal{I}_{K_2} v$, where we extend both local interpolants as polynomials outside K_1 and K_2, respectively. Hence, w is a polyomial of degree k whose restriction $w|_e$ to e is a one-dimensional polynomial having $k+1$ roots (counted by multiplicity). This implies that $w|_e = 0$, and thus the interpolant is continuous across e. □

A similar argument shows that the bilinear and biquadratic Lagrange elements are C^0 as well. Examples of C^1 elements are the Argyris triangle (of degree 5 and 21 nodal variables, including normal derivatives across edges at their midpoints, Figure 4.3a) and the Bogner–Fox–Schmit rectangle (a bicubic Hermite element of dimension 16, Figure 4.3b). It is one of the strengths of the abstract formulation described here that such exotic elements can be treated by the same tools as simple Lagrange elements.

In order to obtain global interpolation error estimates, we need uniform bounds on the local interpolation errors. For this, we need to be able to compare the local interpolation operators on different elements. This can be done with the following notion of equivalence of elements.

Definition 4.12. Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a finite element and $T : \mathbb{R}^n \to \mathbb{R}^n$ be an affine transformation, i.e., $T : v(x) \mapsto v(Ax + b)$ for $A \in \mathbb{R}^{n\times n}$ invertible and $b \in \mathbb{R}^n$. The finite element $(K, \mathcal{P}, \mathcal{N})$ is called *affine equivalent* to $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ if

(i) $K = \left\{Ax + b : x \in \hat{K}\right\}$,

(ii) $\mathcal{P} = \left\{T^{-1}(\hat{p}) : \hat{p} \in \hat{\mathcal{P}}\right\}$,

(iii) $\mathcal{N} = \left\{\{N_i\} : N_i(p) = \hat{N}_i(T(p)) \text{ for all } p \in \mathcal{P}\right\}$.

A triangulation $\mathcal{T}$ consisting of affine equivalent elements is also called *affine*.

It is a straightforward exercise to show that the dual bases of $\hat{\mathcal{P}}$ and $\mathcal{P}$ are related by $\psi_i = T^{-1}(\hat{\psi}_i)$. Hence, if the nodal variables on edges are placed symmetrically, triangular Lagrange elements of the same order are affine equivalent, as are triangular Hermite elements. The same holds true for quadratic elements. Non-affine equivalent elements are useful in treating curved boundaries, but will not be discussed here.

The advantage of this construction is that affine equivalent elements are also interpolation equivalent:

Lemma 4.13. *Let* $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ *and* $(K, \mathcal{P}, \mathcal{N})$ *be two affine equivalent finite elements related by the transformation* T. *Then,*

$$\mathcal{I}_{\hat{K}}(Tv) = T(\mathcal{I}_K v).$$

Proof. Let $\hat{\psi}_i$ and ψ_i be the dual basis of $\hat{\mathcal{P}}$ and $\mathcal{P}$, respectively. By definition,

$$\mathcal{I}_{\hat{K}}(Tv) = \sum_{i=1}^{d} \hat{N}_i(Tv)\hat{\psi}_i = \sum_{i=1}^{d} N_i(v)T(\psi_i) = T(\mathcal{I}_K v).$$

□

Given a reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$, we can thus generate a triangulation $\mathcal{T}$ using affine equivalent elements.

5 POLYNOMIAL INTERPOLATION IN SOBOLEV SPACES

We now come to the heart of the mathematical theory of finite element methods. As we have seen, the error of the finite element solution u_h is determined by the best approximation of the true solution by piecewise polynomials, which in turn is bounded by the interpolating polynomial.It thus remains to derive estimates for the (local and global) interpolation error.

5.1 THE BRAMBLE–HILBERT LEMMA

We start with the error for the local interpolant. The key for deriving error estimates is the *Bramble–Hilbert lemma* [Bramble and Hilbert 1970]. The derivation here follows the original functional-analytic arguments (by way of several results which may be of independent interest); there are also constructive approaches which allow more explicit computation of the constants.[1]

Our first lemma characterizes the kernel of differentiation operators

Lemma 5.1. *If* $v \in W^{k,p}(\Omega)$ *satisfies* $D^\alpha v = 0$ *for all* $|\alpha| = k$, *then* v *is almost everywhere equal to a polynomial of degree* $k-1$.

Proof. If $D^\alpha v = 0$ holds for all $|\alpha| = k$, then so does $D^\beta D^\alpha v = 0$ for any multi-index β. Hence, $v \in \bigcap_{k=1}^\infty W^{k,p}(\Omega)$. The Sobolev embedding theorem 2.2 thus guarantees that $v \in C^k(\Omega)$. The claim then follows using classical (pointwise) arguments. □

The next result concerns projection of Sobolev functions on polynomials.

Lemma 5.2. *For every* $v \in W^{k,p}(\Omega)$ *there is a unique polynomial* $q \in P_{k-1}$ *such that*

$$\int_\Omega D^\alpha(v-q)\,dx = 0 \qquad \text{for all } |\alpha| \leqslant k-1. \tag{5.1}$$

[1] See, e.g., [Süli 2011, § 3.2], [Brenner and Scott 2008, Chap. 4]

Proof. Writing $q = \sum_{|\beta|\leqslant k-1} \xi_\beta x^\beta \in P_{k-1}$ as a linear combination of monomials, the condition (5.1) is equivalent to the linear system

$$\sum_{|\beta|\leqslant k-1} \xi_\beta \int_\Omega D^\alpha x^\beta \,dx = \int_\Omega D^\alpha v \,dx, \qquad |\alpha| \leqslant k-1.$$

It thus remains to show that the matrix

$$\mathbf{M} = \left(\int_\Omega D^\alpha x^\beta \,dx\right)_{|\alpha|,|\beta|\leqslant k-1}$$

is non-singular. Consider $\xi = (\xi_\beta)_{|\beta|\leqslant k-1}$ such that $\mathbf{M}\xi = 0$. The corresponding polynomial q then satisfies

$$\int_\Omega D^\alpha q \,dx = 0 \qquad \text{for all } |\alpha| \leqslant k-1,$$

and hence $\xi_\beta = 0$ for all $|\beta| \leqslant k-1$. Thus, $\mathbf{M}\xi = 0$ implies $\xi = 0$, and therefore $\mathbf{M}$ is invertible. □

Using these two lemmas, we can prove a generalization of Poincaré's inequality.

Lemma 5.3. *For all* $v \in W^{k,p}(\Omega)$ *with*

$$\int_\Omega D^\alpha v \,dx = 0 \qquad \textit{for all } |\alpha| \leqslant k-1, \tag{5.2}$$

the estimate

$$\|v\|_{W^{k,p}(\Omega)} \leqslant c_0 |v|_{W^{k,p}(\Omega)} \tag{5.3}$$

holds, where the constant $c_0 > 0$ *depends only on* Ω, k *and* p.

Proof. We argue by contradiction. Assume the claim does not hold. Then there exists a sequence $\{v_n\} \subset W^{k,p}(\Omega)$ of v_n satisfying (5.2) and

$$1 = \|v_n\|_{W^{k,p}(\Omega)} \geqslant n|v_n|_{W^{k,p}(\Omega)}, \qquad n \in \mathbb{N}.$$

Since the embedding $W^{k,p}(\Omega) \subset W^{k-1,p}(\Omega)$ is compact by Theorem 2.2, there exists a subsequence (also denoted by $\{v_n\}$) converging in $W^{k-1,p}(\Omega)$ to a $v \in W^{k-1,p}(\Omega)$, i.e.,

$$\|v - v_n\|_{W^{k-1,p}(\Omega)} \to 0 \qquad \text{as } n \to \infty. \tag{5.4}$$

Since $|v_n|_{W^{k,p}(\Omega)} \to 0$ by assumption, $\{v_n\}$ is a Cauchy sequence in $W^{k,p}(\Omega)$ as well and thus converges in $W^{k,p}(\Omega)$ to a $\tilde{v} \in W^{k,p}(\Omega)$. Now, (5.4) implies that $\tilde{v} = v$ and hence $|v|_{W^{k,p}(\Omega)} = 0$. It follows from Lemma 5.1 that $v \in P_{k-1}$, and since v satisfies

$$\int_\Omega D^\alpha v \,dx = \lim_{n\to\infty} \int_\Omega D^\alpha v_n \,dx = 0 \qquad \text{for all } |\alpha| \leqslant k-1,$$

Lemma 5.2 then yields $v = 0$, in contradiction to

$$\|v\|_{W^{k,p}(\Omega)} = \lim_{n\to\infty} \|v_n\|_{W^{k,p}(\Omega)} = 1.$$

□

We are now in a position to prove our central result.

Theorem 5.4 (Bramble–Hilbert lemma). *Let* $F : W^{k,p}(\Omega) \to \mathbb{R}$ *satisfy*

(i) $|F(v)| \leqslant c_1 \|v\|_{W^{k,p}(\Omega)}$ *(boundedness),*

(ii) $|F(u+v)| \leqslant c_2(|F(u)| + |F(v)|)$ *(sublinearity),*

(iii) $F(q) = 0$ *for all* $q \in P_{k-1}$ *(annihilation).*

Then there exists a constant $c > 0$ *such that for all* $v \in W^{k,p}(\Omega)$,

$$|F(v)| \leqslant c|v|_{W^{k,p}(\Omega)}.$$

Proof. For arbitrary $v \in W^{k,p}(\Omega)$ and $q \in P_{k-1}$, we have

$$|F(v)| = |F(v-q+q)| \leqslant c_2(|F(v-q)| + |F(q)|) \leqslant c_1 c_2 \|v-q\|_{W^{k,p}(\Omega)}.$$

Given v, we now choose $q \in P_{k-1}$ as the polynomial from Lemma 5.2 and apply Lemma 5.3 to obtain

$$\|v-q\|_{W^{k,p}(\Omega)} \leqslant c_0|v-q|_{W^{k,p}(\Omega)} = c_0|v|_{W^{k,p}(\Omega)},$$

where c_0 is the constant appearing in (5.3). This proves the claim with $c := c_0c_1c_2$. □

5.2 INTERPOLATION ERROR ESTIMATES

We wish to apply the Bramble–Hilbert lemma to the interpolation error. We start with the error on the reference element.

Theorem 5.5. *Let* $(K, \mathcal{P}, \mathcal{N})$ *be a finite element,* $P_{k-1} \subset \mathcal{P}$ *for a* $k \geqslant 1$ *and* $1 \leqslant p \leqslant \infty$. *For any* $v \in W^{k,p}(K)$,

$$|v - \mathcal{I}_K v|_{W^{l,p}(K)} \leqslant c|v|_{W^{k,p}(K)} \quad \textit{for all } 0 \leqslant l \leqslant k \tag{5.5}$$

where the constant $c > 0$ *depends only on* n, k, p, l *and* $\mathcal{P}$.

Proof. It is straightforward to verify that $v \mapsto |v - \mathcal{I}_K v|_{W^{l,p}(K)}$ defines a sublinear functional on $W^{k,p}(K)$ for all $l \leqslant k$. Let $\psi_1, \dots, \psi_d$ be the dual basis $\mathcal{P}$ to $\mathcal{N}$. Since the N_i in $\mathcal{N}$ are bounded on $W^{k,p}(K)$, we have

$$\begin{aligned}
|F(v)|_{W^{l,p}(K)} &\leqslant |v|_{W^{l,p}(K)} + |\mathcal{I}_K v|_{W^{l,p}(K)} \\
&\leqslant \|v\|_{W^{k,p}(K)} + \sum_{i=1}^{d} |N_i(v)||\psi_i|_{W^{l,p}(K)} \\
&\leqslant \|v\|_{W^{k,p}(K)} + \sum_{i=1}^{d} C_i \, \|v\|_{W^{k,p}(K)} \, |\psi_i|_{W^{l,p}(K)} \\
&\leqslant (1 + C \max_{1\leqslant i\leqslant d} |\psi_i|_{W^{l,p}(K)}) \, \|v\|_{W^{k,p}(K)}
\end{aligned}$$

and hence that F is bounded. In addition, $\mathcal{I}_K q = q$ for all $q \in \mathcal{P}$ and therefore $F(q) = 0$. We can now apply the Bramble–Hilbert lemma to F, which proves the claim. □

To estimate the interpolation error on an arbitrary finite element $(K, \mathcal{P}, \mathcal{N})$, we assume that it is generated by the affine transformation

$$T_K : \hat{x} \mapsto A_K \hat{x} + b_K \tag{5.6}$$

from the reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$, i.e., $\hat{v} := v \circ T_K$ is the function v on K expressed in local coordinates on $\hat{K}$. We then need to consider how (5.5) transforms under T_K.

Lemma 5.6. *Let* $k \geqslant 0$ *and* $1 \leqslant p \leqslant \infty$. *There exists* $c > 0$ *such that for all* K *and* $w \in W^{k,p}(K)$, *the function* $\hat{w} = w \circ T_K$ *satisfies*

$$|\hat{w}|_{W^{k,p}(\hat{K})} \leqslant c \, \|A_K\|^k \, |\det(A_K)|^{-\frac{1}{p}} |w|_{W^{k,p}(K)}, \tag{5.7}$$

$$|w|_{W^{k,p}(K)} \leqslant c \, \left\|A_K^{-1}\right\|^k \, |\det(A_K)|^{\frac{1}{p}} |\hat{w}|_{W^{k,p}(\hat{K})}. \tag{5.8}$$

Proof. Let α be a multi-index with $|\alpha| = k$. Using the chain rule and the fact that T_K is affine, we obtain with a constant c depending only on n, k and p

$$\begin{aligned}
\|D^\alpha \hat{w}\|_{L^p(\hat{K})} &\leqslant c \, \|A_K\|^k \sum_{|\beta|=k} \left\|D^\beta w \circ T_K\right\|_{L^p(\hat{K})} \\
&\leqslant c \, \|A_K\|^k \, |\det(A_K)|^{-\frac{1}{p}} |w|_{W^{k,p}(K)}
\end{aligned}$$

by transformation of variables. Summing over all $|\alpha| = k$ yields (5.7). Arguing similarly using T_K^{-1} yields (5.8). □

We now derive a geometrical estimate of the quantities appearing in the right hand side of (5.7) and (5.8). For a given element domain K, we define

- the *diameter* $h_K := \max_{x_1, x_2 \in K} \|x_1 - x_2\|$,

- the *incircle diameter* $\rho_K := 2 \arg\max_r \{x \in K : B_r(x) \subset K\}$ (i.e., the diameter of the largest ball contained in K).
- the *condition number* $\sigma_K := \frac{h_K}{\rho_K}$.

Lemma 5.7. *Let* T_K *defined in* (5.6) *be an affine mapping such that* $K = T_K(\hat{K})$. *Then,*

$$|\det(A_K)| = \frac{\operatorname{vol}(K)}{\operatorname{vol}(\hat{K})}, \qquad \|A_K\| \leqslant \frac{h_K}{\rho_{\hat{K}}}, \qquad \left\|A_K^{-1}\right\| \leqslant \frac{h_{\hat{K}}}{\rho_K}.$$

Proof. The first property is a simple geometrical fact. For the second property, recall that the matrix norm of A_K is given by

$$\|A_K\| = \sup_{\|\hat{x}\|=1} \|A_K \hat{x}\| = \frac{1}{\rho_{\hat{K}}} \sup_{\|\hat{x}\|=\rho_{\hat{K}}} \|A_K \hat{x}\| .$$

Now for any $\hat{x}$ with $\|\hat{x}\| = \rho_{\hat{K}}$, there exists $\hat{x}_1, \hat{x}_2 \in \hat{K}$ with $\hat{x} = \hat{x}_1 - \hat{x}_2$ (e.g., choose $\hat{x}_2$ as the midpoint of the incircle). Then,

$$A_K \hat{x} = T_K \hat{x}_1 - T_K \hat{x}_2 = x_1 - x_2,$$

which implies $\|A_K \hat{x}\| \leqslant h_K$ and thus the desired inequality. The last property is obtained by exchanging the roles of K and $\hat{K}$. □

The local interpolation error can then be estimated as follows

Theorem 5.8 (local interpolation error). *Let* $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ *be a finite element with* $P_{k-1} \subset \hat{\mathcal{P}}$ *for a* $k \geqslant 1$. *For any element* $(K, \mathcal{P}, \mathcal{N})$ *affine equivalent to* $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ *by the affine transformation* T_K *and any* $v \in W^{k,p}(K)$, $1 \leqslant p \leqslant \infty$, *there exists* c *independent of* h_K *such that*

$$|v - \mathcal{I}_K v|_{W^{l,p}(K)} \leqslant c h_K^{k-l} \sigma_K^l |v|_{W^{k,p}(K)}$$

for all $0 \leqslant l \leqslant k$.

Proof. Let $\hat{v} := v \circ T_K$. By Lemma 4.13, $(\mathcal{I}_K v) \circ T_K = \mathcal{I}_{\hat{K}} \hat{v}$ (i.e., interpolating the transformed function is equivalent to transforming the interpolated function). Hence, we can apply Lemma 5.6 to $(v - \mathcal{I}_K v)$ and use Theorem 5.5 to obtain

$$\begin{aligned}
|v - \mathcal{I}_K v|_{W^{l,p}(K)} &\leqslant c \left\|A_K^{-1}\right\|^l |\det(A_K)|^{\frac{1}{p}} |\hat{v} - \mathcal{I}_{\hat{K}} \hat{v}|_{W^{l,p}(\hat{K})} \\
&\leqslant c \left\|A_K^{-1}\right\|^l |\det(A_K)|^{\frac{1}{p}} |\hat{v}|_{W^{k,p}(\hat{K})} \\
&\leqslant c \left\|A_K^{-1}\right\|^l \|A_K\|^k |v|_{W^{k,p}(K)} \\
&\leqslant c(\left\|A_K^{-1}\right\| \|A_K\|)^l \|A_K\|^{k-l} |v|_{W^{k,p}(K)}.
\end{aligned}$$

Using the estimates from Lemma 5.7 and the fact that $h_{\hat{K}}$ is fixed completes the proof. □

To obtain an estimate for the global interpolation error, which converges to zero as $h \to 0$, we need to have a uniform bound (independent of K and h) of the condition number σ_K. This requires a further assumption on the triangulation. A triangulation $\mathcal{T}$ is called *shape regular*, if there exists a constant κ independent of $h := \max_{K\in\mathcal{T}} h_K$ such that

$$\sigma_K \leqslant \kappa \qquad \text{for all } K \in \mathcal{T}.$$

(For triangular elements, e.g., this means that all interior angles are bounded from below.)

Using this upper bound and summing over all elements, we obtain the following estimate for the global interpolation error.

Theorem 5.9 (global interpolation error). *Let $\mathcal{T}$ be a shape regular affine triangulation of $\Omega \subset \mathbb{R}^n$ with $P_{k-1} \subset \hat{\mathcal{P}}$ for a $k \geqslant 1$. Then, there exists a constant $c > 0$ independent of h such that for all $v \in W^{k,p}(\Omega)$,*

$$\|v - \mathcal{I}_{\mathcal{T}}v\|_{L^p(\Omega)} + \sum_{l=1}^{k} h^l \left(\sum_{K\in\mathcal{T}} |v - I_K v|^p_{W^{l,p}(K)}\right)^{\frac{1}{p}} \leqslant ch^k |v|_{W^{k,p}(\Omega)}, \quad 1 \leqslant p < \infty,$$

$$\|v - \mathcal{I}_{\mathcal{T}}v\|_{L^\infty} + \sum_{l=1}^{k} h^l \max_{K\in\mathcal{T}} |v - I_K v|_{W^{l,\infty}(K)} \leqslant ch^k |v|_{W^{k,\infty}(\Omega)}.$$

Similar estimates can be obtained for elements based on tensor product polynomial spaces $\mathcal{Q}_k$.[2]

[2] e.g., [Brenner and Scott 2008, Chap. 3.5]

6 IMPLEMENTATION

This chapter discusses some of the issues involved in the implementation of the finite element method on a computer. It should only serve as a guide for solving model problems and understanding the structure of professional software packages; due to the availability of high-quality free and open source frameworks such as `deal.II`[1] and `FEniCS`[2] there is usually no need to write a finite element solver from scratch.

In the following, we focus on triangular Lagrange and Hermite elements on polygonal domains; the extension to higher-dimensional and quadrilateral elements is fairly straightforward.

6.1 TRIANGULATION

The geometric information on a triangulation is described by a *mesh*, a cloud of connected points in $\mathbb{R}^n$. This information is stored in a collection of two-dimensional arrays, the most fundamental of which are

- *the list of nodes*, which contains the coordinates $z_i = (x_i, y_i)$ for every degree of freedom:

  ```
  nodes(i) = (x_i,y_i);
  ```

- *the list of elements*, which contains for every element in the triangulation the corresponding entries in `nodes` of the nodal variables:

  ```
  elements(i) = (nodes(i_1),nodes(i_2),nodes(i_3)).
  ```

  Care must be taken that the ordering is consistent for each element. Points for which both function and gradient evaluation are given appear twice and are discerned by position in the list (function values first, then gradient).

[1] [*deal.II Differential Equations Analysis Library, Technical Reference*]

[2] [*DOLFIN: A C++/Python finite element library*]

The array `elements` serves as the local-to-global index. Depending on the boundary conditions, the following are also required.

- For Dirichlet conditions, a *list of boundary points* `bdy_nodes`.
- For Neumann conditions, a *list of boundary faces* `bdy_faces` which contain the (consistently ordered) entries in `nodes` of the nodes on each face.

The generation of a good (quasi-uniform) mesh for a given complicated domain is an active research area in itself. For uniform meshes on simple geometries (such as rectangles), it is possible to create the needed data structures by hand. An alternative are *Delaunay triangulations*, which can be constructed (e.g., by the MATLAB command `delaunay`) given a list of nodes. More complicated generators can create meshes from a geometric description of the boundary; an example is the MATLAB package `distmesh`.[3]

6.2 ASSEMBLY

The main effort in implementing lies in assembling the stiffness matrix **K**, i.e., computing its entries $K_{ij} = a(\varphi_i, \varphi_j)$ for all nodal basis element φ_i, φ_j. This is most efficiently done elementwise, where the computation is performed by transformation to a reference element.

THE REFERENCE ELEMENT We consider the reference element domain

$$\hat{K} = \left\{(\xi_1, \xi_2) \in \mathbb{R}^2 : 0 \leqslant \xi_1, \xi_2 \leqslant 1, \text{ and } \xi_1 + \xi_2 \leqslant 1\right\},$$

with the vertices $z_1 = (0,0)$, $z_2 = (1,0)$, $z_3 = (0,1)$ (in this order). For any triangle K defined by the ordered set of vertices $((x_1,y_1),(x_2,y_2),(x_3,y_3))$, the affine transformation T_K from $\hat{K}$ to K is given by

$$T_K(\xi) = A_K\xi + b_K, \quad A_K = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}, \quad b_K = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}.$$

Given a set of nodal variables $\hat{\mathcal{N}} = (\hat{N}_1, \dots, \hat{N}_d)$, it is straightforward (if tedious) to compute the corresponding nodal basis functions $\hat{\psi}_i$ from the conditions $\hat{N}_i(\hat{\psi}_j) = \delta_{ij}$, $1 \leqslant i,j \leqslant d$. (For example, the nodal basis for the linear Lagrange element is $1 - \xi_1 - \xi_2$, ξ_1, ξ_2.)

If the coefficients in the bilinear form a are constant, one can then compute the integrals on the reference element exactly, noting that due to the affine transformation, the partial derivatives of the basis functions change according to

$$\nabla\varphi|_K(x) = A_K^{-T}\nabla\hat{\psi}(\xi).$$

[3] http://persson.berkeley.edu/distmesh; an almost exhaustive list of mesh generators can be found at http://www.robertschneiders.de/meshgeneration/software.html.

QUADRATURE If the coefficients are not given analytically, it is necessary to evaluate the integrals using numerical quadrature, i.e., to compute

$$\int_K v(x)\,dx \approx \sum_{k=1}^{r} w_k v(x_k)$$

using appropriate *quadrature weights* w_k and *quadrature nodes* x_k. Since this amounts to replacing the bilinear form a by a_h (a *variational crime*), care must be taken that the discrete problem is still well-posed and that the quadrature error is negligible compared to the approximation error. It is possible to show that this can be ensured if the quadrature is sufficiently exact and the weights are positive.

Theorem 6.1 (effect of quadrature[4]). *Let $\mathcal{T}_h$ be a shape regular affine triangulation with $P_1 \subset \hat{\mathcal{P}} \subset P_k$ for $k \geqslant 1$. If the quadrature on $\hat{K}$ is of order $2k-2$ and h is small enough, the discrete problem is well-posed. If the surface integrals are approximated by a quadrature rule of order $2k-1$ and the conditions of Theorem 7.1 hold, there exists a $c > 0$ such that for $f \in H^{k-1}(\Omega)$ and $g \in H^k(\partial\Omega)$ and sufficiently small h,*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant ch^{k-1}(\|u\|_{H^k(\Omega)} + \|f\|_{H^{k-1}(\Omega)} + \|g\|_{H^k(\partial\Omega)}).$$

The rule of thumb is that the quadrature should be exact for the integrals involving second-order derivatives if the coefficients were constant. For linear elements (where the gradients are constant), order 0 (i.e., the midpoint rule) is therefore sufficient to obtain an error estimate of order h.

For higher order elements, Gauß quadrature is usually employed. This is simplified by using *barycentric coordinates*: If the vertices of K are $((x_1, y_1), (x_2, y_2), (x_3, y_3))$, the barycentric coordinates $(\zeta_1, \zeta_2, \zeta_3)$ of $(x, y) \in K$ are defined by

- $\zeta_1, \zeta_2, \zeta_3 \in [0, 1]$,
- $\zeta_1 + \zeta_2 + \zeta_3 = 1$,
- $(x, y) = \zeta_1(x_1, y_1) + \zeta_2(x_2, y_2) + \zeta_3(x_3, y_3)$.

Barycentric coordinates are invariant under affine transformations: If $\xi \in \hat{K}$ has the barycentric coordinates $(\zeta_1, \zeta_2, \zeta_3)$ with respect to the vertices of $\hat{K}$, then $x = T_K\xi$ has the same coordinates with respect to the vertices of K. The exact Gauß nodes in barycentric coordinates and the corresponding weights for quadrature of order up to 5 are given in Table 6.1. The element contributions of the local basis functions can then be computed as, e.g., in

$$\int_K \langle A(x)\nabla\varphi_i(x), \nabla\varphi_j(x)\rangle\, dx \approx \det(A_K) \sum_{k=1}^{n_l} w_k \left\langle A(x_k)A_K^{-T}\nabla\hat{\psi}_i(\xi_k), A_K^{-T}\nabla\hat{\psi}_j(\xi_k)\right\rangle,$$

[4] e.g., [Ciarlet 2002, Ths. 4.1.2, 4.1.6]

l	n_l	x_k	w_k
1	1	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$\frac{1}{2}$
2	3	$(\frac{1}{6}, \frac{1}{6}, \frac{2}{3})^\star$	$\frac{1}{6}$
3	7	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$\frac{9}{40}$
		$(\frac{1}{2}, \frac{1}{2}, 0)^\star$	$\frac{2}{30}$
		$(0, 0, 1)^\star$	$\frac{1}{40}$
5	7	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$\frac{9}{80}$
		$(\frac{6-\sqrt{15}}{21}, \frac{6-\sqrt{15}}{21}, \frac{9+2\sqrt{15}}{21})^\star$	$\frac{155-\sqrt{15}}{2400}$
		$(\frac{6+\sqrt{15}}{21}, \frac{6+\sqrt{15}}{21}, \frac{9-2\sqrt{15}}{21})^\star$	$\frac{155+\sqrt{15}}{2400}$

Table 6.1: Gauß nodes x_k (in barycentric coordinates) and weights w_k on the reference triangle. The quadrature is exact up to order l and uses n_l nodes. For starred nodes, all possible permutations appear with identical weights.

where $A = (a_{ij})_{i,j=1}^{n}$, n_l is the number of Gauss nodes, x_k and ξ_k are the Gauß nodes on the element and reference element, respectively, and $\hat{\psi}_i$, $\hat{\psi}_j$ are the basis functions on the reference element corresponding to φ_i, φ_j. The other integrals in a and F are calculated similarly.

The complete procedure for the assembly of the stiffness matrix **K** and right hand side **F** is sketched in Algorithm 6.1.

BOUNDARY CONDITIONS It remains to incorporate the boundary conditions. For Dirichlet conditions $u = g$ on $\partial\Omega$, it is most efficient to assemble the stiffness and mass matrices as above, and replace each row in **K** and entry in **F** corresponding to a node in `bdy_nodes` with the equation for the prescribed nodal value:

```
for $k = 1, \dots, N_{\mathsf{bdynodes}}$ do
    Set $K_{k,i} = 0$ for all $i$
    Set $K_{k,k} = 1$, $F_k = g(\mathsf{nodes}(k))$
end for
```

For inhomogeneous Neumann or for Robin boundary conditions, one assembles the contributions to the boundary integrals from each face similarly to Algorithm 6.1, where the loop over `elements` is replaced by a loop over `bdy_faces` (and one-dimensional Gauß quadrature is used).

Algorithmus 6.1 Finite element method for Lagrange triangles

```
Input: mesh nodes, elements, data $a_{ij}$,$b_j$,c,f
Compute Gauß nodes $\xi_l$ and weights $w_l$ on reference element
Compute values of nodal basis and gradients at Gauß nodes on reference element
Set $K_{ij} = M_{ij} = 0$
for $k = 1, \dots, N_{\mathtt{elements}}$ do
    Compute transformation $T_K$, Jacobian $\det(A_K)$ for element $K = \mathtt{elements}(k)$
    Evaluate coefficients and right hand side at transformed Gauß nodes $T_K(\xi_l)$,
    Compute $a(\psi_i, \psi_j)$, $(f, \psi_j)$ for all nodal basis elements $\psi_i, \psi_j$ using Gauß quadrature on reference element
    for $i, j = 1, \dots, d$ do
        Set $r = \mathtt{elements}(k, i)$, $s = \mathtt{elements}(k, j)$
        Set $K_{r,s} \leftarrow K_{r,s} + a(\psi_i, \psi_j)$, $F_s \leftarrow F_s + (f, \psi_j)$
    end for
end for
Output: $K_{ij}, F_j$
```

7 ERROR ESTIMATES FOR THE FINITE ELEMENT APPROXIMATION

We can now give error estimates for the conforming finite element approximation of elliptic boundary value problems for Lagrange elements. Let a reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ and a triangulation $\mathcal{T}$ using affine equivalent elements be given. Denoting the affine transformation from the reference element to the element $(K, \mathcal{P}, \mathcal{N})$ by $T_K : \hat{x} \mapsto A_K \hat{x} + b_K$, we can define the corresponding C^0 finite element space by

$$V_h := \left\{ v_h \in C^0(\overline{\Omega}) : (v_h \circ T_K)|_{\hat{K}} \in \hat{\mathcal{P}} \text{ for all } K \in \mathcal{T} \right\}.$$

7.1 A PRIORI ERROR ESTIMATES

By Céa's lemma, the discretization error is bounded by the best-approximation error, which in turn can be bounded by the interpolation error. The results of the preceding chapters therefore yield the following a priori error estimates.

Theorem 7.1. *Let $u \in H^1(\Omega)$ be the solution of the boundary value problem (2.2) together with appropriate boundary conditions. Let $\mathcal{T}$ be a shape regular affine triangulation of $\Omega \subset \mathbb{R}^n$ with $P_k \subset \hat{\mathcal{P}}$ for a $k \geqslant 1$, and let $u_h \in V_h$ be the corresponding Galerkin approximation. If $u \in H^m(\Omega)$ for $\frac{n}{2} < m < k$, there exists $c > 0$ such that*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant ch^{m-1}|u|_{H^m(\Omega)}.$$

Proof. Since $m > \frac{n}{2}$, the Sobolev embedding theorem 2.2 implies that $u \in C^0(\overline{\Omega})$ and hence the local (pointwise) interpolant is well defined. In addition, the nodal interpolation preserves Dirichlet boundary conditions. Hence $\mathcal{I}_{\mathcal{T}} u \in V_h$, and Céa's lemma yields

$$\|u - u_h\|_{H^1(\Omega)} \leqslant c \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leqslant c \, \|u - \mathcal{I}_{\mathcal{T}} u\|_{H^1(\Omega)}.$$

Theorem 5.9 for $p = 2$ and $k = m$ implies

$$\|u - \mathcal{I}_{\mathcal{T}} u\|_{H^1(\Omega)} \leqslant ch^{m-1} |u|_{H^m(\Omega)},$$

and the claim follows by combining these estimates. □

If the bilinear form a is symmetric, or if the adjoint problem to (2.2) is well-posed, we can apply the Aubin–Nitsche lemma to obtain better estimates in the L^2 norm.

Theorem 7.2. *Under the assumptions of Theorem 7.1, there exists $c > 0$ such that*

$$\|u - u_h\|_{L^2(\Omega)} \leqslant ch^m |u|_{H^m(\Omega)}.$$

Proof. By the Sobolev embedding theorem 2.2, the embedding $L^2(\Omega) \subset H^1(\Omega)$ is continuous. Thus, the Aubin–Nitsche lemma yields

$$\|u - u_h\|_{L^2(\Omega)} \leqslant c \,\|u - u_h\|_{H^1(\Omega)} \sup_{g \in L^2(\Omega)} \left(\frac{1}{\|g\|_{L^2(\Omega)}} \inf_{v_h \in V_h} \|\varphi_g - v_h\|_{H^1(\Omega)} \right)$$

where φ_g is the solution of the adjoint problem with right hand side g. Estimating the best approximation in V_h by the interpolant and using Theorem 5.9, we obtain

$$\inf_{v_h \in V_h} \|\varphi_g - v_h\|_{H^1(\Omega)} \leqslant \|\varphi_g - \mathcal{I}_{\mathcal{T}} \varphi_g\|_{H^1(\Omega)} \leqslant ch|\varphi_g|_{H^2(\Omega)} \leqslant ch \,\|g\|_{L^2(\Omega)}$$

by the well-posedness of the adjoint problem. Combining this inequality with the one from Theorem 7.1 yields the claimed estimate. □

Using duality arguments based on different adjoint problems, one can derive estimates in other $L^p(\Omega)$ spaces, including $L^\infty(\Omega)$.[1]

7.2 A POSTERIORI ERROR ESTIMATES

It is often the case that the regularity of the solution varies over the domain Ω (for example, near corners or jumps in the right hand side or coefficients). It is then advantageous to make the element size h_K small only where it is actually needed. Such information can be obtained using *a posteriori error estimates*, which can be evaluated for a computed solution u_h to decide where the mesh needs to be refined. Here, we will only sketch *residual-based* error estimates and simple *duality-based* estimates, and refer to the literature for details.[2]

[1] e.g., [Brenner and Scott 2008, Chap. 8]

[2] [Brenner and Scott 2008, Chap. 9], [Ern and Guermond 2004, Chap. 10]

For the sake of presentation, we consider a simplified boundary value problem. Let $f \in L^2(\Omega)$ and $\alpha \in L^\infty(\Omega)$ with $\alpha_1 \geqslant \alpha(x) \geqslant \alpha_0 > 0$ for almost all $x \in \Omega$ be given. Then we search for $u \in H^1_0(\Omega)$ satisfying

$$a(u, v) := (\alpha\nabla u, \nabla v) = (f, v) \qquad \text{for all } v \in H^1_0(\Omega). \tag{7.1}$$

(The same arguments can be repeated for the general boundary value problem (2.2) with homogeneous Dirichlet or Neumann conditions). Let $V_h \subset H^1_0(\Omega)$ be a finite element space and $u_h \in V_h$ the corresponding Ritz–Galerkin approximation.

RESIDUAL-BASED ERROR ESTIMATES Residual-based estimates give an error estimate in the H^1 norm. We first note that the bilinear form a is coercive with constant α_0, and hence we have

$$\begin{aligned}\alpha_0 \|u - u_h\|_{H^1(\Omega)} &\leqslant \frac{a(u - u_h, u - u_h)}{\|u - u_h\|_{H^1(\Omega)}}\\ &\leqslant \sup_{w \in H^1_0(\Omega)} \frac{a(u - u_h, w)}{\|w\|_{H^1(\Omega)}}\\ &= \sup_{w \in H^1_0(\Omega)} \frac{a(u, w) - (\alpha\nabla u_h, \nabla w)}{\|w\|_{H^1(\Omega)}}\\ &= \sup_{w \in H^1_0(\Omega)} \frac{(f, w) - \langle -\nabla \cdot (\alpha\nabla u_h), w\rangle_{H^{-1},H^1}}{\|w\|_{H^1(\Omega)}}\\ &= \sup_{w \in H^1_0(\Omega)} \frac{\langle f + \nabla \cdot (\alpha\nabla u_h), w\rangle_{H^{-1},H^1}}{\|w\|_{H^1(\Omega)}}\\ &= \|f + \nabla \cdot (\alpha\nabla u_h)\|_{H^{-1}(\Omega)}\end{aligned}$$

using integration by parts and the definition of the dual norm. For brevity, we have written $\nabla \cdot w = \sum_{j=1}^n \partial_j w_j$ for the divergence of a vector $w \in C^1(\Omega)^n$. Since all terms on the right hand side are known, this is in principle an a posteriori estimate. However, the H^{-1} norm cannot be localized, so we will perform the integration by parts on each element separately and insert an interpolation error to eliminate the H^1 norm of w (and hence the supremum).

This requires some notation. Let $\mathcal{T}_h$ be the triangulation corresponding to V_h and $\partial\mathcal{T}_h$ the set of faces of all $K \in \mathcal{T}_h$. The set of all interior faces will be denoted by Γ_h, i.e.,

$$\Gamma_h = \{F \in \partial\mathcal{T}_h : F \cap \partial\Omega = \emptyset\}.$$

For $F \in \Gamma_h$ with $F = \overline{K}_1 \cap \overline{K}_2$, let ν_1 and ν_2 denote the unit outward normal to K_1 and K_2, respectively. We define the jump in normal derivative for $w_h \in V_h$ across F as

$$[\![\nabla w_h \cdot \nu]\!] := \nabla w_h|_{K_1} \cdot \nu_1 - \nabla w_h|_{K_2} \cdot \nu_2.$$

We can then integrate by parts elementwise to obtain for $w \in H^1_0(\Omega)$

$$\begin{aligned}
a(u-u_h,w) &= (f,w) - a(u_h,w)\\
&= (f,w) - \sum_{K\in\mathcal{T}_h}\int_K \alpha\nabla(u-u_h)\cdot\nabla w\,dx\\
&= \sum_{K\in\mathcal{T}_h}\left(\int (f+\nabla\cdot(\alpha\nabla u_h))\,w\,dx - \sum_{F\in\partial K}\alpha(\nabla u_h\cdot\nu)\,w\,dx\right)\\
&= \sum_{K\in\mathcal{T}_h}\int (f+\nabla\cdot(\alpha\nabla u_h))\,w\,dx + \sum_{F\in\Gamma_h}\int_F [\![\alpha(\nabla u_h\cdot\nu)]\!]\,w\,dx
\end{aligned}$$

since $w \in H^1_0(\Omega)$ is continuous.

Our next task is to get rid of w by canceling $\|w\|_{H^1(\Omega)}$ in the definition of the dual norm. We do this by applying an interpolation error estimate. The difficulty here is that $w \in H^1_0(\Omega)$ is not sufficiently smooth to allow Lagrange interpolation, since pointwise evaluation is not defined. To circumvent this, one combines interpolation with projection. Assume $v \in V_h$ consists of piecewise polynomials of degree k. For $K \in \mathcal{T}_h$, let ω_K be the set of all elements touching K:

$$\omega_K = \bigcup\left\{\overline{K}' \in \mathcal{T}_h : \overline{K}'\cap\overline{K} \neq 0\right\}.$$

Furthermore, for every node z of K (i.e., there is $N \in \mathcal{N}$ such that $N(v) = v(z)$), denote

$$\omega_z = \bigcup\left\{K' \in \mathcal{T}_h : z \in \overline{K}'\right\} \subset \omega_K.$$

The $L^2(\omega_z)$ projection of $v \in H^1(\Omega)$ onto P_k is then defined as the unique $\pi_z(v)$ satisfying

$$\int_{\omega_z}(\pi_z(v)-v)q\,dx = 0 \qquad \text{for all } q \in P_k.$$

For $z \in \partial\Omega$, we set $\pi_z(v) = 0$ to respect the homogeneous Dirichlet conditions. The local *Clément interpolant* of $v \in H^1(\Omega)$ is then given by

$$\mathcal{I}_C v = \sum_{i=1}^{d} N_i(\pi_{z_i}(v))\varphi_i.$$

Using a variant of the Bramble–Hilbert lemma and a scaling argument, one can show the following interpolation error estimates:[3]

$$\begin{aligned}
\|v-\mathcal{I}_C v\|_{L^2(K)} &\leqslant ch_K\,\|v\|_{H^1(\omega_K)},\\
\|v-\mathcal{I}_C v\|_{L^2(F)} &\leqslant ch_K^{1/2}\,\|v\|_{H^1(\omega_K)},
\end{aligned}$$

[3] e.g., [Braess 2007, Theorem II.6.9]

for all $v \in H_0^1(\Omega)$, $K \in \mathcal{T}_h$ and $F \subset \partial K$.

Using the Galerkin orthogonality for the global Clément interpolant $\mathcal{I}_C w \in V_h$ and the fact that every K appears only in a finite number of ω_K, we thus obtain by the Cauchy–Schwarz inequality

$$\begin{aligned}
\|u-u_h\|_{H^1(\Omega)} &= \frac{1}{\alpha_0} \sup_{w\in H_0^1(\Omega)} \frac{a(u-u_h, w-\mathcal{I}_C w)}{\|w\|_{H^1(\Omega)}} \\
&\leqslant \frac{1}{\alpha_0} \sup_{w\in H_0^1(\Omega)} \frac{1}{\|w\|_{H^1(\Omega)}} \Bigg(\sum_{K\in\mathcal{T}_h} \|f+\nabla\cdot(\alpha\nabla u_h)\|_{L^2(K)} \|w-\mathcal{I}_C w\|_{L^2(K)} \\
&\qquad + \sum_{F\in\Gamma_h} \|[\![\alpha(\nabla u_h\cdot\nu)]\!]\|_{L^2(F)} \|w-\mathcal{I}_C w\|_{L^2(F)} \Bigg) \\
&\leqslant C \sup_{w\in H_0^1(\Omega)} \frac{1}{\|w\|_{H^1(\Omega)}} \Bigg(\sum_{K\in\mathcal{T}_h} h_K \|f+\nabla\cdot(\alpha\nabla u_h)\|_{L^2(K)} \|w\|_{H^1(\Omega)} \\
&\qquad + \sum_{F\in\Gamma_h} h_K^{1/2} \|[\![\alpha(\nabla u_h\cdot\nu)]\!]\|_{L^2(F)} \|w\|_{H^1(\Omega)} \Bigg) \\
&\leqslant C \Bigg(\sum_{K\in\mathcal{T}_h} h_K \|f+\nabla\cdot(\alpha\nabla u_h)\|_{L^2(K)} + \sum_{F\in\Gamma_h} h_K^{1/2} \|[\![\alpha(\nabla u_h\cdot\nu)]\!]\|_{L^2(F)} \Bigg).
\end{aligned}$$

DUALITY-BASED ERROR ESTIMATES The use of Clément interpolation can be avoided if we are satisfied with an a posteriori error estimate in the L^2 norm and $\alpha \in C^1(\Omega)$. We can then apply the Aubin–Nitsche trick. Let $w \in H_0^1(\Omega)$ solve the adjoint problem

$$a(v, w) = (u-u_h, v) \qquad \text{for all } v \in H_0^1(\Omega).$$

Inserting $u-u_h \in H_0^1(\Omega)$ and applying the Galerkin orthogonality $a(u-u_h, w_h) = 0$ for the global interpolant $w_h := \mathcal{I}_{\mathcal{T}} w$ yields

$$\begin{aligned}
\|u-u_h\|_{L^2(\Omega)}^2 &= (u-u_h, u-u_h) = a(u-u_h, w-w_h) \\
&= (f, w-w_h) - a(u_h, w-w_h).
\end{aligned}$$

Now we integrate by parts on each element again and apply the Cauchy–Schwarz inequality to obtain

$$\begin{aligned}
\|u-u_h\|_{L^2(\Omega)}^2 \leqslant &\sum_{K\in\mathcal{T}_h} \|f+\nabla\cdot(\alpha\nabla u_h)\|_{L^2(K)} \|w-w_h\|_{L^2(K)} \\
&+ \sum_{F\in\Gamma_h} \|[\![\alpha(\nabla u_h\cdot\nu)]\!]\|_{L^2(F)} \|w-w_h\|_{L^2(F)}.
\end{aligned}$$

By the symmetry of a and the well-posedness of (7.1), we have that $w \in H^2(\Omega)$ due to Theorem 2.9. We can thus estimate the local interpolation error for w using Theorem 5.8 for $k = 2$, $l = 0$ and $p = 2$ to obtain

$$\|w - w_h\|_{L^2(K)} \leqslant ch_K^2 \|w\|_{H^2(\Omega)} .$$

Similarly, using the Bramble–Hilbert lemma and a scaling argument yields

$$\|w - w_h\|_{L^2(F)} \leqslant ch_K^{3/2} \|w\|_{H^2(\Omega)} .$$

Finally, we have from Theorem 2.9 the estimate

$$\|w\|_{H^2(\Omega)} \leqslant C \|u - u_h\|_{L^2(\Omega)} .$$

Combining these inequalities, we obtain the desired a posteriori error estimate

$$\|u - u_h\|_{L^2(\Omega)} \leqslant C \left(\sum_{K \in \mathcal{T}_h} h_K^2 \|f + \nabla \cdot (\alpha \nabla u_h)\|_{L^2(K)} + \sum_{F \in \Gamma_h} h_K^{3/2} \|[\![\alpha(\nabla u_h \cdot \nu)]\!]\|_{L^2(F)} \right) .$$

Such a posteriori estimates can be used to locally decrease the mesh size in order to reduce the discretization error. This leads to *adaptive finite element methods*, which is a very active area of current research. For details, we refer to, e.g., [Brenner and Scott 2008, Chap. 9].

Part III

NONCONFORMING FINITE ELEMENTS

GENERALIZED GALERKIN APPROACH

8

The results of the preceding chapter depended on the conformality of the Galerkin approach: The discrete problem is obtained by restricting the continuous problem to suitable subspaces. This is too restrictive for many applications beyond standard second order elliptic problems, where it would be necessary to consider

- *Petrov–Galerkin* approaches: The function u satisfying $a(u, v)$ for all $v \in V$ is an element of $U \neq V$,
- *non-conformal* approaches: The discrete spaces U_h and V_h are not subspaces of U and V, respectively,
- *non-consistent* approaches: The discrete problem involves a bilinear form $a_h \neq a$ (and a_h might not be well-defined for all $u \in U$).

We thus need a more general framework which covers these cases as well. Let U, V be Banach spaces, where V is reflexive, and let U^*, V^* denote their duals. Given a bilinear form $a : U \times V \to \mathbb{R}$ and a continuous linear functional $F \in V^*$, we are looking for $u \in U$ satisfying

$$a(u, v) = F(v) \quad \text{for all } v \in V. \tag{8.1}$$

The following generalization of the Lax–Milgram theorem gives sufficient (and, as can be shown, necessary) conditions for the well-posedness of (8.1).

Theorem 8.1 (Banach–Nečas–Babuška). *Let U and V be Banach spaces and V be reflexive. Let a bilinear form $a : U \times V \to \mathbb{R}$ and a linear functional $F : V \to \mathbb{R}$ be given satisfying the following definitions:*

(i) Inf-sup-condition*: There exists a $c_1 > 0$ such that*

$$\inf_{u \in U} \sup_{v \in V} \frac{a(u, v)}{\|u\|_U \|v\|_V} \geqslant c_1.$$

(ii) Continuity: *There exist* c_2, c_3 *such that*

$$\begin{aligned}|a(u,v)| &\leqslant c_2 \|u\|_U \|v\|_V ,\\ |F(v)| &\leqslant c_3 \|v\|_V\end{aligned}$$

for all $u \in U$, $v \in V$.

(iii) Injectivity: *For every* $v \in V$, $v \neq 0$, *there is a* $u \in U$ *such that*

$$a(u,v) \neq 0.$$

Then, there exists a unique solution $u \in U$ *to* (8.1) *satisfying*

$$\|u\|_U \leqslant \frac{1}{c_1} \|F\|_{V^*} .$$

Proof. The proof is essentially an application of the closed range theorem:[1] For a bounded linear functional A between two Banach spaces X and Y, the range $A(X)$ of A is closed in Y if and only if $A(X) = (\ker A^*)^0$, where $A^* : Y^* \to X^*$ is the adjoint of A, $\ker A := \{x \in X : Ax = 0\}$ is the null space of an operator $A : X \to Y$, and for $V \subset X$,

$$V^0 := \{x \in X^* : \langle x, v\rangle_{X^*,X} = 0 \text{ for all } v \in V\}$$

is the polar of V. We apply this theorem to the operator $A : U \to V^*$ defined by

$$\langle Au, v\rangle_{V^*,V} = a(u,v) \qquad \text{for all } v \in V$$

to show that A is an isomorphism (i.e., that A is bijective and A and A^{-1} are continuous), which is equivalent to the claim since (8.1) can be expressed as $Au = f$.

Continuity of A easily follows from continuity of a and the definition of the norm on V^*. We next show injectivity of A. Let $u_1, u_2 \in U$ be given with $Au_1 = Au_2$. By definition, this implies $a(u_1, v) = a(u_2, v)$ and hence $a(u_1 - u_2, v) = 0$ for all $v \in V$. Hence, the inf-sup-condition implies that

$$c_1 \|u_1 - u_2\|_U \leqslant \sup_{v\in V} \frac{a(u_1 - u_2, v)}{\|v\|_V} = 0$$

and therefore $u_1 = u_2$.

Due to the injectivity of A, for any $f \in A(U) \subset V^*$ we have a unique $u := A^{-1}f \in U$, and the inf-sup-condition yields

$$c_1 \|u\|_U \leqslant \sup_{v\in V} \frac{a(u,v)}{\|v\|_V} = \sup_{v\in V} \frac{\langle Au, v\rangle_{V^*,V}}{\|v\|_V} = \sup_{v\in V} \frac{\langle f, v\rangle_{V^*,V}}{\|v\|_V} = \|f\|_{V^*} . \tag{8.2}$$

[1] e.g., [Zeidler 1995b, Theorem 3.E]

Therefore, A^{-1} is continuous on $A(U)$. Let $\{v_n\}_{n\in\mathbb{N}} \subset A(U)$ be a sequence converging to $v \in V^*$. Hence, there exists $u_n \in U$ such that $v_n = Au_n$, and the v_n form a Cauchy sequence. From (8.2), we deduce for all $n, m \in \mathbb{N}$ that

$$\|u_n - u_m\|_U \leqslant \frac{1}{c_1} \|Au_n - Au_m\|_{V^*},$$

which implies that $\{u_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence as well and thus converges to a $u \in U$. The continuity of A then yields

$$v = \lim_{n\to\infty} v_n = \lim_{n\to\infty} Au_n = Au,$$

and we obtain $v \in A(U)$. We can therefore apply the closed range theorem. By the reflexivity of V, we have $A^* : V \to U^*$ and

$$\begin{aligned}\operatorname{null} A^* &= \{v \in V : A^*v = 0\}\\ &= \{v \in V : \langle A^*v, u\rangle_{U^*,U} = 0 \text{ for all } u \in U\}\\ &= \{v \in V : \langle Au, v\rangle_{V^*,V} = 0 \text{ for all } u \in U\}\\ &= \{v \in V : a(u,v) = 0 \text{ for all } u \in U\}.\end{aligned}$$

Due to the injectivity condition (iii), $a(u,v) = 0$ for all $u \in U$ implies $v = 0$. Hence the closed range theorem and reflexivity of V yields

$$A(U) = (\{0\})^0 = \{x \in V^* : \langle x, 0\rangle_{V^*,V} = 0\} = V^*,$$

and therefore surjectivity of A. Thus, A is an isomorphism and the claimed estimate follows from (8.2) applied to $f \in V^*$ defined by $\langle f, v\rangle_{V^*,V} = F(v)$ for all $v \in V$. □

The term "injectivity conditions" is due to the fact that it implies injectivity of the adjoint operator A^* and hence (due to the closed range of A) surjectivity of A. Note that in the symmetric case $U = V$, coercivity of a implies both the inf-sup-condition and the injectivity condition, and we recover the Lax–Milgram lemma.

For the *non-conforming* Galerkin approach, we replace U by U_h and V by V_h, where U_h and V_h are finite-dimensional spaces, and introduce a bilinear form $a_h : U_h \times V_h \to \mathbb{R}$ and a linear functional $F_h : V_h \to \mathbb{R}$. We then search for $u_h \in U_h$ satisfying

$$a_h(u_h, v_h) = F_h(v_h) \quad \text{for all } v_h \in V_h. \tag{8.3}$$

Although we do not require $U_h \subset U$ and $V_h \subset V$, we need to have some way of comparing elements of U and U_h in order to obtain error estimates for the solution u_h. We therefore assume that there exists a subspace $\tilde{U} \subset U$ containing the exact solution such that

$$U(h) := \tilde{U} + U_h = \left\{w + w_h : w \in \tilde{U}, w_h \in U_h\right\}$$

can be endowed with a norm $\|u\|_{U(h)}$ satisfying

(i) $\|u_h\|_{U(h)} = \|u_h\|_{U_h}$ for all $u_h \in U_h$,

(ii) $\|u\|_{U(h)} \leqslant c\,\|u\|_U$ for all $u \in \tilde{U}$.

In contrast to the conformal setting, the well-posedness of (8.3) cannot be deduced from the well-posedness of (8.1), but needs to be proved independently. This is somewhat simpler due to the finite-dimensionality of the spaces.

Theorem 8.2. *Let U_h and V_h be finite-dimensional with $\dim U_h = \dim V_h$. Let a bilinear form $a_h : U_h \times V_h \to \mathbb{R}$ and a linear functional $F_h : V_h \to \mathbb{R}$ be given satisfying the following definitions:*

(i) Inf-sup-condition*: There exists a $c_1 > 0$ such that*

$$\inf_{u_h \in U_h} \sup_{v_h \in V_h} \frac{a_h(u_h, v_h)}{\|u_h\|_{U_h} \|v_h\|_{V_h}} \geqslant c_1 .$$

(ii) Continuity*: There exist c_2, c_3 such that*

$$\begin{aligned} |a_h(u_h, v_h)| &\leqslant c_2 \|u_h\|_{U_h} \|v_h\|_{V_h}, \\ |F_h(v_h)| &\leqslant c_3 \|v_h\|_{V_h} \end{aligned}$$

for all $u_h \in U_h$, $v_h \in V_h$.

Then, there exists a unique solution $u_h \in U_h$ to (8.3) satisfying

$$\|u_h\|_{U_h} \leqslant \frac{1}{c_1} \|F_h\|_{V_h^*} .$$

Proof. Consider a basis $\{\varphi_1, \dots, \varphi_n\}$ of U_h and $\{\psi_1, \dots, \psi_n\}$ of V_h and define the matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, $K_{ij} = a(\varphi_i, \psi_j)$. Then, the claim is equivalent to the invertibility of $\mathbf{K}$. From the inf-sup-condition, we obtain injectivity of $\mathbf{K}$ by arguing as in the continuous case. By the rank theorem and the condition $\dim U_h = \dim V_h$, this implies surjectivity of $\mathbf{K}$ and hence invertibility. □

Note the difference between Theorem 8.2 and the Lax–Milgram theorem in the discrete case: In the latter, the coercivity condition amounts to the assumption that the matrix $\mathbf{K}$ is positive definite, while the inf-sup-condition only requires invertibility.

The error estimates for non-conforming methods are based on the following two generalization of Céa's lemma. The first results concerns non-consistent but conformal approaches, and can be used to prove estimates for the error arising from numerical integration, see Theorem 6.1. In the following, we assume that the conditions of Theorem 8.2 hold.

Theorem 8.3 (first Strang lemma)**.** *Assume that* *(i)* $U_h \subset U = U(h)$ *and* $V_h \subset V$.

(ii) There exists a constant $c_4 > 0$ *independent of* h *such that*

$$|a(u, v_h)| \leqslant c_4 \|u\|_{U(h)} \|v_h\|_{V_h}$$

holds for all $u \in U$ *and* $v_h \in V_h$.

Then, the solutions u *and* u_h *to* (8.1) *and* (8.3) *satisfy*

$$\|u - u_h\|_{U(h)} \leqslant \frac{1}{c_1} \sup_{v_h \in V_h} \frac{|F(v_h) - F_h(v_h)|}{\|v_h\|_{V_h}} + \inf_{w_h \in U_h} \left[\left(1 + \frac{c_4}{c_1}\right) \|u - w_h\|_{U(h)} + \frac{1}{c_1} \sup_{v_h \in V_h} \frac{|a(w_h, v_h) - a_h(w_h, v_h)|}{\|v_h\|_{V_h}} \right].$$

Proof. Let $w_h \in U_h$ be given. By the discrete inf-sup-condition, we have

$$c_1 \|u_h - w_h\|_{U(h)} \leqslant \sup_{v_h \in V_h} \frac{a_h(u_h - w_h, v_h)}{\|v_h\|_{V_h}}.$$

Using (8.1) and (8.3), we can write

$$a_h(u_h - w_h, v_h) = a(u - w_h, v_h) + a(w_h, v_h) - a_h(w_h, v_h) + F_h(v_h) - F(v)$$

Inserting this into the last estimate and applying the assumption on a yields

$$c_1 \|u_h - w_h\|_{U(h)} \leqslant c_4 \|u - w_h\|_{U(h)} + \sup_{v_h \in V_h} \frac{|a(w_h, v_h) - a_h(w_h, v_h)|}{\|v_h\|_{V_h}} + \sup_{v_h \in V_h} \frac{|F(v_h) - F_h(v_h)|}{\|v_h\|_{V_h}}.$$

The claim follows after using the triangle inequality

$$\|u - u_h\|_{U(h)} \leqslant \|u - w_h\|_{U(h)} + \|u_h - w_h\|_{U(h)}$$

and taking the infimum over all $w_h \in U_h$. □

If the bilinear form a_h can be extended to $U(h) \times V_h$ (such that $a_h(u, v_h)$ makes sense), we can dispense with the assumption of conformality.

Theorem 8.4 (second Strang lemma). *Assume that there exists a constant* $c_4 > 0$ *independent of* h *such that*

$$|a_h(u, v_h)| \leqslant c_4 \|u\|_{U(h)} \|v_h\|_{V_h}$$

holds for all $u \in U(h)$ *and* $v_h \in V_h$. *Then, the solutions* u *and* u_h *to* (8.1) *and* (8.3) *satisfy*

$$\|u - u_h\|_{U(h)} \leqslant \left(1 + \frac{c_4}{c_1}\right) \inf_{w_h \in U_h} \|u - w_h\|_{U(h)} + \frac{1}{c_1} \sup_{v_h \in V_h} \frac{|F_h(v_h) - a_h(u, v_h)|}{\|v_h\|_{V_h}}.$$

Proof. Let $w_h \in U_h$ be given. Then,

$$\begin{aligned} a_h(u_h - w_h, v_h) &= a_h(u_h - u, v_h) + a_h(u - w_h, v_h) \\ &= F_h(v_h) - a_h(u, v_h) + a_h(u - w_h, v_h). \end{aligned}$$

The discrete inf-sup-condition and the assumption on a_h imply

$$c_1 \|u_h - w_h\|_{U(h)} \leqslant \sup_{v_h \in V_h} \frac{|F_h(v_h) - a_h(u - w_h, v_h)}{\|v_h\|_{V_h}} + c_4 \|u - w_h\|_{U(h)},$$

and we conclude using the triangle inequality as above. □

To illustrate the application of the first Strang lemma, we consider the effect of quadrature on the Galerkin approximation. For simplicity, we consider for $u, v \in H^1_0(\Omega)$ the continuous bilinear form

$$a(u, v) = (\alpha \nabla u, \nabla v)$$

with $\alpha \in W^{1,\infty}(\Omega) \subset C^0(\Omega)$, $\alpha_1 \geqslant \alpha(x) \geqslant \alpha_0 > 0$. Let $V_h \subset H^1_0(\Omega)$ be constructed from Lagrange elements of degree m on an affine-equivalent triangulation $\mathcal{T}_h$. The discrete bilinear form is then

$$a_h(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \sum_{k=1}^{m} w_k \alpha(x_k) \nabla u_h(x_k) \cdot \nabla v_h(x_k)$$

where w_k and x_k are the Gauß quadrature weights and nodes on each element. We recall that this formula is exact for polynomials of degree up to $2m - 1$, and that all weights are positive. Since ∇u_h is a polynomial of degree $m - 1$, this implies

$$\begin{aligned} \left(\sum_{k=1}^{m} w_k \alpha(x_k) \nabla u_h(x_k) \cdot \nabla v_h(x_k)\right)^2 &\leqslant \alpha_1^2 \left(\sum_{k=1}^{m} w_k |\nabla u_h(x_k)|^2\right) \left(\sum_{k=1}^{m} w_k |\nabla v_h(x_k)|^2\right) \\ &= \alpha_1^2 |\nabla u_h|^2_{H^1(K)} |\nabla v_h|^2_{H^1(K)} \end{aligned}$$

since the quadrature is exact for $|\nabla u_h|^2, |\nabla u_h|^2 \in P_{2m-2}$. Hence, a_h is continuous on $V_h \times V_h$:

$$|a_h(u_h, v_h)| \leqslant C \|u_h\|_{H^1(\Omega)} \|v_h\|_{H^1(\Omega)}.$$

Similarly, a_h is coercive:

$$\begin{aligned} a_h(u_h, u_h) &\geqslant \alpha_0 \sum_{K \in \mathcal{T}_h} \sum_{k=1}^{m} w_k |\nabla u_h(x_k)|^2 = \alpha_0 |u_h|^2_{H^1(\Omega)} \\ &\geqslant C \|u_h\|^2_{H^1(\Omega)} \end{aligned}$$

by Poincaré's inequality (Theorem 2.5). Thus, the discrete problem is well-posed by Theorem 8.2.

We next derive error estimates. Using the first Strang lemma, we find that the discretization error is bounded by the approximation error and the quadrature error. For $m = 1$ (linear Lagrange elements), Theorem 5.9 yields

$$\inf_{w_h \in V_h} \|u - w_h\|_{H^1(\Omega)} \leqslant Ch|u|_{H^2(\Omega)}.$$

For the quadrature error in the bilinear form, we use that for $w_h, v_h \in V_h$, ∇w_h and ∇v_h are constant on each element to write

$$\begin{aligned} |a(w_h, v_h) - a_h(w_h, v_h)| &= \sum_{K \in \mathcal{T}_h} \left(\int_K \alpha \nabla w_h \cdot \nabla v_h \, dx - \sum_{k=1}^m w_k \alpha(x_k) \nabla w_h(x_k) \cdot \nabla v_h(x_k) \right) \\ &= \sum_{K \in \mathcal{T}_h} \nabla w_h \cdot \nabla v_h \left(\int_K \alpha \, dx - \sum_{k=1}^m w_k \alpha(x_k) \right). \end{aligned}$$

Since

$$E_K(v) := \int_K v(x) \, dx - \sum_{k=1}^m w_k v(x_k)$$

is a bounded, sublinear functional on $W^{m,\infty}(K)$ which vanishes for all $v \in P_{m-1} \subset P_{2m-1}$, we can apply the Bramble–Hilbert lemma on the reference element $\hat{K}$ to obtain

$$|E_{\hat{K}}(\hat{v})| \leqslant C|\hat{v}|_{W^{m,\infty}(\hat{K})}.$$

A scaling argument then yields

$$|E_K(v)| \leqslant Ch_K^m \operatorname{vol}(K) \, |v|_{W^{m,\infty}(K)}.$$

Inserting this and using that $\nabla u_h, \nabla v_h$ are constant on each element, we obtain

$$\begin{aligned} |a(w_h, v_h) - a_h(w_h, v_h)| &= \sum_{K \in \mathcal{T}_h} \nabla w_h \cdot \nabla v_h E_K(\alpha) \\ &\leqslant C \sum_{K \in \mathcal{T}_h} h_K |\alpha|_{W^{1,\infty}(K)} \int_K \nabla w_h \cdot \nabla v_h \, dx \\ &\leqslant Ch|\alpha|_{W^{1,\infty}(\Omega)} \|w_h\|_{H^1(\Omega)} \|v_h\|_{H^1(\Omega)}. \end{aligned}$$

For the quadrature error on the right hand side $F_h(v_h)$, we proceed similarly (applying the Bramble–Hilbert lemma to fv_h and using the product rule and equivalence of norms on V_h) to obtain

$$|F(v_h) - F_h(v_h)| \leqslant Ch \, |f|_{W^{1,\infty}(\Omega)} \|v_h\|_{H^1(\Omega)}.$$

Combining these estimates with the first Strang lemma yields

$$\|u - u_h\|_{H^1(\Omega)} \leqslant Ch \left(|f|_{W^{1,\infty}(\Omega)} + |u|_{H^2(\Omega)} \right),$$

where we have used that $\inf_{w_h \in V_h} |\alpha|_{W^{1,\infty}(\Omega)} \|w_h\|_{V_h} = 0$.

DISCONTINUOUS GALERKIN METHODS

9

Discontinuous Galerkin methods are based on nonconforming finite element spaces consisting of piecewise polynomials that are not continuous across elements. This allows them to handle irregular meshes with hanging nodes and different degrees of polynomials on each element. They also provide a natural framework for first order partial differential equations and for imposing Dirichlet boundary conditions in a weak form, on which we will focus here. We consider a simple *advection-reaction* equation

$$\beta \cdot \nabla u + \mu u = f$$

which models the transport of a solute concentration u along the vector field β. The reaction coefficient μ determines the rate with which the solute is destroyed or created due to interaction with its environment, and f is a source term. This is complemented by (for simplicity) homogeneous Dirichlet conditions which will be specified below.

9.1 WEAK FORMULATION OF ADVECTION-REACTION EQUATIONS

We consider $\Omega \subset \mathbb{R}^n$ (polyhedral) with unit outer normal ν and assume

$$\mu \in L^\infty(\Omega), \qquad \beta \in (W^{1,\infty}(\Omega))^n, \qquad f \in L^2(\Omega).$$

Our first task is to define the space in which we look for our solution. Let

$$\partial\Omega^- = \{x \in \partial\Omega : \beta(x) \cdot \nu(x) < 0\}$$

denote the *inflow boundary* and

$$\partial\Omega^+ = \{x \in \partial\Omega : \beta(x) \cdot \nu(x) > 0\}$$

the *outflow boundary*, and assume that they are well-separated:

$$\min_{x \in \partial\Omega^-, y \in \partial\Omega^+} |x - y| > 0.$$

Then we define the so-called *graph space*

$$W = \left\{ v \in L^2(\Omega) : \beta \cdot \nabla v \in L^2(\Omega) \right\} \subset L^2(\Omega),$$

which is a Hilbert space if endowed with the inner product

$$\langle v, w \rangle_W = (v, w) + (\beta \cdot \nabla v, \beta \cdot \nabla w).$$

The latter induces the *graph norm*

$$\|v\|_W = \left(\langle v, v \rangle_W\right)^{\frac{1}{2}}.$$

One can show[1] that functions in W have traces in the space

$$L^2_\beta(\partial\Omega) = \left\{ v \text{ measurable on } \partial\Omega : \int_{\partial\Omega} |\beta \cdot \nu| v^2 \, dx < \infty \right\},$$

and that the following integration by parts formula holds:

$$\int_\Omega (\beta \cdot \nabla v) w + (\beta \cdot \nabla w) v + (\nabla \cdot \beta) v w \, dx = \int_{\partial\Omega} (\beta \cdot \nu) v w \, dx \tag{9.1}$$

for all $v, w \in W$.

We can now define our weak formulation: Set

$$U := \{ v \in W : v|_{\partial\Omega^-} = 0 \}$$

and find $u \in V$ satisfying

$$a(u, v) := (\beta \cdot \nabla u, v) + (\mu u, v) = (f, v) \tag{9.2}$$

for all $v \in L^2(\Omega)$. Note that the test space is now different from the solution space.

Since U is a closed subspace of the Hilbert space W, it is a Banach space. Moreover, $L^2(\Omega)$ is a reflexive Banach space and the right hand side defines a continuous linear functional on $L^2(\Omega)$. We can thus apply the Banach–Nečas–Babuška Theorem to show well-posedness.

Theorem 9.1. *If*

$$\mu(x) - \tfrac{1}{2}\nabla \cdot \beta(x) \geqslant \mu_0 > 0 \quad \textit{for almost all } x \in \Omega$$

holds, there exists a unique $u \in U$ satisfying (9.2). Furthermore, there exists a $c > 0$ such that

$$\|u\|_W \leqslant c \, \|f\|_{L^2(\Omega)}$$

holds.

[1] e.g., [Di Pietro and Ern 2012, Lemma 2.5]

Proof. We begin by showing the continuity of a on $U \times L^2(\Omega)$. For arbitrary $u \in U$ and $v \in L^2(\Omega)$, the Cauchy–Schwarz inequality yields

$$\begin{aligned}|a(u,v)| &\leqslant \|\beta\cdot\nabla u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|\mu v\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)}\\ &\leqslant (1+\|\mu\|_{L^\infty(\Omega)})\|u\|_W\|v\|_{L^2(\Omega)}.\end{aligned}$$

To verify the inf-sup-condition, we first prove coercivity in $L^2(\Omega)$. For any $u \in U$, we integrate by parts using (9.1) for $v = w = u$ to obtain

$$\begin{aligned}a(u,u) &= \int_\Omega (\beta\cdot\nabla u)u + \mu u^2\,dx\\ &= \int_\Omega (\mu - \tfrac{1}{2}\nabla\cdot\beta)u^2\,dx + \int_{\partial\Omega}\tfrac{1}{2}(\beta\cdot\nu)u^2\,dx\\ &\geqslant \mu_0\|u\|^2_{L^2(\Omega)},\end{aligned}$$

where we have used that u vanishes on $\partial\Omega^-$ due to the boundary conditions and that $\beta\cdot\nu > 0$ on $\partial\Omega^+$. This implies

$$\|u\|_{L^2(\Omega)} \leqslant \mu_0^{-1}\frac{a(u,u)}{\|u\|_{L^2(\Omega)}} \leqslant \sup_{w\in L^2(\Omega)}\mu_0^{-1}\frac{a(u,w)}{\|w\|_{L^2(\Omega)}}.$$

For the other term in the graph norm, we use the duality trick

$$\begin{aligned}\|\beta\cdot\nabla u\|_{L^2(\Omega)} &= \sup_{w\in L^2(\Omega)}\frac{(\beta\cdot\nabla u, w)}{\|w\|_{L^2(\Omega)}}\\ &= \sup_{w\in L^2(\Omega)}\frac{a(u,w)-(\mu u,w)}{\|w\|_{L^2(\Omega)}}\\ &\leqslant \sup_{w\in L^2(\Omega)}\frac{a(u,w)}{\|w\|_{L^2(\Omega)}} + \|\mu\|_{L^\infty(\Omega)}\|u\|_{L^2(\Omega)}\\ &\leqslant (1+\mu_0^{-1}\|\mu\|_{L^\infty(\Omega)})\sup_{w\in L^2(\Omega)}\frac{a(u,w)}{\|w\|_{L^2(\Omega)}}.\end{aligned}$$

Summing the last two inequalities and taking the infimum over all $u \in V$ verifies the inf-sup-condition.

For the injectivity condition, we assume that $v \in L^2(\Omega)$ is such that $a(u,v) = 0$ for all $u \in U$ and show that $v = 0$. Since $C_0^\infty(\overline{\Omega}) \subset U$, we deduce that $\nabla\cdot(\beta v)$ exists as a weak derivative and $\mu v - \nabla\cdot(\beta v) = 0$. By the chain rule, we furthermore have $\beta\cdot\nabla v = (\mu - \nabla\cdot\beta)v \in L^2(\Omega)$, which implies $v \in W$. Inserting this into the integration by parts formula (9.1) and adding the productive zero yields for all $u \in U$

$$\begin{aligned}\int_{\partial\Omega}(\beta\cdot\nu)uv\,dx &= \int_\Omega (\beta\cdot\nabla v)u + (\beta\cdot\nabla u)v + (\nabla\cdot\beta)vu\,dx\\ &= a(u,v) - ((\mu-\nabla\cdot\beta)v,u) + (\beta\cdot\nabla v,u)\\ &= 0.\end{aligned}\tag{9.3}$$

Since $\partial\Omega^+$ and $\partial\Omega^-$ are well separated, there exists a smooth cut-off function χ with $\chi(x) = 0$ for $x \in \partial\Omega^-$. Applying (9.3) to $u = \chi v \in U$ yields $\int_{\partial\Omega^+}(\beta \cdot \nu)v^2\,dx = 0$. Using again $\mu v - \nabla\cdot(\beta v) = 0$ and integrating by parts, we deduce that

$$\begin{aligned} 0 &= \int_\Omega \mu v^2 - \nabla\cdot(\beta v)v\,dx \\ &= \int_\Omega (\mu - \tfrac{1}{2}\nabla\cdot\beta)v^2\,dx - \int_{\partial\Omega}\tfrac{1}{2}(\beta\cdot\nu)v^2 \\ &\geqslant \mu_0\,\|v\|_{L^2(\Omega)} \end{aligned}$$

since the remaining boundary integral over $\partial\Omega^-$ is non-negative. This shows that $v = 0$, from which the injectivity condition follows by contraposition. □

Note that the graph norm is the strongest norm in which we could have shown coercivity, and that a would not have been bounded on $U \times U$.

9.2 GALERKIN APPROACH

The *discontinuous Galerkin* approach now consists in choosing our discrete spaces as

$$V_h = \left\{v \in L^2(\Omega) : v|_K \in P_k, K \in \mathcal{T}_h\right\}$$

for $k \geqslant 0$ and a given triangulation $\mathcal{T}_h$ of Ω (no continuity across elements is assumed, hence the name). We then search for $u_h \in V_h$ satisfying

$$a_h(u_h, v_h) = (f, v_h) \qquad \text{for all } v_h \in V_h \tag{9.4}$$

for a bilinear form a_h to be specified. Here, we consider the simplest choice that leads to a convergent scheme. Recall that the set of interior faces of $\mathcal{T}_h$ is denoted by Γ_h. Let $F \in \Gamma_h$ be the face common to the elements $K_1, K_2 \in \mathcal{T}_h$ with exterior normal ν_1 and ν_2, respectively. For a function $u \in L^2(\Omega)$, we denote the *jump* across F as

$$[\![u]\!]_F = u|_{K_1}\nu_1 + u|_{K_2}\nu_2$$

and the *average* as

$$\{\!\{u\}\!\}_F = \tfrac{1}{2}(u|_{K_1} + u|_{K_2}).$$

We will omit the subscript F if it is clear which face is meant. It is also convenient to introduce for $v_h \in V_h$ the *broken gradient* $\nabla_h v_h$ via

$$(\nabla_h v_h)|_K = \nabla(v_h|_K) \qquad \text{for all } K \in \mathcal{T}_h.$$

We then define the bilinear form

$$a_h(u_h,v_h) = (\mu u_h + \beta\cdot\nabla_h u_h, v_h) + \int_{\partial\Omega^-}(\beta\cdot\nu)u_h v_h\,dx \\ - \sum_{F\in\Gamma_h}\int_F \beta\cdot[\![u_h]\!]\{\!\{v_h\}\!\}. \tag{9.5}$$

The second term enforces the homogeneous Dirichlet conditions in a weak sense. The last term can be thought of as weakly enforcing continuity by penalizing the jump across each face; the reason for its specific form will become apparent during the following. Continuity of a on $V_h\times V_h$ will be shown later. To prove well-posedness of (9.4), it remains to verify the discrete inf-sup-condition, which we can do by showing coercitivity in an appropriate norm. We choose

$$|\!|\!|u_h|\!|\!|^2 = \mu_0\,\|u_h\|^2_{L^2(\Omega)} + \int_{\partial\Omega}\tfrac12|\beta\cdot\nu|u_h^2\,dx,$$

which is clearly a norm on $V_h\subset L^2(\Omega)$. We begin by integrating by parts on each element the first term of (9.5) for $v_h = u_h$:

$$\begin{aligned}(\mu u_h + \beta\cdot\nabla_h u_h, u_h) &= \sum_{K\in\mathcal{T}_h}\int_K \mu u_h^2 + (\beta\cdot\nabla u_h)u_h\,dx\\ &= \sum_{K\in\mathcal{T}_h}\int_K \mu u_h^2 - \tfrac12(\nabla\cdot\beta)u_h^2\,dx + \int_{\partial K}\tfrac12(\beta\cdot\nu)u_h^2\,dx.\end{aligned}$$

The last term can be reformulated as a sum over faces. Since $\beta\in W^{1,\infty}(\Omega)$ is continuous, we have

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K}\tfrac12(\beta\cdot\nu)u_h^2 = \sum_{F\in\Gamma_h}\int_F\tfrac12\beta\cdot[\![u_h^2]\!] + \sum_{F\in\partial\mathcal{T}_h\setminus\Gamma_h}\int_F\tfrac12(\beta\cdot\nu)u_h^2.$$

Using

$$\tfrac12[\![w^2]\!]_F = \tfrac12(w|^2_{K_1} - w|^2_{K_2})\nu = \tfrac12(w|_{K_1}+w|_{K_2})(w|_{K_1}-w|_{K_2})\nu = \{\!\{w\}\!\}_F\,[\![w]\!]_F\,,$$

and combining the terms involving integrals over $\partial\Omega$, we obtain

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K}\tfrac12(\beta\cdot\nu)u_h^2\,dx + \int_{\partial\Omega^-}(\beta\cdot\nu)u_h^2\,dx = \sum_{F\in\Gamma_h}\int_F\beta\cdot[\![u_h]\!]\{\!\{u_h\}\!\} + \int_{\partial\Omega}\tfrac12|\beta\cdot\nu|u_h^2\,dx.$$

Note that we have no control over the sign of the first term on the right hand side, which is why we had to introduce the penalty term in a_h to cancel it. Combining these equations yields

$$\begin{aligned}a_h(u_h,u_h) &= \sum_{K\in\mathcal{T}_h}\int_K\left(\mu - \tfrac12(\nabla\cdot\beta)\right)u_h^2\,dx + \int_{\partial\Omega}\tfrac12|\beta\cdot\nu|u_h^2\,dx\\ &\geqslant \mu_0\,\|u_h\|^2_{L^2(\Omega)} + \int_{\partial\Omega}\tfrac12|\beta\cdot\nu|u_h^2\,dx\\ &= |\!|\!|u_h|\!|\!|^2.\end{aligned}$$

Hence, a_h is coercive on V_h, and by Theorem 8.2, there exists a unique solution $u_h \in V_h$ to (9.4).

9.3 ERROR ESTIMATES

To derive error estimates for the discontinuous Galerkin approximation $u_h \in V_h$ to $u \in U$, we wish to apply the second Strang lemma. Our first task is to show boundedness of a_h on a sufficiently large space containing the exact solution. Since the corresponding norm will involve traces on edges, we make the additional assumption that the exact solution satisfies

$$u \in U_* := U \cap H^1(\Omega).$$

By the trace theorem 2.4, $u|_F$ is well-defined in the sense of $L^2(F)$ traces. We then define on $U(h) := U_* + V_h$ the norm

$$|||w|||_*^2 := |||w|||^2 + \sum_{K\in\mathcal{T}_h} \left(\|\beta\cdot\nabla w\|_{L^2(K)}^2 + h_K^{-1} \|w\|_{L^2(\partial K)}^2 \right).$$

We can then show boundedness of a_h:

Lemma 9.2. *There exists a constant $C > 0$ independent of h such that for all $u \in U(h)$ and $v_h \in V_h$,*

$$a_h(u, v_h) \leqslant C\, |||u|||_*\, |||v_h|||$$

holds.

Proof. Using the Cauchy–Schwarz inequality and some generous upper bounds, we immediately obtain

$$(\mu u + \beta\nabla u, v_h) + \int_{\partial\Omega^-} (\beta\cdot\nu) u v_h \, dx \leqslant C\, |||u|||_*\, |||v_h||| , \tag{9.6}$$

with a constant $C > 0$ depending only on μ. For the last term of $a_h(u, v_h)$, we also apply the Cauchy–Schwarz inequality:

$$\sum_{F\in\Gamma_h} \int_F \beta\cdot [\![u]\!]\, \{\!\{v_h\}\!\} \leqslant C \left(\sum_{F\in\Gamma_h} \tfrac{1}{2} \{\!\{h\}\!\}^{-1} \|[\![u]\!]\|_{L^2(F)}^2 \right)^{\frac{1}{2}} \left(\sum_{F\in\Gamma_h} 2\{\!\{h\}\!\} \|\{\!\{v_h\}\!\}\|_{L^2(F)}^2 \right)^{\frac{1}{2}},$$

where $C > 0$ depends only on β. Now we use that

$$\tfrac{1}{2}\, [\![w]\!]_F^2 \leqslant (w|_{K_1}^2 + w|_{K_2}^2), \qquad 2\{\!\{w\}\!\}^2 \leqslant (w|_{K_1}^2 + w|_{K_2}^2)$$

holds, and that for a shape-regular mesh, the element size h_K cannot change arbitrarily between neighboring elements, i.e., there exists a $c > 0$ such that

$$c^{-1}\max(h_{K_1}, h_{K_2}) \leqslant \{\!\{h\}\!\} \leqslant c\min(h_{K_1}, h_{K_2}).$$

This implies

$$\begin{aligned}\text{(9.7)}\qquad \sum_{F\in\Gamma_h}\int_F \beta\cdot [\![u]\!]\,\{\!\{v_h\}\!\} &\leqslant C\left(\sum_{K\in\mathcal{T}_h} h_K^{-1}\,\|u\|^2_{L^2(\partial K)}\right)^{\frac12}\left(\sum_{K\in\mathcal{T}_h} h_K\,\|v_h\|^2_{L^2(\partial K)}\right)^{\frac12}\\ &\leqslant C\,|\!|\!|u|\!|\!|_*\,|\!|\!|v_h|\!|\!|\end{aligned}$$

where we have combined the terms arising from the faces of each element and applied the trace theorem with a scaling argument:

$$h_K^{1/2}\,\|v_h\|_{L^2(\partial K)} \leqslant \|v_h\|_{L^2(K)}\,.$$

Adding (9.6) and (9.7) yields the claim. □

Since $|\!|\!|\cdot|\!|\!|$ and $|\!|\!|\cdot|\!|\!|_*$ are equivalent norms on the (finite-dimensional) space V_h, Lemma 9.2 fills the remaining gap in the well-posedness of (9.4).

We now argue consistency of our discontinuous Galerkin approximation.

Lemma 9.3. *The solution* $u \in U_*$ *to* (9.2) *satisfies*

$$a_h(u, v_h) = (f, v_h)$$

for all $v_h \in V_h$.

Proof. By definition, $u \in U_*$ satisfies $a(u, v_h) = (f, v_h)$ for all $v_h \in V_h$. Furthermore, due to the boundary conditions,

$$\int_{\partial\Omega^-}(\beta\cdot\nu)uv_h\,dx = 0.$$

It remains to show that the penalty term $(\beta\cdot\nu)\,[\![u_h]\!]_F\,\{\!\{v_h\}\!\}_F$ vanishes on each face $F \in \Gamma_h$. Let $\varphi \in C_0^\infty(\overline{\Omega})$ have support contained in $S \subset \overline{K}_1 \cup \overline{K}_2 \subset \Omega$ and intersecting $F = \partial K_1 \cap \partial K_2$. Then the integration by parts formula (9.1) gives

$$\begin{aligned}0 &= \int_\Omega (\beta\cdot\nabla v)\varphi + (\beta\cdot\nabla\varphi)v + (\nabla\cdot\beta)v\varphi\,dx\\ &= \int_{S\cap K_1}(\beta\cdot\nabla v)\varphi + (\beta\cdot\nabla\varphi)v + (\nabla\cdot\beta)v\varphi\,dx\\ &\quad+\int_{S\cap K_2}(\beta\cdot\nabla v)\varphi + (\beta\cdot\nabla\varphi)v + (\nabla\cdot\beta)v\varphi\,dx\\ &= \int_{\partial K_1\cap S}(\beta\cdot\nu)v\varphi\,dx + \int_{\partial K_2\cap S}(\beta\cdot\nu)v\varphi\,dx\\ &= \int_F \beta\cdot[\![v]\!]\,\varphi\,dx.\end{aligned}$$

The claim then follows from a density argument. □

This implies that the consistency error is zero, and we are left with the approximation error in the $U(h)$ norm, which can be estimated using the local interpolant if the exact solution is smooth enough. From the second Strang lemma, we thus obtain the following error estimate.

Theorem 9.4. *Assume that the solution* $u \in U(h)$ *to* (9.2) *satisfies* $u \in H^{k+1}(\Omega)$. *Then there exists a* $c > 0$ *independent of* h *such that*

$$|||u - u_h||| \leqslant ch^k \|u\|_{H^{k+1}(\Omega)}$$

holds.

Note that since we could only show coercivity with respect to $|||\cdot|||$ (and $u - u_h$ is not in a finite-dimensional space), we only get an error estimate in this (weaker) norm of L^2 type, while the approximation error needs to be estimated in the (stronger) H^1-type norm $|||\cdot|||_*$. On the other hand, we would expect a convergence order $h^{k+1/2}$ for the discretization error in an L^2-type norm (involving interface terms). This discrepancy is due to the simple penalty we added, which is insufficient to control oscillations (it cancelled the interface terms arising in the integration by parts, but did not contribute further in the coercivity). A more stable alternative is *upwinding*: Take

$$a_h^+(u_h, v_h) = a_h(u_h, v_h) + \sum_{F \in \Gamma_h} \int_F \frac{\eta}{2} |\beta \cdot \nu| \, [\![u_h]\!] \cdot [\![v_h]\!]$$

for a sufficiently large penalty parameter $\eta > 0$. It can be shown[2] that this bilinear form is consistent as well, and is coercive in the norm

$$|||w|||_+^2 = |||w|||^2 + \sum_{F \in \Gamma_h} \int_F \frac{\eta}{2} |\beta \cdot \nu| \, [\![u_h]\!]^2 + \sum_{K \in \mathcal{T}_h} h_K \|\beta \cdot \nabla w\|_{L^2(K)}^2$$

and continuous in

$$|||w|||_{+,*}^2 = |||w|||_+^2 + \sum_{K \in \mathcal{T}_h} \left(h_K^{-1} \|w\|_{L^2(K)}^2 + \|w\|_{L^2(\partial K)}^2 \right),$$

which can be used to obtain the expected convergence order of $h^{k+1/2}$.

[2] e.g., [Di Pietro and Ern 2012, Chapter 2.3]

MIXED METHODS

10

We now consider variational problems with constraints. Such problems arise, e.g., in the variational formulation of incompressible flow problems (where incompressibility of the solution u can be expressed as the condition $\nabla \cdot u = 0$) or when explicitly enforcing boundary conditions in the weak formulation. To motivate the general problem we will study in this chapter, consider two reflexive Banach spaces V and M and the symmetric and coercive bilinear form $a : V \times V \to \mathbb{R}$. We know (cf. Theorem 3.3) that the solution $u \in V$ to $a(u, v) = 0$ for all $v \in V$ is the unique minimizer of $J(v) = \frac{1}{2}a(v, v)$. If we want u to satisfy the additional condition $b(u, \mu) = 0$ for all $\mu \in M$ and a bilinear form $b : V \times M \to \mathbb{R}$ (e.g., $b(u, \mu) = (\nabla \cdot u, \mu)$), we can introduce the Lagrangian

$$L(u, \lambda) = J(u) + b(u, \lambda)$$

and consider the saddle point problem

$$\inf_{v \in V} \sup_{\mu \in M} L(v, \mu).$$

Taking the derivative with respect to v and μ, we obtain the (formal) first order optimality conditions for the saddle point $(u, \lambda) \in V \times M$:

$$\begin{cases} a(u, v) + b(v, \lambda) = 0 & \text{for all } v \in V, \\ \qquad\qquad b(u, \mu) = 0 & \text{for all } \mu \in M \end{cases}$$

This can be made rigorous; the existence of a Lagrange multiplier λ however requires some assumptions on b. In the next section, we will see that these can be expressed in the form of an inf-sup condition.

10.1 ABSTRACT SADDLE POINT PROBLEMS

Let V and M be two reflexive Banach spaces,

$$a : V \times V \to \mathbb{R}, \qquad b : V \times M \to \mathbb{R}$$

be two continuous (not necessarily symmetric) bilinear forms, and $f \in V^*$ and $g \in M^*$ be given. Then we search for $(u, \lambda) \in V \times M$ satisfying the saddle point problem

$$(\mathcal{S}) \qquad \begin{cases} a(u, v) + b(v, \lambda) = \langle f, v\rangle_{V^*,V} & \text{for all } v \in V, \\ b(u, \mu) = \langle g, \mu\rangle_{M^*,M} & \text{for all } \mu \in M. \end{cases}$$

In principle, we can obtain existence and uniqueness of (u, λ) by considering ($\mathcal{S}$) as a bilinear form on $V \times M$ and verifying a suitable inf-sup condition. It is, however, more convenient to express this condition in terms of the original bilinear forms a and b. For this purpose, we first reformulate ($\mathcal{S}$) as an operator equation by introducing the operators

$$\begin{aligned} &A : V \to V^*, && \langle Au, v\rangle_{V^*,V} = a(u, v) && \text{for all } v \in V,\\ &B : V \to M^*, && \langle Bu, \mu\rangle_{M^*,M} = b(u, \mu) && \text{for all } \mu \in M,\\ &B^* : M \to V^*, && \langle B^*\lambda, v\rangle_{V^*,V} = b(v, \lambda) && \text{for all } v \in V. \end{aligned}$$

Then, ($\mathcal{S}$) is equivalent to

$$(10.1) \qquad \begin{cases} Au + B^*\lambda = f,\\ Bu = g. \end{cases}$$

From this, we can see the following: If B were invertible, the existence and uniqueness of (u, λ) would follow immediately. In the (more realistic case) that B has a nontrivial null space

$$\ker B = \{x \in V : b(x, \mu) = 0 \text{ for all } \mu \in M\},$$

we have to require that A is injective on it to obtain a unique u. Existence of λ then follows from surjectivity of B^*. To verify these conditions, we follow the general approach of the Banach–Nečas–Babuška theorem.

Theorem 10.1 (Brezzi splitting theorem). *Assume*

(i) $a : V \times V \to \mathbb{R}$ *satisfies the conditions of Theorem 8.1 for* $U = V = \ker B$,

(ii) $b : V \times M \to \mathbb{R}$ *satisfies for* $\beta > 0$ *the inf-sup condition*

$$(10.2) \qquad \inf_{\mu \in M} \sup_{v \in V} \frac{b(v, \mu)}{\|v\|_V \|\mu\|_M} \geqslant \beta.$$

Then, there exists a unique solution $(u, \lambda) \in V \times M$ *to* (10.1) *satisfying*

$$\|u\|_V + \|\lambda\|_M \leqslant C(\|f\|_{V^*} + \|g\|_{M^*}).$$

Condition (ii) is known as the *Ladyžhenskaya–Babuška–Brezzi* (LBB) condition. Note that a only has to satisfy an inf-sup condition on the null space of B, not on all of V, which is crucial in many applications.

Proof. First, by inspecting the proof of Theorem 8.1, we note that the LBB condition implies that B^* has closed range, is injective on M, and is surjective on

$$(\ker B)^0 = \{x \in V^* : \langle x, v\rangle_{V^*,V} = 0 \text{ for all } v \in \ker B\}.$$

In addition,

$$\beta \|\mu\|_M \leqslant \|B^*\mu\|_{V^*}$$

holds for all $\mu \in M$. By reflexivity of V, M and the closed range theorem, $B = (B^*)^*$ has closed range as well and hence is surjective on $(\ker B^*)^0 = (\{0\})^0 = M^*$. For all $x \in V$, we have

$$\begin{aligned}\|Bx\|_{M^*} &= \sup_{\mu\in M} \frac{\langle Bx, \mu\rangle_{M^*,M}}{\|\mu\|_M} = \sup_{\mu\in M} \frac{b(x,\mu)}{\|\mu\|_M}\\ &\geqslant \inf_{\mu\in M} \frac{b(x,\mu)}{\|\mu\|_M} \geqslant \beta \|x\|_V\end{aligned}$$

by the LBB condition. Hence, for $g \in M^*$, there exists a $u_g \in V$ satisfying $Bu_g = g$ and

$$\|u_g\|_V \leqslant \frac{1}{\beta} \|g\|_{M^*}. \tag{10.3}$$

Due to condition (i), A is an isomorphism on $\ker B$. Considering $f - Au_g$ as a bounded linear form on $\ker B \subset V$, we thus obtain a unique $u_f \in \ker B$ satisfying $Au_f = f - Au_g$ and

$$\|u_f\|_V \leqslant \frac{1}{\alpha}(\|f\|_{V^*} + C\|u_g\|_V), \tag{10.4}$$

where $\alpha > 0$ and $C > 0$ are the constants in the inf-sup and continuity conditions for a, respectively.

Now set $u = u_f + u_g \in V$ and consider $f - Au \in V^*$, which by construction satisfies

$$\langle f - Au, v\rangle_{V^*,V} = 0 \quad \text{for all } v \in \ker B,$$

i.e., $f - Au \in (\ker B)^0$. Since B^* is surjective on $(\ker B)^0$, we obtain existence of a $\lambda \in M$ satisfying $B^*\lambda = f - Au$ and

$$\|\lambda\|_M \leqslant \frac{1}{\beta}(\|f\|_{V^*} + C\|u\|_V). \tag{10.5}$$

We have thus found $(u, \lambda) \in V \times M$ satisfying

$$Au + B^*\lambda = f$$

and

$$Bu = Bu_g = g.$$

The claimed estimate follows by combining (10.3), (10.4) and (10.5).

To show uniqueness of the solution, consider the difference $(\overline{u},\overline{\lambda})$ of two solutions (u_1,λ_1) and (u_2,λ_2), which solves the homogeneous problem (10.1) with $f=0$ and $g=0$, i.e., $B\overline{u}=0$ and $A\overline{u}+B^*\overline{\lambda}=0$. Then, $\overline{u}\in\ker B$ and the bijectivity of A on $\ker B$ implies $\overline{u}=0$, since

$$\alpha\,\|\overline{u}\|_V^2 \leqslant a(\overline{u},\overline{u}) = a(\overline{u},\overline{u}) + b(\overline{u},\overline{\lambda}) = 0.$$

Hence $B^*\overline{\lambda}=0$. Similarly, from the injectivity of B^* it follows that $\overline{\lambda}=0$. □

10.2 GALERKIN APPROXIMATION OF SADDLE POINT PROBLEMS

To simplify matters, we consider a conforming Galerkin approximation of ($\mathcal{S}$): Choose finite-dimensional subspaces $V_h\subset V$ and $M_h\subset M$ and look for $(u_h,\lambda_h)\in V_h\times M_h$ satisfying

$$(\mathcal{S}_h)\qquad \begin{cases} a(u_h,v_h)+b(v_h,\lambda_h) = \langle f,v_h\rangle_{V^*,V} & \text{for all } v_h\in V_h,\\ b(u_h,\mu_h) = \langle g,\mu_h\rangle_{M^*,M} & \text{for all } \mu_h\in M_h. \end{cases}$$

This approach is called a *mixed finite element method*. It is clear that the choice of V_h and of M_h cannot be independent of each other but must satisfy a compatibility condition similar to Theorem 10.1. Define the operator $B_h : V_h\to M_h^*$ analogously to B.

Theorem 10.2. *Assume there exist constants $\alpha_h,\beta_h>0$ such that*

$$\inf_{u_h\in\ker B_h}\sup_{v_h\in\ker B_h}\frac{a(u_h,v_h)}{\|u_h\|_V\,\|v_h\|_V}\geqslant\alpha_h, \tag{10.6}$$

$$\inf_{\mu_h\in M_h}\sup_{v_h\in V_h}\frac{b(v_h,\mu_h)}{\|v_h\|_V\,\|\mu_h\|_M}\geqslant\beta_h. \tag{10.7}$$

Then, there exists a unique solution $(u_h,\lambda_h)\in V_h\times M_h$ to ($\mathcal{S}_h$) satisfying

$$\|u_h\|_{V_h}+\|\lambda_h\|_{M_h}\leqslant C(\|f_h\|_{V^*}+\|g_h\|_{M^*}).$$

Proof. The claim follows immediately from Theorem 10.1 and the fact that in finite dimensions, the inf-sup condition for a is sufficient to apply Theorem 8.1. □

Note that in general, this is a non-conforming approach since even for $V_h\subset V$ and $M_h\subset M$, we do not have that B_h is the restriction of B to V_h, i.e., $B(V_h)\not\subset M_h^*$ and that $\ker B_h$ need not be a subspace of $\ker B$. Hence, the discrete inf-sup conditions do not follow from the continuous conditions. However, if the subspace V_h is chosen suitably, it is possible to deduce the discrete LBB condition from the continuous one.

Theorem 10.3 (Fortin criterion)**.** *Assume the LBB condition* (10.2) *is satisfied. Then the discrete LBB condition* (10.7) *is satisfied if and only if there exists a linear operator* $\Pi_h : V \to V_h$ *such that*

$$b(\Pi_h v, \mu_h) = b(v, \mu_h) \quad \textit{for all } \mu_h \in M_h$$

and there exists a $\gamma_h > 0$ *such that*

$$\|\Pi_h v\|_V \leqslant \gamma_h \|v\|_V \quad \textit{for all } v \in V$$

holds.

Proof. Assume that such a projector exists. Since $\Pi_h(V) \subset V_h$, we have for all $\mu_h \in M_h$

$$\sup_{v_h \in V_h} \frac{b(v, \mu_h)}{\|v\|_V} \geqslant \sup_{v \in V} \frac{b(\Pi_h v, \mu_h)}{\|\Pi_h v\|_V} \geqslant \sup_{v \in V} \frac{b(v, \mu_h)}{\gamma_h \|v\|_V} \geqslant \frac{\beta}{\gamma_h} \|\mu_h\|_M ,$$

which implies the discrete LBB condition. Conversely, if the discrete LBB condition holds, the operator $B_h : V_h \to M_h^*$ as defined above is surjective and has continuous right inverse, hence for any $v \in V$, there exists a $\Pi_h v \in V_h$ such that $B_h(\Pi_h v) = B_h v \in M_h^*$ and

$$\beta_h \|\Pi_h v\|_V \leqslant \|B_h v\|_{M^*} \leqslant C \|v\|_V .$$

□

A priori error estimates can be obtained using the following variant of Céa's lemma.

Theorem 10.4. *Assume the conditions of Theorem 10.2 are satisfied. Let* $(u, \lambda) \in V \times M$ *and* $(u_h, \lambda_h) \in V_h \times M_h$ *be the solutions to* ($\mathcal{S}$) *and* ($\mathcal{S}_h$)*, respectively. Then there exists a constant* $C > 0$ *such that*

$$\|u - u_h\|_V + \|\lambda - \lambda_h\|_M \leqslant C \left(\inf_{v_h \in V_h} \|u - v_h\|_V + \inf_{\mu_h \in M_h} \|\lambda - \mu_h\|_M \right) .$$

Proof. Due to the discrete LBB condition, the operator $B_h : V_h \to M_h^*$ is surjective and has continuous right inverse. For arbitrary $v_h \in V_h$, consider $B(u - v_h)$ as a linear form on M_h. Hence, there exists $r_h \in V_h$ satisfying $B_h r_h = B(u - v_h)$, i.e.,

$$b(r_h, \mu_h) = b(u - v_h, \mu_h) \qquad \text{for all } \mu_h \in M_h,$$

and

$$\beta_h \|r_h\|_V \leqslant C \|u - v_h\|_V .$$

Furthermore, $w_h = r_h + v_h$ satisfies

$$b(w_h, \mu_h) = b(u, \mu_h) = \langle g, \mu_h \rangle_{M^*, M} \qquad \text{for all } \mu_h \in M_h,$$

hence $u_h - w_h \in \ker B_h$. The discrete inf-sup condition (10.6) thus implies

$$\begin{aligned}\alpha_h \|u_h - w_h\|_V &\leqslant \sup_{x_h \in \ker B_h} \frac{a(u_h - w_h, x_h)}{\|x_h\|_V} \\ &\leqslant \sup_{x_h \in \ker B_h} \frac{a(u_h - u, x_h) + a(u - w_h, x_h)}{\|x_h\|_V} \\ &\leqslant \sup_{x_h \in \ker B_h} \frac{b(x_h, \lambda - \lambda_h) + a(u - w_h, x_h)}{\|x_h\|_V},\end{aligned} \tag{10.8}$$

by taking the difference of the first equations of ($\mathcal{S}$) and ($\mathcal{S}_h$). For any $x_h \in \ker B_h$ and $\mu_h \in M_h$, we have

$$b(x_h, \lambda_h) = 0 = b(x_h, \mu_h)$$

and hence

$$\alpha_h \|u_h - w_h\|_V \leqslant C(\|u - w_h\|_V + \|\lambda - \mu_h\|_M)$$

for arbitrary $\mu_h \in M_h$. Using the triangle inequality, we thus obtain

$$\begin{aligned}\|u - u_h\|_V &\leqslant \|u - w_h\|_V + \|w_h - u_h\|_V \\ &\leqslant (1 + \frac{C}{\alpha_h}) \|u - w_h\|_V + \frac{C}{\alpha_h} \|\lambda - \mu_h\|_M\end{aligned} \tag{10.9}$$

and

$$\|u - w_h\|_V \leqslant \|u - v_h\|_V + \|r_h\|_V \leqslant (1 + \frac{C}{\beta_h}) \|u - v_h\|_V. \tag{10.10}$$

To estimate $\|\lambda - \lambda_h\|_M$, we again use that

$$a(u - u_h, v_h) = b(v_h, \lambda - \lambda_h) = b(v_h, \lambda - \mu_h) + b(v_h, \mu_h - \lambda_h)$$

holds for all $v_h \in V_h$ and $\mu_h \in M_h$. The discrete LBB condition thus implies

$$\beta_h \|\lambda_h - \mu_h\|_M \leqslant C(\|u - u_h\|_V + \|\lambda - \mu_h\|_M).$$

Applying the triangle inequality again, we obtain

$$\begin{aligned}\|\lambda - \lambda_h\|_M &\leqslant \|\lambda - \mu_h\|_M + \|\lambda_h - \mu_h\|_M \\ &\leqslant (1 + \frac{C}{\beta_h}) \|\lambda - \mu_h\|_M + \frac{C}{\beta_h} \|u - u_h\|_V.\end{aligned} \tag{10.11}$$

Combining (10.9), (10.10), and (10.11) yields the claimed estimate. □

This estimate is optimal if the constants α_h, β_h can be chosen independently of h.

If $\ker B_h \subset \ker B$ (i.e., $b(v_h, \mu_h) = 0$ for all $\mu_h \in M_h$ implies $b(v_h, \mu) = 0$ for all $\mu \in M$), we can improve the estimate for u.

Corollary 10.5. *If* $\ker B_h \subset \ker B$,

$$\|u - u_h\|_V \leqslant C \inf_{v_h \in V_h} \|u - v_h\|_V$$

holds.

Proof. If $\ker B_h \subset \ker B$, we have $b(v_h, \lambda - \lambda_h) = 0$ for all $v_h \in \ker B_h$, and hence (10.8) implies

$$\alpha_h \|u_h - w_h\|_V \leqslant C \|u - w_h\|_V .$$

Continuing as above, we obtain the claimed estimate. □

10.3 A MIXED METHOD FOR THE POISSON EQUATION

The classical application of mixed finite element methods is the Stokes equation,[1] which describes the flow of an incompressible fluid. Here, we want to illustrate the theory using a very simple example. Consider the Poisson equation $-\Delta u = f$ on $\Omega \subset \mathbb{R}^n$ with homogeneous Dirichlet conditions. If we introduce $\sigma = \nabla u \in L^2(\Omega)^n$, we can write it as

$$\begin{cases} \nabla u - \sigma = 0, \\ -\nabla \cdot \sigma = f. \end{cases}$$

The weak solution $(\sigma, u) \in L^2(\Omega)^n \times H^1_0(\Omega)$ of this system satisfies

$$\begin{cases} (\sigma, \tau) - (\tau, \nabla u) = 0 & \text{for all } \tau \in L^2(\Omega)^n, \\ -(\sigma, \nabla v) = -(f, v) & \text{for all } v \in H^1_0(\Omega). \end{cases} \tag{10.12}$$

This fits into the abstract framework of § 10.1 by setting $V := L^2(\Omega)^n$, $M := H^1_0(\Omega)$,

$$a(\sigma, \tau) = (\sigma, \tau), \qquad b(\sigma, v) = -(\sigma, \nabla v).$$

Clearly, a is coercive on the whole space V with constant $\alpha = 1$. To verify the LBB condition, we insert $\tau = -\nabla v \in L^2(\Omega)^n = V$ for given $v \in H^1_0(\Omega) = M$ in

$$\sup_{\tau \in V} \frac{b(\tau, v)}{\|\tau\|_V} \geqslant \frac{b(-\nabla v, v)}{\|\nabla v\|_V} = \frac{(\nabla v, \nabla v)}{\|\nabla v\|_{L^2(\Omega)}} = |v|_{H^1(\Omega)} \geqslant C^{-1} \|v\|_{H^1(\Omega)}$$

[1] see, e.g., [Braess 2007, Chapter III.6], [Ern and Guermond 2004, Chapter 4]

using the Poincaré inequality 2.5. Theorem 10.1 thus yields the existence and uniqueness of the solution (σ, u) to (10.12).

To obtain a stable mixed finite element method, we take a shape-regular affine triangulation $\mathcal{T}_h$ of Ω and set for $k \geqslant 1$

$$\begin{aligned} V_h &:= \left\{\tau_h \in L^2(\Omega)^n : \tau_h|_K \in P_{k-1}(K) \text{ for all } K \in \mathcal{T}_h\right\}, \\ M_h &:= \left\{v_h \in C^0(\Omega) : v_h|_K \in P_k(K) \text{ for all } K \in \mathcal{T}_h\right\}. \end{aligned}$$

Since $V_h \subset V$, the coercivity of a on V_h follows as above with constant $\alpha_h = \alpha$. Furthermore, it is easy to verify that $\nabla M_h \subset V_h$, i.e., the gradient of any piecewise linear continuous function is piecewise constant. (We are thus in the special situation of a conforming Galerkin approximation.) Hence, the $L^2(\Omega)^n$ projection from V on V_h verifies the Fortin criterion: If $\Pi_h\sigma \in V_h$ satisfies $(\Pi_h\sigma - \sigma, \tau_h) = 0$ for all $\tau_h \in V_h$ and given $\sigma \in V$, then

$$b(\Pi_h\sigma, v_h) = -(\Pi_h\sigma, \nabla v_h) = -(\sigma, \nabla v_h) = b(\sigma, v_h) \qquad \text{for all } v_h \in M_h$$

since $\nabla v_h \in V_h$. Theorem 10.3 therefore yields the discrete LBB condition and we obtain existence of and (from Corollary 10.5) a priori estimates for the mixed finite element discretization ($\mathcal{S}_h$).

Part IV

TIME-DEPENDENT PROBLEMS

VARIATIONAL THEORY OF PARABOLIC PDES

11

In this chapter, we study evolution equations. For example, if $-\Delta u = f$ (together with appropriate boundary conditions) describes the temperature distribution u in a body due to the heat source f, the *heat equation*

$$\begin{cases} \partial_t u(t,x) - \Delta u(t,x) = f(t,x), \\ \qquad u(0,x) = u_0(x) \end{cases}$$

describes the evolution in time of the temperature distribution u starting from the given initial condition u_0. This is a *parabolic* equation, since the spatial partial differential operator $-\Delta$ is elliptic and only the first time derivative of u appears.

11.1 FUNCTION SPACES

To specify the weak formulation of parabolic problems, we first need to fix the proper functional analytic framework. Let $T > 0$ be a fixed time and $\Omega \subset \mathbb{R}^n$ be a domain, and set $Q := (0,T) \times \Omega$. To respect the special role of the time variable, we consider a real-valued function $u(t,x)$ on Q as a function of t with values in a Banach space V that consists of functions depending on x only:

$$u : (0,T) \to V, \qquad t \mapsto u(t,\cdot) \in V.$$

Similarly to the real-valued case, we can define the following function spaces:

- *Hölder spaces*: For $k \geqslant 0$, define $C^k(0,T;V)$ as the space of all V-valued functions on $[0,T]$ which are k times continuously differentiable with respect to t. Denote by $d_t^j u$ the jth derivative of u. Then $C^k(0,T;V)$ is a Banach space when equipped with the norm

$$\|u\|_{C^k(0,T;V)} := \sum_{j=1}^{k} \sup_{t\in[0,T]} \left\| d_t^j u(t) \right\|_V$$

- *Lebesgue spaces* (also called *Bochner spaces*):[1] For $1 \leqslant p \leqslant \infty$, define $L^p(0,T;V)$ as the space of all V-valued functions on $(0,T)$ for which $t \mapsto \|u(t)\|_V$ is in $L^p(0,T)$, which is a Banach space if equipped with the norm

$$\|u\|_{L^p(0,T;V)} = \begin{cases} \left(\int_0^T \|u(t)\|_V^p \, dt\right)^{\frac{1}{p}} & \text{if } p < \infty, \\ \operatorname{ess\,sup}_{t\in(0,T)} \|u(t)\|_V & \text{if } p = \infty. \end{cases}$$

- *Sobolev spaces*: If $u \in L^p(0,T;V)$ has a weak derivative $d_t u$ (defined in the usual fashion) in $L^p(0,T;V)$, we say that $u \in W^{1,p}(0,T;V)$, which is a Banach space if equipped with the norm

$$\|u\|_{W^{1,p}(0,T;V)} := \|u\|_{L^p(0,T;V)} + \|d_t u\|_{L^p(0,T;V)}\,.$$

  More generally, for $1 < p < \infty$ and two reflexive Banach spaces V_0, V_1 with continuous embedding $V_0 \hookrightarrow V_1$, we set $q = p/(p-1)$ and

$$W^{1,p}(V_0, V_1) := \{v \text{ measurable} : v \in L^p(0,T;V_0) \text{ and } d_t v \in L^q(0,T;V_1)\}.$$

  This is a Banach space if equipped with the norm

$$\|u\|_{W(V_0,V_1)} := \|u\|_{L^p(0,T;V_0)} + \|d_t u\|_{L^q(0,T;V_1)}\,.$$

Let V be a reflexive Banach space with continuous and dense embedding into a Hilbert space H. Identifying H^* with H using the Riesz representation theorem, we have

$$V \subset H \equiv H^* \subset V^*$$

with dense embeddings. We call (V, H, V^*) *Gelfand* or *evolution triple*. We can then transfer (via mollifiers)[2] the usual calculus rules to $W^{1,p}(V, V^*)$. Similarly to the Rellich–Kondrachov theorem, the following embedding tells us that sufficiently smooth functions are continuous in time.

Theorem 11.1. *Let $1 < p < \infty$, $q = p/(p-1)$, and (V, H, V^*) a Gelfand triple. Then, the embedding*

$$W^{1,p}(V, V^*) \hookrightarrow C(0,T;H)$$

is continuous.

This result guarantees that functions in $W^{1,p}(V, V^*)$ have well-defined traces $u(0)$, $u(T)$ in V. We also need the following integration by parts equalities.

[1] For a rigorous definition, see [Wloka 1987, § 24]

[2] For proofs of this and the following result, see, e.g., [Showalter 1997, Proposition III.1.2, Corollary III.1.1], [Wloka 1987, Theorem 25.5 (with obvious modifications)]

Lemma 11.2. *Let* (V, H, V^*) *be a Gelfand triple. For every* $u, v \in W^{1,p}(V, V^*)$,

$$\frac{d}{dt}\langle u, v\rangle_H = \langle d_t u, v\rangle_{V^*,V} + \langle d_t v, u\rangle_{V^*,V}$$

and

$$\int_0^T \langle d_t u(t), v(t)\rangle_{V^*,V}\, dt = \langle u(T), v(T)\rangle_H - \langle u(0), v(0)\rangle_H - \int_0^T \langle d_t v(t), u(t)\rangle_{V^*,V}\, dt.$$

In the following, we restrict ourselves to the case $p = q = 2$, for which $W(V, V^*) := W^{1,2}(V, V^*)$ is a Hilbert space.

11.2 WEAK SOLUTION OF PARABOLIC PDES

We can now formulate our parabolic evolution problem. Given $a : (0, T) \times V \times V \to \mathbb{R}$ such that $a(t, \cdot, \cdot)$ is bilinear for almost all $t \in (0, T)$, $f \in L^2(0, T; V^*)$, and $u_0 \in H$, find $u \in W(V, V^*)$ such that

$$\begin{cases} \langle d_t u, v\rangle_{V^*,V} + a(t, u, v) = \langle f, v\rangle_{V^*,V} & \text{for a.e. } t \in (0, T),\ \text{all } v \in V \\ u(0) = u_0 \end{cases} \tag{11.1}$$

(For the heat equation, e.g., $V = H^1_0(\Omega) \subset L^2(\Omega) = H$ and $a(t, u, v) = (\nabla u, \nabla v)$.) Just as in the stationary case, this can be expressed equivalently in weak form (using the fact that functions in $W(V, V^*)$ are continuous in time). For simplicity, assume $u_0 = 0$ (the inhomogeneous case can be treated in the same fashion as inhomogeneous Dirichlet conditions) and consider the Hilbert spaces

$$Y = L^2(0, T; V), \qquad X = \{v \in W(V, V^*) : v(0) = 0\}.$$

Setting

$$b : X \times Y \to \mathbb{R}, \qquad b(u, y) = \int_0^T \langle d_t u, y\rangle_{V^*,V} + a(t, u, y)\, dt$$

and

$$\langle f, y\rangle_{Y^*,Y} = \int_0^T \langle f, y\rangle_{V^*,V}\, dt,$$

we look for $u \in X$ such that

$$b(u, y) = \langle f, y\rangle_{Y^*,Y} \qquad \text{for all } y \in Y. \tag{11.2}$$

Well-posedness of (11.1) can thus be shown using the Banach–Nečas–Babuška theorem.

Theorem 11.3. *Assume that* $a : (0,T) \times V \times V$ *satisfies the following properties:*

(i) $t \mapsto a(t,u,v)$ *is measurable for all* $u, v \in V$.

(ii) There exists $M > 0$ *such that* $|a(t,u,v)| \leqslant M \|u\|_V \|v\|_V$ *for almost all* $0 < t < T$ *and all* $u, v \in V$.

(iii) There exists $\alpha > 0$ *such that* $a(t,u,u) \geqslant \alpha \|u\|_V^2$ *for almost all* $0 < t < T$ *and all* $u \in V$.

Then, (11.2) *has a unique solution* $u \in W(V, V^*)$ *satisfying*

$$\|u\|_{W(V,V^*)} \leqslant \frac{1}{\alpha} \|f\|_Y .$$

Proof. Continuity of b and $y \mapsto \langle f, y\rangle_{Y^*,Y}$ follows from their definition and the continuity of a. To verify the inf-sup condition, we define for almost every $t \in (0,T)$ the operator $A(t) : V \to V^*$ by $\langle A(t)u, v\rangle_{V^*,V} = a(t,u,v)$ for all $u, v \in V$. Continuity of a implies that $A(t)$ is a bounded operator with constant M. Furthermore, coercivity of a and the Lax–Milgram theorem yields that $A(t)$ is an isomorphism, hence $A(t)^{-1}$ is bounded as well with constant α^{-1}. Therefore, for all $x \in V^*$,

$$\begin{aligned}(11.3)\qquad \left\langle x, A(t)^{-1}x\right\rangle_{V^*,V} &= \left\langle A(t)A(t)^{-1}x, A(t)^{-1}x\right\rangle_{V^*,V} \geqslant \alpha \left\|A(t)^{-1}x\right\|_V^2 \\ &\geqslant \frac{\alpha}{M^2} \|x\|_{V^*}^2 .\end{aligned}$$

For arbitrary $u \in X$ and $\mu > 0$, set $w = A(t)^{-1} d_t u + \mu u \in Y$. By (11.3), the triangle inequality, and the definition of X and Y we have that

$$\|w\|_Y \leqslant \alpha^{-1} \|d_t u\|_Y + \mu \|u\|_Y \leqslant c \|u\|_X .$$

Moreover, using (11.3), integration by parts, continuity of $A(t)$ and $A(t)^{-1}$ and coercivity of a, respectively, we can estimate

$$\begin{aligned} b(u,w) &= \int_0^T \left\langle d_t u + A(t)u, A(t)^{-1} d_t u + \mu u\right\rangle_{V^*,V} \,dt \\ &\geqslant \frac{\alpha}{M^2} \int_0^T \|d_t u\|_{V^*}^2 \,dt + \frac{\mu}{2} \|u(T)\|_H^2 - \frac{M}{\alpha} \int_0^T \|u\|_V \|d_t u\|_{V^*} \,dt + \mu\alpha \int_0^T \|u\|_V^2 \,dt \\ &\geqslant \frac{\alpha}{2M^2} \int_0^T \|d_t u\|_{V^*}^2 \,dt + \left(\mu\alpha - \tfrac{M^4}{2\alpha^3}\right) \int_0^T \|u\|_V^2 \,dt \end{aligned}$$

by Young's inequality. Taking $\mu = M^4 \alpha^{-4}$ yields

$$b(u,w) \geqslant c \|u\|_X^2 \geqslant c \|u\|_X \|w\|_Y$$

and hence

$$\inf_{u\in X} \sup_{v\in Y} \frac{b(u,v)}{\|u\|_X \|v\|_Y} \geqslant \inf_{u\in X} \frac{b(u,w)}{\|u\|_X \|w\|_Y} \geqslant c.$$

It remains to show that the injectivity condition holds. Assume $v \in Y$ is such that $b(u, v) = 0$ for all $u \in X$. Inserting $\varphi \in C_0^\infty(0, T; V) \subset X$ into (11.2) and integrating by parts, we see that

$$\left|\int_0^T \langle d_t v, \varphi\rangle_{V^*,V}\, dt\right| = \left|\int_0^T \langle d_t \varphi, v\rangle_{V^*,V}\, dt\right| = \left|\int_0^T a(t, \varphi, v)\, dt\right| \leqslant C\, \|\varphi\|_X\, \|v\|_Y\,,$$

hence $d_t v \in L^2(0, T; V^*)$ and

$$\int_0^T \langle -d_t v, u\rangle_{V^*,V} + a(t, u, v)\, dt = 0 \qquad \text{for all } u \in C_0^\infty(0, T; V). \tag{11.4}$$

By density of $C_0^\infty(0, T; V)$ in Y, this holds for all $u \in Y$. Inserting $u = t\varphi \in Y$ for arbitrary $\varphi \in V$ and integrating by parts yields

$$\begin{aligned} 0 &= \int_0^T \langle -d_t v, t\varphi\rangle_{V^*,V} + a(t, t\varphi, v)\, dt \\ &= -\langle v(T), T\varphi\rangle_H + \int_0^T \langle d_t(t\varphi), v\rangle_{V^*,V} + a(t, t\varphi, v)\, dt \\ &= -T\,\langle v(T), \varphi\rangle_H \end{aligned}$$

by noting that $t\varphi \in X$ and hence $b(t\varphi, v) = 0$. Since V is dense in H, this implies that $v(T) = 0$. Finally, we can insert $v \in Y$ into (11.4) and use Lemma 11.2 to obtain

$$\begin{aligned} 0 &= \int_0^T -\langle d_t v, v\rangle_{V^*,V} + a(t, v, v)\, dt \\ &\geqslant -\int_0^T \frac{d}{dt}\left(\frac{1}{2}\|v(t)\|_V^2\right) + \alpha\, \|v(t)\|_V^2\, dt \\ &= \frac{1}{2}\|v(0)\|_V^2 + \alpha\, \|v\|_Y^2 \end{aligned}$$

and hence $v = 0$. We can thus apply the Banach–Nečas–Babuška theorem, and the claim follows. □

GALERKIN APPROACH FOR PARABOLIC PROBLEMS

12

To obtain a finite-dimensional approximation of (11.1), we need to discretize in time and space.

12.1 TIME STEPPING METHODS

Based on the order of operations, we can discern three popular approaches:

METHOD OF LINES. This method starts with a discretization in space to obtain a system of ordinary differential equations, which are then solved with one of the vast number of available methods. In the context of finite element methods, we use a discrete space V_h of piecewise polynomials defined on the triangulation $\mathcal{T}_h$ of the domain Ω. Given a nodal basis $\{\varphi_j\}$ of V_h, we approximate the unknown solution as $u_h(t,x) = \sum_j U_j(t)\varphi_j(x)$. Letting $\mathcal{P}_h$ denote the L^2 projection on V_h and using the mass matrix $\mathbf{M}_{ij} = (\varphi_i, \varphi_j)$ and the (time-dependent) stiffness matrix $\mathbf{K}(t)_{ij} = a(t; \varphi_i, \varphi_j)$ yields the following linear system of ordinary differential equations for the coefficient vector $U(t) = (U_1(t), \dots)^T$:

$$\begin{cases} \mathbf{M}\dfrac{d}{dt}U(t) + \mathbf{K}U(t) = \mathbf{M}F(t), \\ \qquad\qquad\qquad\qquad U(0) = U_0, \end{cases}$$

where U_0 and $F(t)$ are the vectors of coefficients of $\mathcal{P}_h u_0$ and $\mathcal{P}_h f(t)$, respectively. The choice of integration method for this system depends on the properties of $\mathbf{K}$, such as its stiffness, which can lead to numerical instability. Some details can be found, e.g., in [Ern and Guermond 2004, Chapter 6.1].

ROTHE'S METHOD. This method consists in treating (11.1) as an ordinary differential equation in the Banach space V, which is discretized in time by replacing the time derivative $d_t u$ by a difference quotient:

- The *implicit Euler scheme* uses the backward difference quotient

  $$d_t u(t+h) \approx \frac{u(t+h)-u(t)}{h}$$

  for $h > 0$ at time $t+h$ to obtain for given $u(t)$ and unknown $u(t+h) \in V$ the *stationary* partial differential equation

  $$\langle u(t+h), v\rangle_H + h\, a(t+h; u(t+h), v) = \langle u(t), v\rangle_H + h\,\langle f(t+h), v\rangle_{V^*,V}$$

  for all $v \in V$.

- The *Crank–Nicolson scheme* uses the central difference quotient

  $$d_t u(t+\tfrac{h}{2}) \approx \frac{u(t+h)-u(t)}{h}$$

  for $h > 0$ at time $t + \frac{h}{2}$ to obtain

  $$\begin{aligned}\langle u(t+h), v\rangle_H + \frac{h}{2}\, a(t+\frac{h}{2}; u(t+h), v) = \langle u(t), v\rangle_H - \frac{h}{2}\, a(t+\frac{h}{2}; u(t), v)\\ + h\,\big\langle f(t+\tfrac{h}{2}), v\big\rangle_{V^*,V}\end{aligned}$$

  for all $v \in V$.

Starting with $t = 0$, these are then approximated and solved in turn using a finite element discretization in space. This approach is discussed in detail in [Thomée 2006, Chapters 7–9]. The advantage of this approach is that at each time step $t_m := t + mh$, a different spatial discretization can be used.

SPACE-TIME GALERKIN SCHEMES. Finally, we can proceed as in the stationary case and apply a Galerkin approximation to (11.2): Choose finite-dimensional subspaces $X_h \subset X$ and $Y_h \subset Y$ and find $u_h \in X_h$ such that

$$\int_0^T \langle d_t u_h(t), y_h(t)\rangle_{V^*,V} + a(t; u_h(t), y_h(t))\, dt = \int_0^T \langle f(t), y_h(t)\rangle_{V^*,V}\, dt \tag{12.1}$$

for all $y_h \in Y_h$. This approach is closely related to Rothe's method, if we choose the discrete spaces as tensor products in space and time: Let

$$0 = t_0 < t_1 < \cdots < t_N = T$$

and choose for each t_m, $1 \leqslant m \leqslant N$, a (possibly different) finite-dimensional subspace $V_m \subset V$. Let $P_r(t_{m-1}, t_m; V_m)$ denote the space of polynomials on the interval $[t_{m-1}, t_m]$ with degree up to r with values in V_m. Then we define

$$\begin{aligned} X_h &= \left\{ y_h \in C(0,T;V) : y_h|_{[t_m,t_{m+1}]} \in P_r(t_{m-1},t_m;V_m),\ 1 \leqslant m \leqslant N,\ v_h(0) = 0 \right\}, \\ Y_h &= \left\{ y_h \in L^2(0,T;V) : y_h|_{[t_m,t_{m+1}]} \in P_{r-1}(t_{m-1},t_m;V_m),\ 1 \leqslant m \leqslant N \right\}. \end{aligned}$$

Since this is a conformal approximation, we can deduce well-posedness of the corresponding discrete problem in the usual fashion (noting that $d_t u_h \in Y_h$).

To see the relation to Rothe's method, consider the case $r = 1$ (i.e., piecewise linear in time) and, for simplicity, a time-independent bilinear form. Since functions in X_h are continuous at $t = t_m$ for all $0 \leqslant m \leqslant N$ and linear on each intervall $[t_{m-1}, t_m]$, we can write

$$u_h(t) = \frac{t_m - t}{t_m - t_{m-1}} u_h(t_{m-1}) + \frac{t - t_{m-1}}{t_m - t_{m-1}} u_h(t_m), \qquad t \in [t_{m-1}, t_m].$$

Similarly, functions in Y_h are constant and thus

$$y_h(t) = y_h(t_{m-1}) =: v_h \in V_m.$$

Inserting this into (12.1) and setting $k_m := t_m - t_{m-1}$ yields

$$\langle u_h(t_m) - u_h(t_{m-1}), v_h \rangle_{V^*,V} + \frac{k_m}{2} a(u_h(t_{m-1}) + u_h(t_m), v_h) = \int_{t_{m-1}}^{t_m} \langle f(t), v_h \rangle_{V^*,V}\, dt,$$

which is a modified Crank–Nicolson scheme (which, in fact, can be obtained by approximating the integral on the right hand side using the midpoint rule which is exact for $y_h \in Y_h$). For this method, one can show error estimates of the form[1]

$$\|u_h(t_m) - u(t_m)\|_{L^2(\Omega)} \leqslant C(h^r \|u_0\|_{H^r(\Omega)} + k^2 \|u_0\|_{H^4(\Omega)}),$$

for $f = 0$ and $u_0 \neq 0$, where r depends on the accuracy of the spatial discretization, and $k = \max k_m$.

12.2 DISCONTINUOUS GALERKIN METHODS FOR PARABOLIC PROBLEMS

Just as for stationary first order equations, however, the *discontinuous Galerkin* method has proved to be very successful for parabolic problems, so we shall focus our analysis on these methods. Let $J_m := (t_{m-1}, t_m]$ denote the half-open interval between two time steps of length $k_m = t_m - t_{m-1}$. Then we set for $r \geqslant 0$

$$X_h = Y_h = \left\{ y_h \in L^2(0,T;V) : y_h|_{J_m} \in P_r(t_{m-1}, t_m; V_m),\ 1 \leqslant m \leqslant N \right\} \subset Y,$$

[1] [Thomée 2006, Theorem 7.8]

where V_m is again a finite-dimensional subspace of V. Note that functions in X_h can be discontinuous at the points t_m, but are continuous from the left with limits from the right, and so we will write

$$u_m := u_h(t_m) = \lim_{\varepsilon\to 0} u_h(t_m - \varepsilon), \qquad u_m^+ := \lim_{\varepsilon\to 0} u_h(t_m + \varepsilon)$$

and

$$[\![u_h]\!]_m = u_m^+ - u_m.$$

As in the stationary case, we now define (by integration by parts on each interval J_m and rearranging the jump terms) the bilinear form

$$b_h(u,y) = \sum_{m=1}^{N} \int_{J_m} \langle d_t u, y\rangle_{V^*,V} + a(t;u,y)\,dt + \sum_{m=1}^{N} \left\langle [\![u]\!]_{m-1}, y_{m-1}^+ \right\rangle_H .$$

Note that as $0 \notin J_1$, we will need to specify $u_0 = u_h(0)$ separately, which we do by setting $[\![u_h]\!]_0 := u_0^+ - u_0$. Since the exact solution $u \in X$ is continuous and satisfies $u(0) = u_0$, we have

$$b_h(u, y_h) = \langle f, y_h\rangle_{Y^*,Y} \qquad \text{for all } y_h \in X_h,$$

and hence this is a consistent approximation. We next show existence and uniqueness of the discrete solution u_h.

Theorem 12.1. *Under the assumptions of Theorem 11.3, there exists a unique solution* $u_h \in X_h$ *satisfying*

$$b_h(u_h, y_h) = \langle f, y_h\rangle_{Y^*,Y} \qquad \textit{for all } y_h \in X_h. \tag{12.2}$$

Proof. Continuity of b_h is clear. To apply the Banach–Nečas–Babuška theorem, it thus remains to verify either the inf-sup or the injectivity condition (since X_h is finite-dimensional). We choose the latter. Let $y_h \in X_h$ satisfy $b_h(u_h, y_h) = 0$ for all $u_h \in X_h$. Since functions in X_h can be discontinuous at the time points t_m, we can insert $u_h = \chi_{J_m}(t)v_h \in X_h$ for each $1 \leqslant m \leqslant N$, where $\chi_{J_m}(t) = 1$ if $t \in J_m$ and zero else. We start with $J_N = (t_{N-1}, t_N]$. Since χ_{J_N} is constant on J_N and zero outside J_N, we have $u_{N-1} = 0$ and thus

$$\begin{aligned}
0 &= b_h(y_h\chi_{J_N}, y_h)\\
&= \int_{J_N} \langle d_t y_h, y_h\rangle_{V^*,V} + a(t;y_h,y_h)\,dt + \left\langle y_{N-1}^+ - y_{N-1}, y_{N-1}^+\right\rangle_H\\
&\geqslant \frac{1}{2}\|y_N\|_H^2 - \frac{1}{2}\left\|y_{N-1}^+\right\|_H^2 + \alpha\int_{J_N} \|y_h\|_V^2\,dt + \left\|y_{N-1}^+\right\|_H^2\\
&\geqslant \frac{1}{2}\left\|y_{N-1}^+\right\|_H^2 + \alpha\int_{J_N}\|y_h\|_V^2\,dt.
\end{aligned}$$

Hence, $y_h|_{J_N} = 0$ and $y_{N-1}^+ = 0$, and we can proceed in the same way for $J_{N-1}, J_{N-2}, \dots, J_1$ to deduce that $y_h = 0$. □

We next show a stability result for the discontinuous Galerkin approximation. For simplicity, we assume from now on that the bilinear form a is time-independent and symmetric, and that $V_1 = \cdots = V_N = V_h$. Let $A : V \to V^*$ again denote the operator corresponding to the bilinear form a, i.e., $\langle Au, v\rangle_H = a(u, v)$ for all $u, v \in V$, where we assume that $Au \in H$ due to the higher regularity (e.g., from Theorem 2.8 or Theorem 2.9) of the corresponding stationary equation.

Theorem 12.2. *For given* $f \in L^2(0, T; H)$ *and* $u_0 \in H$*, the solution* $u_h \in X_h$ *of* (12.2) *satisfies*

$$\sum_{m=1}^{N} \int_{J_m} \|d_t u_h\|_H^2 + \|Au_h\|_H^2 \, dt + \sum_{m=1}^{N} \left\|[\![u_h]\!]_{m-1}\right\|_H^2 \leqslant C \left(\int_0^T \|f\|_H^2 \, dt + \|u_0\|_H^2 \right).$$

Proof. We estimate in turn each term on the left hand side by inserting suitable test functions y_h in (12.2).

Step 1. To estimate $\|Au_h\|_H$, we set $y_h = \chi_{J_m}(t) Au_h$ for $1 \leqslant m \leqslant N$ to obtain

$$\int_{J_m} \langle d_t u_h, Au_h\rangle_H + \|Au_h\|_H^2 \, dt + \left\langle [\![u_h]\!]_{m-1}, (Au_h)_{m-1}^+ \right\rangle_H = \int_{J_m} \langle f, Au_h\rangle_H \, dt.$$

Due to the bilinearity and symmetry of a, we have

$$\begin{aligned}\int_{J_m} \langle d_t u_h, Au_h\rangle_H \, dt = \int_{J_m} a(u_h, d_t u_h) \, dt &= \int_{J_m} \frac{d}{dt} \left(\frac{1}{2} a(u_h, u_h) \right) dt \\ &= \frac{1}{2} a(u_m, u_m) - \frac{1}{2} a(u_{m-1}^+, u_{m-1}^+).\end{aligned}$$

Similarly, since A is time-independent,

$$\begin{aligned}\left\langle [\![u_h]\!]_{m-1}, (Au_h)_{m-1}^+ \right\rangle_H &= a([\![u_h]\!]_{m-1}, u_{m-1}^+) \\ &= \frac{1}{2} a([\![u_h]\!]_{m-1}, u_{m-1}^+ + u_{m-1} + [\![u_h]\!]_{m-1}) \\ &= \frac{1}{2} a(u_{m-1}^+, u_{m-1}^+) - \frac{1}{2} a(u_{m-1}, u_{m-1}) \\ &\quad + \frac{1}{2} a([\![u_h]\!]_{m-1}, [\![u_h]\!]_{m-1}).\end{aligned}$$

Inserting these into the bilinear form $b_h(u_h, y_h)$ yields

$$\begin{aligned}a([\![u_h]\!]_{m-1}, [\![u_h]\!]_{m-1}) + a(u_m, u_m) - a(u_{m-1}, u_{m-1}) + 2\int_{J_m} \|Au_h\|_H^2 \, dt& \\ = 2\int_{J_m} \langle f, Au_h\rangle_H \, dt.&\end{aligned}$$

Summing over all $1 \leqslant m \leqslant N$ yields

$$\sum_{m=1}^{N} a(\llbracket u_h \rrbracket_{m-1}, \llbracket u_h \rrbracket_{m-1}) + \sum_{m=1}^{N} \int_{J_m} 2 \|Au_h\|_H^2 \, dt \\ \leqslant \sum_{m=1}^{N} \int_{J_m} 2 \langle f, Au_h \rangle_H \, dt + a(u_0, u_0).$$

For $2 \leqslant m \leqslant N$, we can simply use coercivity of a to eliminate the jump terms and apply Young's inequality to $\langle f, Au_h \rangle_H$ to absorb the norm of Au_h on J_m in the left hand side. For $m = 1$, we use that

$$a(\llbracket u_h \rrbracket_0, \llbracket u_h \rrbracket_0) - a(u_0, u_0) = a(u_0^+, u_0^+) - 2a(u_0, u_0^+)$$

and for $\varepsilon > 0$ the generalized Young's inequality

$$a(u_0, u_0^+) = \langle u_0, Au_0^+ \rangle_H \leqslant \frac{\varepsilon}{2} \|Au_0^+\|_H^2 + \frac{1}{2\varepsilon} \|u_0\|_H^2 .$$

Since $\|Au_h\|_H^2$ is a polynomial in t of degree up to $2r$ on J_1, we have the estimate

$$k_1 \|Au_0^+\|_H^2 \leqslant C \int_{t_0}^{t_1} \|Au_h\|_H^2 \, dt.$$

Choosing $\varepsilon > 0$ small enough such that $\varepsilon C k_1^{-1} < 1$ yields

$$\sum_{m=1}^{N} \int_{J_m} \|Au_h\|_H^2 \, dt \leqslant C \left(\int_0^T \|f\|_H^2 \, dt + \|u_0\|_H^2 \right). \tag{12.3}$$

Step 2. For the bound on $d_t u_h$, we use the inverse estimate

$$\int_{J_m} \|y_h\|_H^2 \, dt \leqslant C k_m^{-1} \int_{J_m} (t - t_{m-1}) \|y_h\|_H^2 \, dt$$

for all $y_h \in P_r(t_{m-1}, t_m; V_h)$, which follows from a scaling argument in time and equivalence of norms on the finite-dimensional space $P_r(0, 1; V_h)$. Now choose $y_h = \chi_{J_m}(t)(t - t_{m-1}) d_t u_h$ for $1 \leqslant m \leqslant N$. Since $y_{m-1}^+ = 0$, we have using the Cauchy–Schwarz inequality that

$$\int_{J_m} (t - t_{m-1}) \|d_t u_h\|_H^2 \, dt = \int_{J_m} (t - t_{m-1}) \langle f - Au_h, d_t u_h \rangle_H \, dt \\ \leqslant \left(\int_{J_m} (t - t_{m-1}) \|f - Au_h\|_H^2 \, dt \right)^{\frac{1}{2}} \\ \left(\int_{J_m} (t - t_{m-1}) \|d_t u_h\|_H^2 \, dt \right)^{\frac{1}{2}} .$$

Applying the inverse estimate for $y_h = d_t u_h$, the Cauchy–Schwarz inequality for the first integral and estimating the norm there using (12.3) yields

$$\sum_{m=1}^{N} \int_{J_m} \|d_t u_h\|_H^2 \, dt \leqslant C \left(\int_0^T \|f\|_H^2 \, dt + \|u_0\|_H^2 \right). \tag{12.4}$$

Step 3. It remains to estimate the jump terms. For this, we set $y_h = \chi_{J_m}(t) \, [\![u_h]\!]_{m-1}$ for $1 \leqslant m \leqslant N$. This yields

$$\begin{aligned} \left\| [\![u_h]\!]_{m-1} \right\|_H^2 &= \int_{J_m} \left\langle f - Au_h, [\![u_h]\!]_{m-1} \right\rangle_H - \left\langle d_t u_h, [\![u_h]\!]_{m-1} \right\rangle_H \, dt \\ &\leqslant \frac{k_m}{2} \int_{J_m} \|f - Au_h - d_t u_h\|_H^2 \, dt + \frac{1}{2k_m} \int_{J_m} \left\| [\![u_h]\!]_{m-1} \right\|_H^2 \, dt, \end{aligned}$$

where we have used the generalized Young's inequality. Since $[\![u_h]\!]_{m-1}$ is constant in time, we have

$$\int_{J_m} \left\| [\![u_h]\!]_{m-1} \right\|_H^2 \, dt = k_m \left\| [\![u_h]\!]_{m-1} \right\|_H^2 .$$

From (12.3) and (12.4), we thus obtain

$$\sum_{m=1}^{N} k_m^{-1} \left\| [\![u_h]\!]_{m-1} \right\|_H^2 \leqslant C \left(\int_0^T \|f\|_H^2 \, dt + \|u_0\|_H^2 \right),$$

which completes the proof. □

Before we address a priori error estimates, we show how to formulate discontinuous Galerkin methods as time stepping methods. First consider the case $r = 0$, i.e., piecewise constant functions in time. Then, $d_t u_h = 0$ and $u_h(t)|_{J_m} = u_m = u_{m-1}^+ \in V_h$. Using as test functions $y_h = \chi_{J_m}(t) v_h$ for arbitrary $v_h \in V_h$ and $m = 1, \ldots, N$, we obtain

$$\langle u_m, v_h \rangle_H + k_m \, a(u_m, v_h) = \langle u_{m-1}, v_h \rangle_H + k_m \int_{J_m} \langle f, v_h \rangle_{V^*,V} \, dt$$

for all $v_h \in V_h$, which is a variant of the implicit Euler scheme.[2]

For $r = 1$ (piecewise linear functions), we make the ansatz

$$u_h|_{J_m}(t) = u_m^0 + \frac{t - t_{m-1}}{k_m} u_m^1$$

[2] If the discrete spaces are different for each time interval, we need to use the H-projection of u_{m-1} on V_m.

for $u_m^0, u_m^1 \in V_h$. Again, we choose for each J_m test functions which are zero outside J_m; specifically, we take $\chi_{J_m}(t)v_h$ and $\chi_{J_m}(t)\frac{t-t_{m-1}}{k_m}w_h$ for arbitrary $v_h, w_h \in V_h$. Inserting these into the bilinear form and computing the integrals yields the coupled system

$$\begin{aligned}
\langle u_m^0, v_h\rangle_H + k_m\, a(u_m^0, v_h) + \langle u_m^1, v_h\rangle_H + \frac{k_m}{2}\, a(u_m^1, v_h)& \\
= \langle u_{m-1}, v_h\rangle_H + k_m \int_{J_m} \langle f, v_h\rangle_{V^*,V}\, dt&,\\
\frac{k_m}{2}\, a(u_m^0, v_h) + \frac{1}{2}\langle u_m^1, v_h\rangle_H + \frac{k_m}{3}\, a(u_m^1, v_h)& \\
= \frac{1}{k_m}\int_{J_m} (t-t_{m-1})\, \langle f, v_h\rangle_{V^*,V}\, dt&
\end{aligned}$$

for all $v_h, w_h \in V_h$. By solving this system successively at each time step and setting $u_m = u_m^0 + u_m^1$, we obtain the approximate solution u_h.[3]

12.3 A PRIORI ERROR ESTIMATES

As before, we will estimate the error $u - u_h$ using the approximation properties of the space X_h. Due to the discontinuity of the functions in X_h, we can use a local projection on each time intervall J_m to bound the approximation error. It will be convenient to split this error into two parts: one due to the temporal and one due to the spatial discretization.

We first consider the temporal discretization error. Let

$$X_r = \left\{y_r \in L^2(0,T;V) : y_r|_{J_m} \in P_r(t_{m-1}, t_m; V),\ 1 \leqslant m \leqslant N\right\}$$

and consider the local projection $\pi_r(u) \in X_r$ of $u \in X$ defined by $\pi_r(u)(t_0) = u(t_0)$ and

$$\begin{cases}
\pi_r(u)(t_m) = u(t_m),\\
\displaystyle\int_{J_m} (u(t) - \pi_r(u)(t))\varphi(t)\, dt = 0 \qquad \text{for all } \varphi \in P_{r-1}(J_m; V),
\end{cases}$$

for all $1 \leqslant m \leqslant N$. (For $r = 0$, the second condition is void.) This projection is well-defined since $u \in X$ is continuous in time, and hence the interpolation conditions make sense. Using the Bramble–Hilbert lemma and a scaling argument, we obtain for sufficiently smooth u the following error estimate for every $t \in J_m$, $1 \leqslant m \leqslant N$:

$$\|u(t) - \pi_r(u)(t)\|_H \leqslant C k_m^{r+1} \int_{J_m} \left\|d_t^{r+1} u(t)\right\|_H\, dt.$$

[3] Similarly, discontinuous Galerkin methods for $r \geqslant 2$ lead to $r+1$-stage implicit Runge–Kutta time-stepping schemes.

Similarly, we assume that for each $t \in [0,T]$ the spatial interpolation error in V_h satisfies the estimate

$$\|u(t) - \mathcal{I}_h u(t)\|_H + h\,\|u(t) - \mathcal{I}_h u(t)\|_V \leqslant Ch^{s+1}\,\|u(t)\|_{H^{s+1}(\Omega)}\,.$$

(This is the case, e.g., if $H = L^2(\Omega)$, $V = H_0^1(\Omega)$, and V_h consists of continuous piecewise polynomials of degree $s \geqslant 1$, cf. Theorem 5.9.)

Finally, we will make use of a duality argument, which requires considering for given $\varphi \in H$ the solution of the adjoint equation

$$b_h(y_h, z_h) = \langle y_N, \varphi\rangle_H\,.$$

Integrating by parts on each interval J_m and rearranging the jump terms, we can express the adjoint equation as

$$\sum_{m=1}^{N}\int_{J_m} -\langle y_h, d_t z_h\rangle_H + a(y_h, z_h)\,dt + \sum_{m=1}^{N-1}\big\langle y_m, [\![z_h]\!]_m\big\rangle_H + \langle y_N, z_N\rangle_H = \langle y_N, \varphi\rangle_H\,. \tag{12.5}$$

This can be interpreted as a backwards in time equation with "initial value" $z_h(t_N) = \varphi$. Making the substitution $\tau = t_N - t$, we can apply Theorem 12.2 to obtain

$$\sum_{m=1}^{N}\int_{J_m} \|d_t z_h\|_H^2 + \|Az_h\|_H^2\,dt + \sum_{m=1}^{N}\big\|[\![z_h]\!]_{m-1}\big\|_H^2 \leqslant C\,\|\varphi\|_H^2\,. \tag{12.6}$$

Now everything is in place to show the following a priori estimate for the solution at the time steps.

Theorem 12.3. *For $r = 0$, the solutions $u \in X$ to (11.2) and $u_h \in X_h$ to (12.2) satisfy*

$$\|u(t_m) - u_m\|_H \leqslant C \max_{1\leqslant n\leqslant m}\left(h^{s+1}\sup_{t\in J_n}\|u(t)\|_{H^{s+1}(\Omega)} + k_n\int_{J_n}\|d_t u\|_H\,dt\right)$$

for all $1 \leqslant m \leqslant N$.

Proof. Let $r = 0$. We write the error $e(t)$ at each time t as

$$\begin{aligned} e(t) = u(t) - u_h(t) &= (u(t) - \mathcal{I}_h\pi_r(u)(t)) + (\mathcal{I}_h\pi_r(u)(t) - u_h(t)) \\ &:= e_1(t) + e_2(t). \end{aligned}$$

For $t = t_m$, we have $\pi_r(u)(t_m) = u(t_m)$ by construction, and hence

$$\|e_1(t_m)\|_H = \|\mathcal{I}_h u(t_m) - u(t_m)\|_H \leqslant Ch^{s+1}\,\|u(t_m)\|_{H^{s+1}(\Omega)}\,. \tag{12.7}$$

To bound $e_2(t_m)$, we use the duality trick. For arbitrary $\varphi \in H$, let z_h denote the solution to (12.5) with $N = m$. Since we have a consistent approximation, we can use the Galerkin orthogonality to deduce

$$0 = b_h(e, y_h) = b_h(e_1, y_h) + b_h(e_2, y_h) \qquad \text{for all } y_h \in X_h.$$

From this and $d_t(z_h|_{J_n}) = 0$ we obtain with $y_h = e_2 \in X_h$ that

$$\begin{aligned} \langle e_2(t_m), \varphi\rangle_H &= b_h(e_2, z_h) = -b_h(e_1, z_h) \\ &= -\sum_{n=1}^{m} \int_{J_n} a(e_1, z_h)\,dt + \sum_{n=1}^{m-1} \langle e_1(t_n), [\![z_h]\!]_n\rangle_H - \langle e_1(t_m), \varphi\rangle_H . \end{aligned}$$

Introducing $\langle Az_h, e_1\rangle_H = a(e_1, z_h)$ as above and estimating e_1 by its pointwise in time maximum yields

$$|\langle e_2(t_m), \varphi\rangle_H| \leqslant \left(\sup_{t\leqslant t_m} \|e_1(t)\|_H\right)\left(\sum_{n=1}^{m}\int_{J_n} \|Az_h\|_H\,dt + \sum_{n=1}^{m-1} \left\|[\![z_h]\!]_n\right\|_H + \|\varphi\|_H\right).$$

From the dual definition of the norm in H and estimate (12.6), we obtain

$$\|e_2(t_m)\|_H \leqslant C \max_{1\leqslant n\leqslant m} \sup_{t\in J_n} \|e_1(t)\|_H . \tag{12.8}$$

It remains to bound $e_1(t)$ for arbitrary $t \in J_n$, which we do by estimating

$$\begin{aligned} \|e_1(t)\|_H &= \|u(t) - \mathcal{I}_h\pi_r(u)(t)\|_H \\ &\leqslant \|u(t) - \pi_r(u)(t)\|_H + \|\pi_r(u)(t) - \mathcal{I}_h\pi_r(u)(t)\|_H \\ &\leqslant Ck_n \int_{J_n} \|d_t u\|_H\,dt + Ch^{s+1}\|u(t)\|_{H^{s+1}(\Omega)} . \end{aligned} \tag{12.9}$$

Combining (12.7), (12.8) and (12.9) yields the claim. □

For $r = 1$, one can proceed similarly (using that $d_t z_h|_{J_m} \in P_{r-1}(J_m, V_h)$, and hence that $\int_{J_n} \langle d_t z_h, e_1\rangle_H\,dt$ vanishes by definition of π_r) to obtain

Theorem 12.4. *For $r = 1$, the solutions $u \in X$ to (11.2) and $u_h \in X_h$ to (12.2) satisfy*

$$\|u(t_m) - u_m\|_H \leqslant C \max_{1\leqslant n\leqslant m}\left(h^{s+1}\sup_{t\in J_n}\|u(t)\|_{H^{s+1}(\Omega)} + k_n^3\int_{J_n}\left\|d_t^2 u\right\|_{H^2(\Omega)}\,dt\right)$$

for all $1 \leqslant m \leqslant N$.

The general case (including time-dependent bilinear form a and different discrete spaces V_m) can be found in [Chrysafinos and Walkington 2006].

BIBLIOGRAPHY

R. A. Adams and J. J. F. Fournier (2003). *Sobolev Spaces*. 2nd ed. Elsevier/Academic Press, Amsterdam.

W. Bangerth, R. Hartmann, and G. Kanschat. *deal.II Differential Equations Analysis Library, Technical Reference*. URL: http://www.dealii.org.

D. Braess (2007). *Finite Elements*. 3rd ed. Cambridge University Press, Cambridge.

J. H. Bramble and S. R. Hilbert (1970). *Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation.* SIAM J. Numer. Anal. 7, pp. 112–124. DOI: 10.1137/0707006.

S. C. Brenner and L. R. Scott (2008). *The Mathematical Theory of Finite Element Methods*. 3rd ed. Vol. 15. Texts in Applied Mathematics. Springer, New York.

K. Chrysafinos and N. J. Walkington (2006). *Error estimates for the discontinuous Galerkin methods for parabolic equations*. SIAM J. Numer. Anal. 44.1, 349–366 (electronic).

P. G. Ciarlet (2002). *The Finite Element Method for Elliptic Problems*. Vol. 40. Classics in Applied Mathematics. Reprint of the 1978 original [North-Holland, Amsterdam]. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

R. Courant (1943). *Variational methods for the solution of problems of equilibrium and vibrations*. Bull. Amer. Math. Soc. 49, pp. 1–23.

D. A. Di Pietro and A. Ern (2012). *Mathematical Aspects of Discontinuous Galerkin Methods*. Vol. 69. Mathématiques et Applications. Springer, New York.

A. Ern and J.-L. Guermond (2004). *Theory and Practice of Finite Elements*. Vol. 159. Applied Mathematical Sciences. Springer-Verlag, New York.

L. C. Evans (2010). *Partial Differential Equations*. 2nd ed. Vol. 19. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI.

P. Grisvard (1985). *Elliptic Problems in Nonsmooth Domains*. Vol. 24. Monographs and Studies in Mathematics. Pitman (Advanced Publishing Program), Boston, MA.

J. Hake, A. Logg, G. N. Wells, et al. *DOLFIN: A C++/Python finite element library*. URL: http://fenicsproject.org.

O. A. Ladyzhenskaya and N. N. Ural'tseva (1968). *Linear and Quasilinear Elliptic Equations*. Translated from the Russian by Scripta Technica, Inc. Translation editor: Leon Ehrenpreis. Academic Press, New York.

R. Rannacher (2008). “Numerische Mathematik 2”. Lecture notes. URL: http://numerik.iwr.uni-heidelberg.de/~lehre/notes/num2/numerik2.pdf.

M. Renardy and R. C. Rogers (2004). *An Introduction to Partial Differential Equations*. 2nd ed. Vol. 13. Texts in Applied Mathematics. Springer-Verlag, New York.

R. E. Showalter (1997). *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*. Vol. 49. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI.

E. Süli (2011). “Finite Element Methods for Partial Differential Equations”. Lecture notes. URL: http://people.maths.ox.ac.uk/suli/fem.pdf.

V. Thomée (2006). *Galerkin Finite Element Methods for Parabolic Problems*. 2nd ed. Vol. 25. Springer Series in Computational Mathematics. Springer-Verlag, Berlin.

G. M. Troianiello (1987). *Elliptic Differential Equations and Obstacle Problems*. The University Series in Mathematics. Plenum Press, New York.

J. Wloka (1987). *Partial Differential Equations*. Translated from the German by C. B. Thomas and M. J. Thomas. Cambridge University Press, Cambridge.

E. Zeidler (1995a). *Applied Functional Analysis*. Vol. 108. Applied Mathematical Sciences. Applications to mathematical physics. Springer-Verlag, New York.

E. Zeidler (1995b). *Applied Functional Analysis*. Vol. 109. Applied Mathematical Sciences. Main principles and their applications. Springer-Verlag, New York.

M. Zlámal (1968). *On the finite element method*. Numer. Math. 12, pp. 394–409.