NUMERICAL PARTIAL DIFFERENTIAL EQUATIONS

Lecture Notes, Summer 2021

Christian Clason

August 19, 2021

Institute of Mathematics and Scientific Computing
University of Graz

CONTENTS

I BACKGROUND

II CONFORMING FINITE ELEMENTS

III NONCONFORMING FINITE ELEMENTS

IV TIME-DEPENDENT PROBLEMS

PREFACE

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 in [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, starting with [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] D. Braess (2007), *Finite Elements*, 3rd ed., Cambridge University Press, Cambridge, doi: 10.1017/cbo9780511618635

[2] D. Boffi, F. Brezzi & M. Fortin (2013), *Mixed and Finite Element Methods and Applications*, vol. 44, Springer Series in Computational Mathematics, Springer, New York, doi: 10.1007/978-3-642-36519-5

[3] S. C. Brenner & L. R. Scott (2008), *The Mathematical Theory of Finite Element Methods*, 3rd ed., vol. 15, Texts in Applied Mathematics, Springer, New York, doi: 10.1007/978-0-387-75934-0

[4] A. Ern & J.-L. Guermond (2004), *Theory and Practice of Finite Elements*, vol. 159, Applied Mathematical Sciences, Springer, New York, doi: 10.1007/978-1-4757-4355-5

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

[6] V. Thomée (2006), *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., vol. 25, Springer Series in Computational Mathematics, Springer, Berlin, doi: 10.1007/3-540-33122-0

Part I

BACKGROUND

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 ELLIPTIC PDES

Consider for a given function $f : (0,1) \to \mathbb{R}$ the solution $u : (0,1) \to \mathbb{R}$ of the two-point boundary value problem

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

The idea is to pass from (BVP) to a system of linear equations – which can be solved on a computer – by projection onto a finite-dimensional subspace. Any projection requires some kind of inner product, which we introduce now. We begin by multiplying the differential equation with any sufficiently regular *test 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

$$V := \{v : (0,1) \to \mathbb{R} \text{ integrable} : a(v,v) < \infty,\ v(0) = 0\}.$$

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 a discontinuous right-hand side f. Since then the second derivative of u is discontinuous, u is not necessarily 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

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

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 III.)

We first have to ask whether (W_S) has a unique solution. 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 that $u_S = \sum_{i=1}^n U_i\varphi_i \in S$, $U_i \in \mathbb{R}$ for $i = 1, \dots, n$, satisfies

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

This is now a system of linear equations for the unknown coefficients U_i. If we define

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

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 $\mathbf{KV} = 0$ implies $\mathbf{V} = 0$. To show this, we set $v_S := \sum_{i=1}^n V_i\varphi_i \in S$. Then,

$$0 = \mathbf{KV} = (a(v_S, \varphi_1), \dots, a(v_S, \varphi_n))^T$$

implies that

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

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

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 $\mathbf{V}^T\mathbf{K}\mathbf{V} = a(v_S, v_S) > 0$ 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 $v_S \in S$, we obtain

$$a(u - u_S, v_S) = 0 \qquad \text{for all } v_S \in S. \tag{1.1}$$

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*

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

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

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

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

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

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

$$\|u - u_S\|_E \le \inf_{v_S \in S} \|u - v_S\|_E \,,$$

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:

$$\|u - u_S\|_E = \min_{v_S \in S} \|u - v_S\|_E \,. \tag{1.2}$$

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*

$$\begin{cases} -w''(x) = u(x) - u_S(x) & \text{for } x \in (0,1), \\ w(0) = 0, \qquad w'(1) = 0. \end{cases} \tag{1.3}$$

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

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

Dividing by $\|u - u_S\|_{L^2} = \|w''\|_{L^2}$ from (1.3) and taking the infimum over all $v_S \in S$ yields

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

To continue, we require an *approximation property* for S: There exists a constant $c_S > 0$ such that

$$\inf_{v_S \in S} \|g - v_S\|_E \leq c_S \, \|g''\|_{L^2} \tag{1.4}$$

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

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

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. Of course, we wish that c_S can be made arbitrarily small by choosing S sufficiently large. 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*)

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

for $1 \le i \le n$, which satisfy $\varphi_i(0) = 0$ and hence $\varphi_i \in S$. Furthermore,

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

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

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

For $v_S \in S$, the interpolation error $v_S - (v_S)_I$ is piecewise linear as well, and since $(v_S)_I(x_i) = v_S(x_i)$ for all $1 \le i \le n$, this implies that $v_S - (v_S)_I \equiv 0$. Any $v_S \in S$ can thus be written as a unique 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 (i.e., $\mathcal{I} \circ \mathcal{I} = \mathcal{I}$).

We are now in a position to prove the approximation property (1.4) of S. Let

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

denote the *mesh size*. Since the best approximation error is certainly not bigger than the interpolation error, it suffices to show that there exists a constant $C > 0$ such that for all sufficiently smooth $u \in V$,

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

We now consider this error separately on each *element* $[x_{i-1}, x_i]$, i.e., we show that

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

First, 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*), the previous estimate is equivalent to

$$\int_0^1 \tilde{e}'(t)^2 \, dt \le C^2 \int_0^1 \tilde{e}''(t)^2 \, dt. \tag{1.5}$$

(This is an elementary version of *Poincaré's inequality*). Since u_I is the nodal interpolant of u, the error satisfies $e(x_{i-1}) = e(x_i) = 0$. In addition, u_I is linear and u continuously differentiable on $[x_{i-1}, x_i]$. Hence, $\tilde{e}$ is continuously differentiable on $[0,1]$ with $\tilde{e}(0) = \tilde{e}(1) = 0$, and Rolle's theorem yields 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 $a > b$)

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

We can now use the Cauchy–Schwarz inequality to estimate

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

Integrating both sides with respect to y and taking the supremum over all $\xi \in (0,1)$ yields (1.5) with

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

Summing over all elements and estimating h_i by h shows the approximation property (1.4) for S with $c_S := Ch$. For this choice of S, the solution u_S of (W_S) satisfies

$$\|u - u_S\|_E \leq \min_{v_S \in S} \|u - v_S\|_E \leq \|u - u_I\|_E \leq Ch\,\|u''\|_{L^2}$$

as well as

$$\|u - u_S\|_{L^2} \leq C^2 h^2\,\|u''\|_{L^2}\,. \tag{1.6}$$

These are called *a priori estimates*, since they only require knowledge of the given data $f = u''$ but not of the solution u_S. They tell 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, which represents the fact that it is more difficult to control errors in the derivative than errors in the function value.

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 $\mathbf{U}$. 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.,

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

and the contributions from a single element for all (i, j) are computed simultaneously. Here we can exploit that by its 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 $k - 1 \leq i, j \leq k$. 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, e.g., by the method of conjugate gradients (since $\mathbf{K}$ is also symmetric and positive definite).

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 $\varphi_k(x) = \hat{\varphi}_2(\xi(x))$. Using $\xi'(x) = (x_k - x_{k-1})^{-1} = h_k^{-1}$, the integrals for $i, j \in \{k - 1, k\}$ can therefore be computed via

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

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*.) Since the derivatives of $\hat{\varphi}_1, \hat{\varphi}_2$ are constant, the contribution from the element $[x_{k-1}, x_k]$ to $K_{ij} = a(\varphi_i, \varphi_j)$ for $i, j \in \{k - 1, k\}$ (the contribution for all other pairs (i, j) being zero) 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

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

The elements $M_{ij} := \left(\varphi_i, \varphi_j\right)$ 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 $\mathbf{K}$. Setting $\mathbf{f} := (f(x_1), \dots, f(x_n))^T$, the right-hand side of the linear system is then given by $\mathbf{F} = \mathbf{Mf}$.

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

Algorithm 1.1 Finite element method in 1D

```
Require: $0 = x_0 < \cdots < x_n = 1$, $F := (f(x_0), \dots, f(x_n))^T$
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 $KU = MF$
Ensure: $U$
```

1.5 A POSTERIORI ERROR ESTIMATES AND ADAPTIVITY★

The a priori estimate (1.6) 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 are localized error estimates for each element but involve the computed solution u_S. This gives 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

$$\begin{cases} -w''(x) = u(x) - u_S(x) & \text{for } x \in (0,1), \\ w(0) = 0, \qquad w'(1) = 0, \end{cases} \tag{1.7}$$

and proceed as before, yielding

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

for all $v_S \in S$. We now choose $v_S = w_I \in S$, 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_S 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

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

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 a priori error estimate (1.6)) that

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

holds, from which we obtain

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

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

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

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 $\sum_{i=1}^{n^{m+1}} (h_i^{(m)})^2 \left\| R_{h^{(m)}} \right\|_{L^2(x_{i-1}^{(m)}, x_i^{(m)})} \geq \tau$ 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 $m \longleftarrow m + 1$
end while
```

There are different strategies to choose the new mesh. A common requirement is that the strategy 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, as well as *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 ELLIPTIC 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 & 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 $\mathbb{R}^n$, $n \in \mathbb{N}$. We recall that for $1 \le p \le \infty$,

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

with

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

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*

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

if $p^{-1} + q^{-1} = 1$ (with $\infty^{-1} := 0$). For bounded Ω, this implies that $L^p(\Omega) \hookrightarrow L^q(\Omega)$ for $p \ge q$. 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 $|\alpha| \leq k$. 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

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

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 compact subset of Ω, as well as

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

(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 $Df(x) = \text{sign}(x)$, 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 $1 \leq p \leq \infty$:

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

which are Banach spaces when endowed with the norm

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

We shall also use the corresponding semi-norms

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

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 – if not otherwise stated – 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 $n = 3$.) Similarly, a C^m *boundary* can be parametrized by a finite set of m times continuously differentiable functions. A fundamental result is then the following approximation property (which does not hold for arbitrary domains).

Theorem 2.1 (density[1]). *For $1 \leq p < \infty$ 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.*

Using a density argument, one can show that Sobolev spaces behave well under sufficiently smooth coordinate transformations.

Theorem 2.2 (coordinate transformation[2]). *Let $\Omega, \Omega' \subset \mathbb{R}^n$ be two domains, and $T : \Omega \to \Omega'$ be a k-diffeomorphism (i.e., T is a bijection, T and its inverse T^{-1} are continuous with k bounded and continuous derivatives on $\overline{\Omega}$ and $\overline{\Omega}'$, and the determinant of the Jacobian of T is uniformly bounded from above and below). Then the mapping $v \mapsto v \circ T$ is bounded from $W^{k,p}(\Omega)$ to $\mathrm{W}^{k,p}(\Omega')$ and has a bounded inverse.*

[1] The key result was shown by Meyers and Serrin in a paper 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, Theorem 3], [Adams & Fournier 2003, Theorem 3.17]

[2] e.g.,[Adams & Fournier 2003, Theorem 3.41]

Corresponding chain rules for weak derivatives can be obtained from the classical ones using a density argument as well. Theorem 2.2 can also be used to define Sobolev spaces on (sufficiently smooth) manifolds via a local coordinate charts. In particular, if Ω has a C^k boundary, $k \geq 1$, we can define $W^{k,p}(\partial\Omega)$ by (local) transformation to $W^{k,p}(D)$, where $D \subset \mathbb{R}^{n-1}$.

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

Theorem 2.3 (Sobolev[3], Rellich–Kondrachov[4] embedding). *Let* $1 \leq p, q < \infty$ *and* $\Omega \subset \mathbb{R}^n$ *be a bounded open set with Lipschitz boundary. Then the following embeddings are continuous:*

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

Moreover, the following embeddings are compact:

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

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

We can also ask if conversely, continuous functions are weakly differentiable. Intuitively, this is the case if the points of (classical) non-differentiability form a set of Lebesgue measure zero. Indeed, continuous and piecewise differentiable functions are weakly differentiable.

Theorem 2.4. *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.,* $\overline{\Omega} = \bigcup_{j=1}^N \overline{\Omega}_j$ *and* $\Omega_i \cap \Omega_j = \emptyset$ *for all* $i \neq j$*). Then for every* $k \geq 1$ *and* $1 \leq p \leq \infty$,

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

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 $1 \leq j \leq N$. We need to show that $\partial_i v$ exists as a weak derivative for all $1 \leq i \leq n$ and that $\partial_i v \in L^p(\Omega)$. An obvious candidate is

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

[3] e.g., [Evans 2010, § 5.6], [Adams & Fournier 2003, Theorem 4.12]

[4] e.g., [Evans 2010, § 5.7], [Adams & Fournier 2003, Theorem 6.3]

for arbitrary $c \in \mathbb{R}$. By the embedding $C^0(\overline{\Omega}_j) \hookrightarrow L^\infty(\Omega_j)$ and the boundedness of Ω, we have that $w_i \in L^p(\Omega)$ for any $1 \leq p \leq \infty$. 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 $\varphi \in C_0^\infty(\overline{\Omega})$

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

where $\nu_j = ((\nu_j)_1, \dots, (\nu_j)_n)$ is the outer normal vector to 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 approximating continous functions) which maps a function f on a bounded domain $\Omega \subset \mathbb{R}^n$ to a function Tf on $\partial\Omega$.

Theorem 2.5 (trace theorem[5]). *Let $kp < n$ and $p \leq q \leq (n-1)p/(n-kp)$, and $\Omega \subset \mathbb{R}^n$ be a bounded open set with C^m boundary or a polygon in $\mathbb{R}^2$. Then $T : W^{k,p}(\Omega) \to L^q(\partial\Omega)$ is a bounded linear operator, i.e., there exists a constant $C > 0$ depending only on p and Ω such that for all $f \in W^{k,p}(\Omega)$,*

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

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

This implies (although it is not obvious)[6] 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 2.6 (Poincaré's inequality[7]). *Let $1 \leq p < \infty$ and let Ω be a bounded open set. Then there exists a constant $c_\Omega > 0$ depending only on Ω and p such that for all $f \in W_0^{1,p}(\Omega)$,*

$$\|f\|_{W^{1,p}(\Omega)} \leq c_\Omega |f|_{W^{1,p}(\Omega)}.$$

[5] e.g., [Evans 2010, § 5.5], [Adams & Fournier 2003, Theorem 5.36], [Grisvard 2011, Theorem 1.5.2.8]

[6] e.g., [Evans 2010, § 5.5, Theorem 2], [Adams & Fournier 2003, Theorem 5.37]

[7] e.g, [Adams & Fournier 2003, Corollary 6.31]

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 $f \in W^{1,p}(\Omega)$, $1 \leq p \leq \infty$, for which $D^\alpha f = 0$ almost everywhere in Ω for all $|\alpha| = 1$ must be constant (cf. Lemma 5.1).

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

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

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

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

for the semi-norm on $H^1(\Omega)$ (which, by the Poincaré inequality (Theorem 2.6), is equivalent to the 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

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

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 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 2.4 guarantees that $S \subset V$.

2.2 WEAK SOLUTION OF ELLIPTIC PDES

In the first two parts, 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

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

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

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

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

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

where g should be understood in the sense of traces, i.e., as Tg. 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 $u \in V$ – for a suitably chosen function space 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 $g = 0$ (a *homogeneous* Dirichlet condition), 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

$$a(u, v) := \sum_{j,k=1}^{n} \left(a_{jk}\partial_j u, \partial_k v\right) + \sum_{j=1}^{n} \left(b_j\partial_j u, v\right) + (cu, v) = (f, v) \tag{2.5}$$

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),[8] 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

$$a(\tilde{u}, v) = (f, v) - a(u_g, v)$$

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

[8] [Renardy & Rogers 2004, Theorem 7.40]

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 equation in the boundary integral in (2.4) and take $V = H^1(\Omega)$. We then look for $u \in H^1(\Omega)$ satisfying

$$a(u,v) = (f,v) + (g,v)_{\partial\Omega} \tag{2.6}$$

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

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

$$a_R(u,v) := a(u,v) + (du,v)_{\partial\Omega} = (f,v) + (g,v)_{\partial\Omega} \tag{2.7}$$

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

$$a(u,v) = F(v), \qquad \text{for all } v \in V. \tag{2.8}$$

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 2.7 (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 $c_1 > 0$ such that*

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

for all $v \in V$.

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

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

for all $v, w \in V$.

Then there exists a unique solution $u \in V$ to (2.8), and

$$\|u\|_V \leq \frac{1}{c_1} \|F\|_{V^*}. \tag{2.9}$$

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,[9] there exist unique $\varphi_u, \varphi_F \in V$ such that

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

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, a solution $u \in V$ satisfies

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

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.[10] For $\delta > 0$, consider the mapping

$$T_\delta : V \to V, \qquad T_\delta(v) = v - \delta(\varphi_v - \varphi_F).$$

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

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

We can thus choose $0 < \delta < 2\frac{c_1}{c_2}$ 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 (2.9), 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

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

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.

[9] e.g., [Zeidler 1995a, Theorem 2.E]

[10] e.g., [Zeidler 1995a, Theorem 1.A]

Theorem 2.8 (well-posedness). *Let $a_{jk} \in L^\infty(\Omega)$ satisfy the ellipticity condition (2.3) with constant $\alpha > 0$, let $b_j, c \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ be given, and set $\beta = \alpha^{-1} \sum_{j=1}^n \left\|b_j\right\|_{L^\infty(\Omega)}^2$.*

a) The homogeneous Dirichlet problem has a unique solution $u \in H_0^1(\Omega)$ if

$$c(x) - \frac{\beta}{2} \geq 0 \qquad \text{for almost all } x \in \Omega.$$

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

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

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

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

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

$$c(x) - \frac{\beta}{2} \geq \gamma > 0 \qquad \text{for almost all } x \in \Omega.$$

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

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

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

$$\begin{aligned} c(x) - \frac{\beta}{2} \geq \gamma \geq 0 \qquad & \text{for almost all } x \in \Omega, \\ d(x) \geq \delta \geq 0 \qquad & \text{for almost all } x \in \partial\Omega, \end{aligned}$$

and either $\gamma > 0$ or $\delta > 0$. In this case, there exists a $C > 0$ such that

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

Proof. We 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 other cases being similar). Let $v \in H_0^1(\Omega)$ be given. First, the ellipticity of a_{jk} implies that

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

We then have by Young's inequality $ab \leq \frac{\alpha}{2}a^2 + \frac{1}{2\alpha}b^2$ for $a = |v|_{H^1(\Omega)}$, $b = \|v\|_{L^2(\Omega)}$ and $\alpha > 0$ as well as repeated application of Hölder's inequality that

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

Under the assumption that $c - \frac{\beta}{2} \geq 0$, the second term is non-negative and we deduce using Poincaré's inequality that

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

for $C := \alpha/(4 + 4c_\Omega^2)$, where c_Ω is the constant from Poincaré's inequality. □

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 $\beta = \sum_{j=1}^n \left\|\partial_j b_j\right\|_{L^\infty(\Omega)}$.

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 results which will be relevant in the following.

Theorem 2.9 (higher regularity[11]). *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with C^{k+1} boundary, $k \geq 0$, $a_{jk} \in C^k(\overline{\Omega})$ 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 $H^{k+2}(\Omega) \cap H_0^1(\Omega)$, and there exists a $C > 0$ such that*

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

Theorem 2.10 (higher regularity[12]). *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 $b_j, c \in C^0(\overline{\Omega})$. If $f \in L^{(}(\Omega)$, then the solution of the homogeneous Dirichlet problem is in $H^2(\Omega)$, and there exists a $C > 0$ such that*

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

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.

[11] [Troianiello 1987, Theorem 2.24]

[12] [Grisvard 2011, Theorem 5.2.2], [Ladyzhenskaya & Ural'tseva 1968, pp. 169–189]

Part II

CONFORMING FINITE ELEMENTS

3 GALERKIN APPROACH FOR ELLIPTIC PROBLEMS

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

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

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

$$a(v,v) \geq c_1 \|v\|_V^2, \tag{3.1}$$

$$a(u,v) \leq c_2 \|u\|_V \|v\|_V, \tag{3.2}$$

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]

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

Since we have chosen a closed $V_h \subset V$, the subspace V_h is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_V$ and norm $\|\cdot\|_V$. 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 2.7, for any closed subspace $V_h \subset V$, there exists a unique solution $u_h \in V_h$ of (W$_h$) satisfying*

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

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

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

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,*

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

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 $v = v_h \in V_h$) the *Galerkin orthogonality*

$$a(u - u_h, v_h) = 0 \qquad \text{for all } v_h \in V_h. \tag{3.3}$$

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

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

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 (1.2) 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 coercive and 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 bilinearity and symmetry of a. Assume now that u satisfies $a(u, v) - F(v) = 0$ for all $v \in V$. Then setting $t = 1$, we deduce that for all $v \ne 0$,

$$J(u + v) = J(u) + \tfrac{1}{2} a(v, v) \ge J(u) + \frac{c_1}{2} \|v\|_V^2 > J(u).$$

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 that

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

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 $-F(v)$ represents its potential energy under external load.)

Arguing as in Section 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. Using the equivalence of norms, this implies that the infimum in Lemma 3.2 is attained for symmetric bilinear forms. 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 $(\cdot, \cdot)_H$ and V be a closed subspace satisfying the conditions of the Lax–Milgram theorem theorem such that the embedding $V \hookrightarrow H$ is continuous (e.g., $V = H^1(\Omega) \hookrightarrow L^2(\Omega) = H$). 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). *For any $g \in H$, let φ_g be the unique solution of the* adjoint problem

$$a(w, \varphi_g) = (g, w)_H \quad \text{for all } w \in V. \tag{3.4}$$

Then there exists a $C > 0$ such that

$$\|u - u_h\|_H \leq C \|u - u_h\|_V \sup_{g \in H \setminus \{0\}} \left(\frac{1}{\|g\|_H} \inf_{v_h \in V_h} \|\varphi_g - v_h\|_V \right).$$

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

$$\|w\|_H = \sup_{g \in H} \frac{(g, w)_H}{\|g\|_H}. \tag{3.5}$$

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

$$\begin{aligned}(g, u - u_h)_H &= a(u - u_h, \varphi_g)\\ &= a(u - u_h, \varphi_g - v_h)\\ &\leq c_2 \left\|u - u_h\right\|_V \left\|\varphi_g - v_h\right\|_V .\end{aligned}$$

Inserting $w = u - u_h$ into (3.5), we thus obtain

$$\begin{aligned}\left\|u - u_h\right\|_H &= \sup_{g \in H} \frac{(g, u - u_h)_H}{\left\|g\right\|_H}\\ &\leq c_2 \left\|u - u_h\right\|_V \sup_{g \in H \setminus \{0\}} \frac{\left\|\varphi_g - v_h\right\|_V}{\left\|g\right\|_H}\end{aligned}$$

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

Note that the existence of a unique solution of the adjoint problem is an assumption here that needs to be verified. If a is symmetric, this is guaranteed by the Lax–Milgram theorem. Otherwise, both the original and the adjoint problem need to satisfy the conditions of the Lax–Milgram theorem (which is the case, e.g., for constant coefficients b_j).

4 FINITE ELEMENT SPACES

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 transformed copies of the reference element to partition the given domain and discuss how to construct a global interpolant from 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. A *finite element* is a triple $(K, \mathcal{P}, \mathcal{N})$ where

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

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

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

Here $\mathcal{P}^*$ denotes the algebraic dual of $\mathcal{P}$, i.e., the space of linear functionals on $\mathcal{P}$. 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

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

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. A basis $\{\psi_1, \dots, \psi_d\}$ of $\mathcal{P}$ is called *dual basis* or *nodal basis* to $\mathcal{N}$ if $N_i(\psi_j) = \delta_{ij}$.

For example, for the linear finite elements in one dimension, $K = (0, 1)$, $\mathcal{P} = P_1$ 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 $1 \le i \le d$, 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, $1 \le i \le d$, 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 $1 \le i \le d$. Define the (square) *Vandermonde 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 $\mathbf{B}\mathbf{a} = \mathbf{L}$ 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

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

implying $v = 0$, or, in matrix form, that $\mathbf{B}^T\mathbf{b} = 0$ implies $0 = \mathbf{b} := (\beta_1, \dots, \beta_d)^T$, i.e., that B^T is injective. But this too is equivalent to the fact that $\mathbf{B}$ is invertible (since any square matrix is invertible if and only if it is surjective). □

Note that (b) in particular implies that the interpolation problem using functions in $\mathcal{P}$ with interpolation conditions $\mathcal{N}$ is uniquely solvable. To construct a finite element, one usually proceeds in the following way:

1. choose an element domain K (e.g., a triangle),
2. choose a polynomial space $\mathcal{P}$ of a given degree k (e.g., linear functions),
3. choose d degrees of freedom $\mathcal{N} = \{N_1, \dots, N_d\}$, where d is the dimension of $\mathcal{P}$, such that the corresponding interpolation problem has a unique solution,
4. compute the nodal basis of $\mathcal{P}$ with respect to $\mathcal{N}$.

The last step amounts to solving for $1 \leq j \leq d$ the concrete interpolation problems $N_i(\psi_j) = \delta_{ij}$, e.g., using the Vandermonde matrix. A useful tool to verify the unique solvability of the interpolation problem for polynomials is the following lemma, which is a multidimensional form of polynomial division. Recall that for multivariate polynomials, the *(total) degree* is the maximal sum of all occuring powers in a term (e.g., $p(x) = x_1x_2^2$ has degree 3). It is convenient to write such a polynomial p of degree k on $\mathbb{R}^n$ as $p(x) = \sum_{|\alpha|\leq k} c_\alpha x^\alpha$ using a *multi-index* $\alpha \in \mathbb{N}_0^{n-1}$ with the convention that $x^\alpha := x_1^{\alpha_1} \cdot x_n^{\alpha_n}$ and $|\alpha| := \sum_{i=1}^n \alpha_i$.

Lemma 4.4. *Let $L \neq 0$ be a linear-affine functional on $\mathbb{R}^n$ and P be a polynomial of total degree $d \geq 1$ with $P(x) = 0$ for all x with $L(x) = 0$. Then there exists a polynomial Q of total 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

$$P(\hat{x}, x_n) = \sum_{j=0}^{d} \left[\sum_{|\alpha|\leq d-j} c_{\alpha,j}\hat{x}^\alpha \right] x_n^j.$$

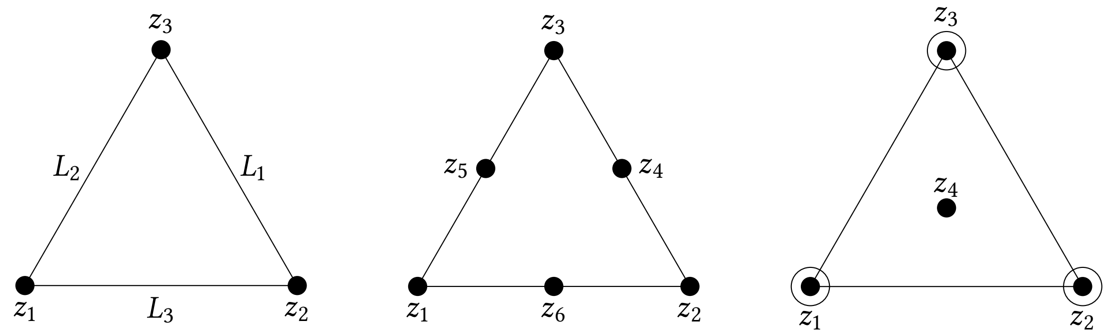
For $x_n = 0$, this implies that

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

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

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

where Q is of degree $d - 1$. □



(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.

4.2 EXAMPLES OF FINITE ELEMENTS

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

Triangular elements Let K be a triangle and

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

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 Fig. 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 (non-constant) 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, e.g., $v = cL_1$. Now let z_1 be the vertex not on the edge defined by L_1. Then

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

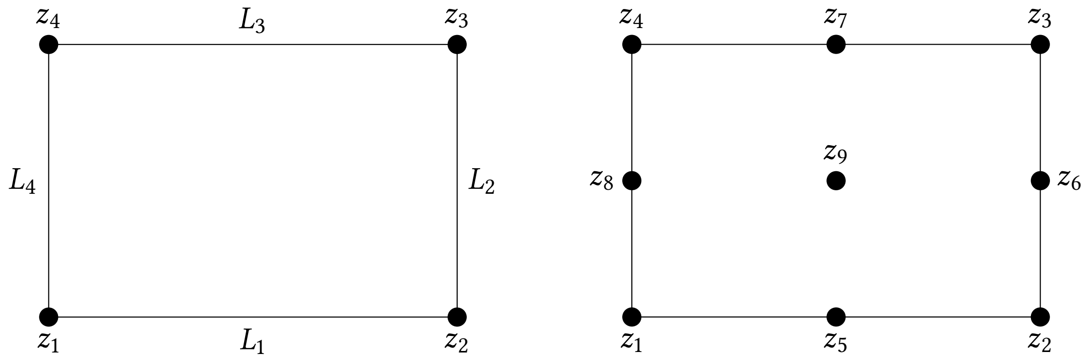
  Since $L_1(z_1) \neq 0$ (otherwise the linear functional L_1 would be identically zero), this implies $c = 0$ and thus $v = 0$.

- *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 Fig. 4.1b). To show that condition (iii) holds, we argue as above. Let $v \in P_2$ vanish at z_i, $1 \leq i \leq 6$. 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 functional 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 $v = L_1 Q_1$. Now consider one of the remaining edges with corresponding functional, e.g., L_2. Since $v(z_5) = v(z_6) = 0$ by assumption and L_2 cannot be zero there (otherwise it would be constant), we have that $Q_1(z_5) = Q_1(z_6) = 0$, i.e., Q_1 is a linear polynomial on this edge with two roots and hence vanishes. Applying Lemma 4.4 to Q_1, we thus obtain a constant c such that $v = L_1 Q_1 = c L_1 L_2$. Taking the midpoint of the remaining edge, z_6, we have
$$0 = v(z_6) = c L_1(z_6) L_2(z_6),$$
and since neither L_1 nor L_2 are zero in 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 ten suitable points, we take N_i, $1 \leq i \leq 4$ 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 Fig. 4.1c) and take the remaining nodal variables as gradient evaluations:
$$N_{i+4}(v) = \partial_1 v(z_i), \qquad N_{i+7} = \partial_2 v(z_i), \quad 1 \leq i \leq 3.$$
Now we again consider $v \in P_3$ with $N_i(v) = 0$ for all $1 \leq i \leq 10$. On each edge, v is a cubic polynomial with double roots at each vertex, and hence must vanish. By considering successively each edge, we find that $v = c L_1 L_2 L_3$ which implies that
$$0 = v(z_4) = c L_1(z_4) L_2(z_4) L_3(z_4)$$
and hence $c = 0$ since the barycenter z_4 lies on neither of the edges. Therefore, $v = 0$.

The interpolation points z_i are called *nodes* (not to be confused with the *vertices* defining the element domain). Both types of elements can be defined for arbitrary degree k. It should be clear from the above that our 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
$$Q_k = \left\{ \sum_j c_j p_j(x_1) q_j(x_2) : c_j \in \mathbb{R}, p_j, q_j \in P_k \right\}$$



(a) bilinear Lagrange element

(b) biquadratic Lagrange element

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

of products of univariate polynomials of degree up to k, which has dimension $(k+1)^2$ (e.g., $Q_2 = \operatorname{span}\{1, x_1, x_2, x_1x_2, x_1^2x_2, x_1x_2^2, x_1^2, x_2^2\}$). By the same 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 Fig. 4.2a).
- *Biquadratic Lagrange elements*: Let $k = 2$ and take $\mathcal{P} = Q_2$ (hence the dimension of $\mathcal{P}$ and $\mathcal{P}^*$ is 9) and $\mathcal{N} = \{N_1, \dots, N_9\}$ 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 Fig. 4.2b).

The above construction is easy to generalize for arbitrary k and n: Let $t_1, \dots, t_{k+1}$ be distinct points on (say) $[0, 1]$ with $t_1 = 0$ and $t_{k+1} = 1$. Then the nodes $z_1, \dots, z_d$ for the rectangular Lagrange element on $K = [0, 1]^n$ are given by the *tensor product*

$$\left\{(t_{i_1}, \dots, t_{i_n}) : i_j = 1, \dots, k+1 \text{ for } j = 1, \dots, n\right\}.$$

This straightforward construction is the main advantage of rectangular elements; on the other hand, triangular elements give more flexibility for handling complicated domains.

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 $\{\psi_1, \dots, \psi_d\}$ be the corresponding nodal basis of $\mathcal{P}$. For a given function v such that $N_i(v)$ is defined for all $1 \leq i \leq d$, the *local interpolant* of v is defined as

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

The local interpolant can be explicitly constructed once the nodal 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*

1. *the mapping $v \mapsto \mathcal{I}_K$ is linear;*
2. *$N_i(\mathcal{I}_K v)) = N_i(v)$, $1 \leq i \leq d$;*
3. *$\mathcal{I}_K(v) = v$ for all $v \in \mathcal{P}$, i.e., $\mathcal{I}_K$ is a projection.*

Proof. The claim (i) follows directly from the linearity of the N_i. For (ii), 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 $1 \leq i \leq d$ and arbitrary v. This implies that $N_i(v - \mathcal{I}_K v) = 0$ for all $1 \leq i \leq d$, and hence by Lemma 4.3 that $\mathcal{I}_K v = v$ and hence (iii) holds. □

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

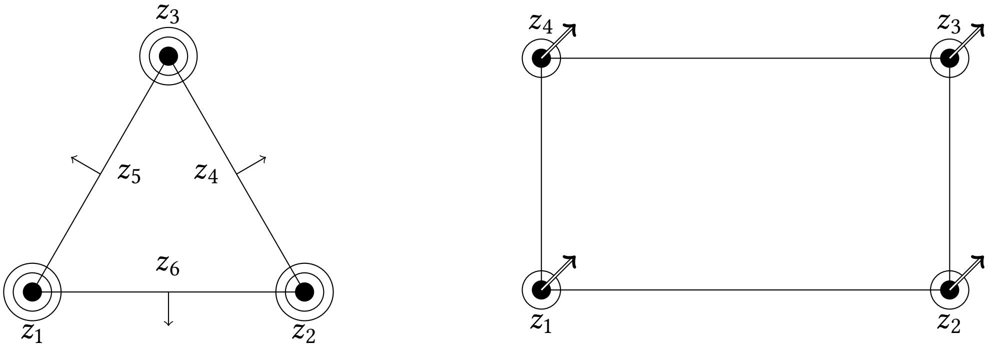
Definition 4.7. A *subdivision* of a bounded open set $\Omega \subset \mathbb{R}^n$ is a finite collection $\mathcal{T}$ of open sets K_i such that

(i) $K_i \cap K_j = \emptyset$ if $i \neq j$;

(ii) $\bigcup_i \overline{K_i} = \overline{\Omega}$.

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

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



(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}.

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;

(ii) no vertex of any triangle lies on an edge of another triangle (i.e., no *hanging nodes*).

Similar conditions can be given for $n \geq 3$ (tetrahedra, simplices), in which case one usually also speaks of triangulations. Note that this supposes that Ω is polyhedral itself. (For non-polyhedral domains, it is possible to use curved elements near the boundary.)

Definition 4.10. A global interpolant $\mathcal{I}_\mathcal{T}$ has *continuity order m* (in short, "is C^m") if $\mathcal{I}_\mathcal{T} v \in C^m(\overline{\Omega})$ for all $v \in C^m(\overline{\Omega})$ (for which the interpolation is well-defined). In this case, the space

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

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

In particular, 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.

Theorem 4.11. *The triangular Lagrange and Hermite elements of fixed degree are all C^0 elements (i.e., lead to C^0 finite element space). 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 polynomial 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, Fig. 4.3a) and the Bogner–Fox–Schmit rectangle (a bicubic Hermite element of dimension 16, Fig. 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., $T : \hat{x} \mapsto A\hat{x} + b$ 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) $K = \left\{A\hat{x} + b : \hat{x} \in \hat{K}\right\}$,

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

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

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

It is a straightforward exercise to show that the nodal bases of $\hat{\mathcal{P}}$ and $\mathcal{P}$ are related by $\hat{\psi}_i = \psi_i \circ T$. 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 rectangular elements. Non-affine equivalent elements (such as *isoparametric elements*[2]) are useful in treating elements with curved boundaries (for non-polyhedral domains)

The advantage of this construction is that affine equivalent elements are also interpolation equivalent in the following sense.

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_K. Then,*

$$\mathcal{I}_{\hat{K}}(v \circ T_K) = (\mathcal{I}_K v) \circ T_K.$$

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

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

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

[2] see, e.g., [Braess 2007, § III.2]

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 distance of the finite element solution to the true solution is determined by the distance to the best approximation by piecewise polynomials, which in turn is bounded by the distance to the corresponding interpolant. 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 & 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]

The 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 also $D^\beta D^\alpha v = 0 \in L^p(\Omega)$ for any multi-index β. Hence, $v \in \bigcap_{k=1}^\infty W^{k,p}(\Omega)$. The Sobolev embedding Theorem 2.3 thus guarantees that $v \in C^k(\Omega)$ for all $k \in \mathbb{N}$. The claim then follows using classical (pointwise) arguments, e.g., by Taylor series expansion. □

The next result concerns moment interpolation 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*

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

[1] see, e.g., [Süli 2011, § 3.2], [Brenner & Scott 2008, Chapter 4]

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

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

It thus remains to show that the quadratic matrix

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

is non-singular, which we do by showing injectivity. Consider $\xi = (\xi_\beta)_{|\beta|\leq k-1}$ such that $\mathbf{M}\xi = 0$. This implies that the corresponding polynomial q satisfies

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

Inserting in turn for α all possible multi-indices in descending (lexicographical) order (such that $D^\alpha x^\beta$ is constant) yields $\xi_\beta = 0$ for all $|\beta| \leq k-1$. Thus, $\mathbf{M}\xi = 0$ implies $\xi = 0$, and therefore $\mathbf{M}$ is invertible. □

The last lemma is a generalization of Poincaré's inequality.

Lemma 5.3. *Let $v \in W^{k,p}(\Omega)$ such that*

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

Then

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

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 $\{v_n\}_{n\in\mathbb{N}} \subset W^{k,p}(\Omega)$ of functions satisfying (5.2) and

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

Since the embedding $W^{k,p}(\Omega) \hookrightarrow W^{k-1,p}(\Omega)$ is compact by Theorem 2.3, there exists a subsequence (also denoted by $\{v_n\}_{n\in\mathbb{N}}$) converging in $W^{k-1,p}(\Omega)$ to a $v \in W^{k-1,p}(\Omega)$, i.e.,

$$\|v - v_n\|_{W^{k-1,p}(\Omega)} \to 0 \qquad \text{as } n \to \infty. \tag{5.5}$$

Since in addition $|v_n|_{W^{k,p}(\Omega)} \to 0$ by assumption (5.4), $\{v_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $W^{k,p}(\Omega)$ as well and thus converges in $W^{k,p}(\Omega)$ to a $\tilde{v} \in W^{k,p}(\Omega)$ which must satisfy $\tilde{v} = v$

(otherwise we would have a contradiction to (5.5)). By continuity, we then obtain that $|v|_{W^{k,p}(\Omega)} = 0$, and Lemma 5.1 yields that $v \in P_{k-1}$. Furthermore, v satisfies

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

by assumption (5.2), which as in the proof of Lemma 5.2 implies that $v = 0$. But this is a contradiction to

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

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) $|F(v)| \leq c_1 \|v\|_{W^{k,p}(\Omega)}$ for all $v \in W^{k,p}(\Omega)$ (boundedness),

(ii) $|F(u+v)| \leq c_2(|F(u)| + |F(v)|)$ for all $u, v \in W^{k,p}(\Omega)$ (sublinearity),

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

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

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

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

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

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

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

where c_0 is the constant appearing in (5.3) and we have used that $D^\alpha q = 0$ for $q \in P_{k-1}$ and all $|\alpha| = k$. This proves the claim with $c := c_0 c_1 c_2$. $\square$

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 with $P_{k-1} \subset \mathcal{P}$ for some $k \geq 1$ and all $N \in \mathcal{N}$ bounded on $W^{k,p}(K)$, $1 \leq p \leq \infty$. Then for any $v \in W^{k,p}(K)$,*

$$|v - \mathcal{I}_K v|_{W^{l,p}(K)} \leq c|v|_{W^{k,p}(K)} \quad \textit{for all} \quad 0 \leq l \leq k \tag{5.6}$$

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

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

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

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

$$T_K : \hat{K} \to K, \qquad \hat{x} \mapsto A_K \hat{x} + b_K \tag{5.7}$$

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 the estimate (5.6) transforms under T_K. For this, we recall that for sufficiently smooth v, the chain rule for weak derivatives is given by

$$\frac{\partial \hat{v}}{\partial \hat{x}_i} = \sum_{j=1}^{n} \frac{\partial v}{\partial x_j} \frac{\partial x_j}{\partial \hat{x}_i} = \sum_{j=1}^{n} (A_K)_{ij} \frac{\partial v}{\partial x_j}, \tag{5.8}$$

and the transformation rule for integrals by

$$\int_{T_K(\hat{K})} v \, dx = \int_{\hat{K}} (v \circ T_K) |\det(A_K)| \, d\hat{x}. \tag{5.9}$$

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

$$|\hat{v}|_{W^{k,p}(\hat{K})} \leq c \, \|A_K\|^k \, |\det(A_K)|^{-\frac{1}{p}} |v|_{W^{k,p}(K)}, \tag{5.10}$$

$$|v|_{W^{k,p}(K)} \leq c \left\|A_K^{-1}\right\|^k |\det(A_K)|^{\frac{1}{p}} |\hat{v}|_{W^{k,p}(\hat{K})}. \tag{5.11}$$

Proof. First, we have by Theorem 2.2 that $\hat{v} \in W^{k,p}(\hat{K})$. Let now α be a multi-index with $|\alpha| = k$, and let $\hat{D}^\alpha$ denote the corresponding weak derivative with respect to $\hat{x}$. Applying the chain and transformation rule, we obtain with a constant c depending only on n, k, and p that

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

Summing over all $|\alpha| = k$ yields (5.10). Arguing similarly using T_K^{-1} yields (5.11). □

We now derive a geometrical estimate of the quantities appearing in the right-hand side of (5.10) and (5.11). For a given element domain K, we define

- the *diameter* $h_K := \max_{x_1,x_2\in K} \|x_1 - x_2\|$,
- the *insphere diameter* $\rho_K := 2\max\{\rho > 0 : B_\rho(x) \subset K \text{ for some } x \in K\}$ (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 be an affine mapping defined as in (5.7) such that $K = T_K(\hat{K})$. Then*

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

Proof. The first property follows from the transformation rule (5.9) applied to the constant function $v \equiv 1$. 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 a suitable $\hat{x}_1$ on the insphere and $\hat{x}_2$ as its antipodal point). Then

$$A_K\hat{x} = T_K\hat{x}_1 - T_K\hat{x}_2 = x_1 - x_2 \quad \text{for some } x_1, x_2 \in K,$$

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

Note that since the insphere of diameter ρ_K is contained in K, which in turn is contained in the surrounding sphere of diameter h_K, we can further estimate (with a constant c depending only on n)

$$c h_K^n \geq \operatorname{vol}(K) \geq c\rho_K^n = c\frac{h_K^n}{\sigma_K^n}.$$

The local interpolation error can then be estimated by transforming to the reference element, bounding the error there, and transforming back (a so-called *scaling argument*).

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 some $k \geq 1$ and $\hat{\mathcal{N}}$ bounded on $W^{k,p}(\hat{K})$, $1 \leq p \leq \infty$. 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, there exists a constant $c > 0$ independent of K such that for any $v \in W^{k,p}(K)$,*

$$|v - \mathcal{I}_K v|_{W^{l,p}(K)} \leq c h_K^{k-l} \sigma_K^l |v|_{W^{k,p}(K)} \quad \text{for all } 0 \leq l \leq k. \tag{5.12}$$

Proof. Let $\hat{v} := v \circ T_K$. By Lemma 4.13, $\mathcal{I}_{\hat{K}}\hat{v} = (\mathcal{I}_K v) \circ T_K$ (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 (with a generic constant c that can change from line to line)

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

The claim now follows from Lemma 5.7 and the fact that $h_{\hat{K}}$ and $\rho_{\hat{K}}$ are independent of K. □

To obtain an estimate for the global interpolation error, which should converge 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

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

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

Using this upper bound and summing over all elements, we obtain an 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 the reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ satisfying the requirements of Theorem 5.8 for some $k \geq 1$. Then there exists a constant $c > 0$ independent of h such that for all $v \in W^{k,p}(\Omega)$,*

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

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

Similar estimates can be obtained for elements based on the tensor product spaces Q_k.[2]

[2] e.g., [Brenner & Scott 2008, Chapter 4.6]

5.3 INVERSE ESTIMATES

The above theorems estimated the interpolation error in a coarser norm (i.e., $l \leq k$) than than the given function to be interpolated. In general, the converse (estimating a finer norm by a coarser one) is not possible; however, for the discrete approximations $v_h \in V_h$, such so-called *inverse estimates* can be established.

Local estimates follow as above from a scaling argument, using the equivalence of norms on the finite dimensional space $\hat{\mathcal{P}}$ in place of the Bramble–Hilbert lemma.

Theorem 5.10 (local inverse estimate[3]). *Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a finite element with $\hat{\mathcal{P}} \subset W^{l,p}(\hat{K})$ for an $l \geq 0$ and $1 \leq p \leq \infty$. For any element $(K, \mathcal{P}, \mathcal{N})$ with $h_K \leq 1$ affine equivalent to $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ by the affine transformation T_K, there exists a constant $c > 0$ independent of K such that for any $v_h \in \mathcal{P}$,*

$$\|v_h\|_{W^{l,p}(K)} \leq c h_K^{k-l} \|v_h\|_{W^{k,p}(K)}$$

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

For uniform global estimates, we need a lower bound on h_K^{-1}. A triangulation $\mathcal{T}$ is called *quasi-uniform* if it is shape regular and there exists a $\tau \in (0, 1]$ such that $h_K \geq \tau h$ for all $K \in \mathcal{T}$. By summing over the local estimates, we obtain the following global estimate.

Theorem 5.11 (global inverse estimate[4]). *Let $\mathcal{T}$ be a quasi-uniform affine triangulation of $\Omega \subset \mathbb{R}^n$ with the reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ satisfying the requirements of Theorem 5.10 for an $l \geq 0$. Then there exists a constant $c > 0$ independent of h such that for all $v_h \in V_h := \{v \in L^p(\Omega) : v|_K \in \mathcal{P}, K \in \mathcal{T}\}$,*

$$\left(\sum_{K\in\mathcal{T}} \|v_h\|^p_{W^{l,p}(K)}\right)^{\frac{1}{p}} \leq c h^{k-l} \left(\sum_{K\in\mathcal{T}} \|v_h\|^p_{W^{k,p}(K)}\right)^{\frac{1}{p}}, \quad 1 \leq p < \infty,$$

$$\max_{K\in\mathcal{T}} \|v_h\|_{W^{l,\infty}(K)} \leq c h^{k-l} \left(\max_{K\in\mathcal{T}} \|v_h\|_{W^{k,\infty}(K)}\right),$$

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

[3] e.g., [Ern & Guermond 2004, Lemma 1.138]

[4] e.g., [Ern & Guermond 2004, Corollary 1.141]

6 ERROR ESTIMATES FOR THE FINITE ELEMENT APPROXIMATION

We can now give error estimates for the conforming finite element approximation of elliptic boundary value problems using 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

$$V_h := \left\{v_h \in C^0(\overline{\Omega}) : (v_h|_K \circ T_K) \in \hat{\mathcal{P}} \text{ for all } K \in \mathcal{T}\right\} \cap V \tag{6.1}$$

(the intersection being necessary in case of Dirichlet conditions).

6.1 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 6.1. *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 the reference element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ satisfying $P_{k-1} \in \hat{\mathcal{P}}$ for some $k \geq 1$, and let $u_h \in V_h$ be the corresponding Galerkin approximation. If $u \in H^m(\Omega)$ for $\frac{n}{2} < m < k$, then there exists $c > 0$ independent of h and u such that*

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

Proof. Since $m > \frac{n}{2}$, the Sobolev embedding Theorem 2.3 implies that $u \in C^0(\overline{\Omega})$ and hence that the local (pointwise) interpolant is well defined. In addition, the nodal interpolation preserves homogeneous Dirichlet boundary conditions. Hence $\mathcal{I}_{\mathcal{T}}u \in V_h$, and Céa's lemma yields

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

Theorem 5.9 for $p = 2$, $l = 1$, and $k = m$ further implies

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

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 6.2. *Under the assumptions of Theorem 6.1, there exists $c > 0$ such that*

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

Proof. By the Sobolev embedding Theorem 2.3, the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ is continuous. Thus, the Aubin–Nitsche lemma yields

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

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

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

by the well-posedness of the adjoint problem. Combining this inequality with the one from Theorem 6.1 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]

6.2 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 & Scott 2008, Chapter 8]

[2] e.g., [Brenner & Scott 2008, Chapter 9], [Ern & Guermond 2004, Chapter 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 $\alpha_1 \geq \alpha(x) \geq \alpha_0 > 0$ for almost all $x \in \Omega$ be given. Then we search for $u \in H_0^1(\Omega)$ satisfying

$$a(u,v) := (\alpha\nabla u, \nabla v) = (f,v) \qquad \text{for all } v \in H_0^1(\Omega). \tag{6.2}$$

(The same arguments can be carried out for the general boundary value problem (2.2) with homogeneous Dirichlet or Neumann conditions). Let $V_h \subset H_0^1(\Omega)$ be a finite element space and let $u_h \in V_h$ be 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

$$\begin{aligned}
\alpha_0 \|u-u_h\|_{H^1(\Omega)} &\leq \frac{a(u-u_h,u-u_h)}{\|u-u_h\|_{H^1(\Omega)}} \\
&\leq \sup_{w\in H_0^1(\Omega)} \frac{a(u-u_h,w)}{\|w\|_{H^1(\Omega)}} \\
&= \sup_{w\in H_0^1(\Omega)} \frac{a(u,w)-(\alpha\nabla u_h,\nabla w)}{\|w\|_{H^1(\Omega)}} \\
&= \sup_{w\in H_0^1(\Omega)} \frac{(f,w)-\langle -\nabla\cdot(\alpha\nabla u_h),w\rangle_{H^{-1},H^1}}{\|w\|_{H^1(\Omega)}} \\
&= \sup_{w\in H_0^1(\Omega)} \frac{\langle f+\nabla\cdot(\alpha\nabla u_h),w\rangle_{H^{-1},H^1}}{\|w\|_{H^1(\Omega)}} \\
&= \|f+\nabla\cdot(\alpha\nabla u_h)\|_{H^{-1}(\Omega)}
\end{aligned}$$

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 (distributional) divergence of $w \in L^2(\Omega)^n$. Since all terms on the right-hand side are known, this is in principle already 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 (noting that $\nu_1 = -\nu_2$) as

$$[\![\alpha\nabla w_h]\!]_F := (\alpha\nabla w_h)|_{K_1}\cdot\nu_1 + (\alpha\nabla w_h)|_{K_2}\cdot\nu_2 \in L^2(F).$$

Assume now that α is piecewise smooth and that the triangulation is chosen such that $\alpha|_K \in C^1(\overline{K})$ for every $K \in \mathcal{T}_h$. We can then perform the above integration by parts elementwise to obtain for $w \in H_0^1(\Omega)$

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

since $w \in H_0^1(\Omega)$ is continuous almost everywhere and u_h is a polynomial on K and hence $(\alpha\nabla u_h)|_K$ is in fact differentiable.

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 inserting (via Galerkin orthogonality) the interpolant of w and applying an interpolation error estimate. The difficulty here is that $w \in H_0^1(\Omega)$ is not sufficiently smooth to allow Lagrange interpolation, since pointwise evaluation is not well-defined. To circumvent this, we combine interpolation with projection. For $K \in \mathcal{T}_h$, let ω_K be the union of all elements touching K, i.e.,

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

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

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

The $L^2(\omega_z)$ projection of $v \in H^1(\Omega)$ onto P_0 is then defined as the unique $\pi_z(v) \in P_0$ satisfying

$$\int_{\omega_z}(\pi_z(v)-v)q\,dx = 0 \qquad \text{for all } q \in P_0,$$

see Lemma 5.2 for $k = 0$. For $z \in \partial\Omega$, we set $\pi_z(v) = 0$ to respect the homogeneous Dirichlet conditions. The local *Clément interpolant* $\mathcal{I}_C v \in V_h$ of $v \in H_0^1(\Omega)$ is then given by

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

Since the $L^2(\omega_z)$ projection is continuous on $H^1(\omega_K)$ for every $z \in \overline{K}$, we can apply the Bramble–Hilbert lemma together with a scaling argument to obtain as for the standard interpolation the error estimates[3]

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

for all $v \in H^1_0(\Omega)$, $K \in \mathcal{T}_h$ and $F \subset \partial K$. (Note $\operatorname{vol}(F)$ scales with h_K, while $\operatorname{vol}(K)$ scales with h_K^2.)

Using the Galerkin orthogonality for the global Clément interpolant $\mathcal{I}_C w \in V_h$, we can proceed as before, use the Cauchy–Schwarz inequality, these interpolation error estimates, and the fact that every K appears only in a finite number of ω_K, to arrive at the estimate

$$\begin{aligned}\|u-u_h\|_{H^1(\Omega)} &\le \frac{1}{\alpha_0}\sup_{w\in H^1_0(\Omega)} \frac{a(u-u_h, w-\mathcal{I}_C w)}{\|w\|_{H^1(\Omega)}}\\
&\le \frac{1}{\alpha_0}\sup_{w\in H^1_0(\Omega)} \frac{1}{\|w\|_{H^1(\Omega)}}\left(\sum_{K\in\mathcal{T}_h}\|f+\nabla\cdot(\alpha\nabla u_h)\|_{L^2(K)}\|w-\mathcal{I}_C w\|_{L^2(K)}\right.\\
&\qquad\qquad\left.+\sum_{F\in\Gamma_h}\left\|[\![\alpha\nabla u_h]\!]_F\right\|_{L^2(F)}\|w-\mathcal{I}_C w\|_{L^2(F)}\right)\\
&\le C\sup_{w\in H^1_0(\Omega)} \frac{1}{\|w\|_{H^1(\Omega)}}\left(\sum_{K\in\mathcal{T}_h}h_K\|f+\nabla\cdot(\alpha\nabla u_h)\|_{L^2(K)}\|w\|_{H^1(\Omega)}\right.\\
&\qquad\qquad\left.+\sum_{F\in\Gamma_h}h_K^{1/2}\left\|[\![\alpha\nabla u_h]\!]_F\right\|_{L^2(F)}\|w\|_{H^1(\Omega)}\right)\\
&\le C\left(\sum_{K\in\mathcal{T}_h}h_K\|f+\nabla\cdot(\alpha\nabla u_h)\|_{L^2(K)}+\sum_{F\in\Gamma_h}h_K^{1/2}\left\|[\![\alpha\nabla u_h]\!]_F\right\|_{L^2(F)}\right).\end{aligned}$$

All terms on the right-hand side are now fully computable given a discrete solution u_h. To obtain a minimal upper bound (as a proxy for minimizing the error itself), we are thus lead to make h_K small(er) (by subdividing K into a number of smaller elements) where the *finite element residual* is large or the normal derivative has a large jump.

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 as we can then apply the Aubin–Nitsche trick. Let $w \in H^1_0(\Omega)\cap H^2(\Omega)$ solve the adjoint problem

$$a(v,w) = (u-u_h, v) \qquad \text{for all } v \in H^1_0(\Omega).$$

Inserting $v = u-u_h \in H^1_0(\Omega)$ and applying the Galerkin orthogonality $a(u-u_h, w_h) = 0$ for the global interpolant $w_h := \mathcal{I}_{\mathcal{T}} w$ then yields

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

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

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

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

By the symmetry of a and the well-posedness of (6.2), we have $w\in H^2(\Omega)$ due to Theorem 2.10. We can thus estimate the local interpolation error for w using Theorem 5.8 for $k=2$, $l=0$ and $p=2$ to obtain

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

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

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

Finally, we have from Theorem 2.10 the estimate

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

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

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

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 & Scott 2008, Chapter 9], [Verfürth 2013].

7 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.

7.1 TRIANGULATION

The geometric information on a triangulation is described by a *mesh*, a cloud of connected points in $\mathbb{R}^2$. This information is usually stored in a collection of two-dimensional arrays, the most fundamental of which are

- *the list of nodes*, which contains the coordinates $z_i = (x_i, y_i)$ of each node corresponding to a 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) = (i_1,i_2,i_3),
  ```

  where z_{i_1} =`nodes(i_1)`. 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 (usually function values first, then gradient).

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

[1] [Bangerth, Hartmann & Kanschat 2013], http://www.dealii.org

[2] [Logg, Mardal, Wells, et al. 2012], http://fenicsproject.org

- 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]

7.2 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 basis elements φ_i, φ_j. This is most efficiently done element-wise, where the computation is performed by transformation to a reference element.

The reference element

We consider the reference element domain

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

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 $T_K(\xi) = A_K\xi + b_K$ with

$$A_K = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}, \qquad b_K = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}.$$

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 $\hat{N}_i(\hat{\psi}_j) = \delta_{ij}$, $1 \leq i, j \leq d$. (For example, the nodal basis for the linear Lagrange element is $\{1 - \xi_1 - \xi_2, \xi_1, \xi_2\}$.)

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

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

[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*[4]), 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 (see Chapter 8).

Theorem 7.1 (effect of quadrature[5]). *Let $\mathcal{T}_h$ be a shape regular affine triangulation with $P_1 \subset \hat{\mathcal{P}} \subset P_k$ for $k \geq 1$. If the quadrature on $\hat{K}$ is of order $2k-2$, all weights are positive, and h is small enough, then the discrete problem is well-posed.*

If in addition the surface integrals are approximated by a quadrature rule of order $2k-1$ and the conditions of Theorem 6.1 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,

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

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), and (x_3, y_3), the barycentric coordinates $(\zeta_1, \zeta_2, \zeta_3)$ of $(x, y) \in K$ are defined by

(i) $\zeta_1, \zeta_2, \zeta_3 \in [0, 1]$,

(ii) $\zeta_1 + \zeta_2 + \zeta_3 = 1$,

(iii) $(x, y) = \zeta_1(x_1, y_1) + \zeta_2(x_2, y_2) + \zeta_3(x_3, y_3)$.

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 Gauß nodes in barycentric coordinates and the corresponding weights for quadrature of order up to 5 are given

[4] [Strang 1972]

[5] e.g., [Ciarlet 2002, Theorems 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 7.1: 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.

in Table 7.1. The element contributions of the local basis functions can then be computed as, e.g., in

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

where $A(x) = (a_{ij}(x))_{i,j=1}^2$ is the matrix of coefficients for the second-order derivatives, 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 $\mathbf{K}$ and right-hand side $\mathbf{F}$ is sketched in Algorithm 7.1.

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 matrices and right-hand side as above, and replace each row in $\mathbf{K}$ and entry in $\mathbf{F}$ corresponding to a node in `bdy_nodes` with the equation for the prescribed nodal value:

```
for $i = 1, \dots,$ length(bdy_nodes) do
    Set $k$ = bdy_nodes($i$)
    Set $K_{k,j} = 0$ for all $j$
    Set $K_{k,k} = 1$, $F_k = g($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 7.1, where the loop over `elements` is replaced by a loop over `bdy_faces` (and one-dimensional Gauß quadrature is used).

Algorithm 7.1 Finite element method for Lagrange triangles

```
Require: 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 elements and their gradients at Gauß nodes on reference element
Set $K_{ij} = F_j = 0$
for $k = 1, \dots,$ length(elements) do
    Compute transformation $T_K$, Jacobian $\det(A_K)$ for element $K =$ elements$(k)$
    Evaluate coefficients and right-hand side at transformed Gauß nodes $T_K(\xi_l)$
    Compute $a(\varphi_i, \varphi_j)$, $\left(f, \varphi_j\right)$ for all nodal basis elements $\varphi_i, \varphi_j$ using transformation rule and Gauß quadrature on reference element
    for $i, j = 1, \dots, d$ do
        Set $r =$ elements$(k, i)$, $s =$ elements$(k, j)$
        Set $K_{r,s} \leftarrow K_{r,s} + a(\varphi_i, \varphi_j)$, $F_s \leftarrow F_s + \left(f, \varphi_j\right)$
    end for
end for
Ensure: $K_{ij}, F_j$
```

Part III

NONCONFORMING FINITE ELEMENTS

8 GENERALIZED GALERKIN APPROACH

The results of the preceding chapters depended on the conformity 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, where the function u satisfying $a(u,v)$ for all $v \in V$ is an element of $U \neq V$;
- *non-conforming* approaches, where the discrete spaces U_h and V_h are not subspaces of U and V, respectively; and
- *non-consistent* approaches, where 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 that covers these cases as well. Let U, V be Banach spaces, where V is reflexive, and let U^*, V^* denote their topological 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

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

The following generalization of the Lax–Milgram theorem gives sufficient (and, as can be shown, necessary) conditions for the well-posedness of ($\mathcal{W}$).

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 assumptions:*

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

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

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

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

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

(iii) Injectivity: *for any* $v \in V$,

$$a(u,v) = 0 \text{ for all } u \in U \quad \text{implies} \quad v = 0.$$

Then there exists a unique solution $u \in U$ *to* ($\mathcal{W}$), *which satisfies*

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

Proof. The proof is essentially an application of the closed range theorem:[1] For a bounded linear operator A between two Banach spaces X and Y, the range $\operatorname{ran} A$ of A is closed in Y if and only if $\operatorname{ran} A = (\ker A^*)^\perp$, 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 $V \subset X$,

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

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 ($\mathcal{W}$) 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

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

and therefore $u_1 = u_2$.

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

$$c_1 \|u\|_U \leq \sup_{v \in V} \frac{a(u,v)}{\|v\|_V} = \sup_{v \in V} \frac{\langle Au, v\rangle_{V^*,V}}{\|v\|_V} = \sup_{v \in V} \frac{\langle v^*, v\rangle_{V^*,V}}{\|v\|_V} = \|v^*\|_{V^*}. \tag{8.1}$$

Any preimage thus satisfies the claimed inequality; it remains to show that every $v^* \in V^*$ has a preimage. We next show that $\operatorname{ran} A$ is closed. Let $\{v_n^*\}_{n \in \mathbb{N}} \subset \operatorname{ran} A \subset V^*$ be a sequence converging to a $v^* \in V^*$, i.e., there exists $u_n \in U$ such that $v_n^* = Au_n$, and the v_n^* form a Cauchy sequence. From (8.1), we deduce for all $n, m \in \mathbb{N}$ that

$$\|u_n - u_m\|_U \leq \frac{1}{c_1} \|A(u_n - u_m)\|_{V^*} = \frac{1}{c_1} \left\|v_n^* - v_m^*\right\|_{V^*},$$

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

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 $v^* \in \operatorname{ran} A$. We can therefore apply the closed range theorem. By the reflexivity of V, we have $A^* : V \to U^*$ and

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

due to the injectivity condition (iii). Hence the closed range theorem and the reflexivity of V yields

$$\operatorname{ran} A = (\{0\})^\perp = \left\{v^* \in V^* : \langle v^*, 0\rangle_{V^*,V} = 0\right\} = V^*,$$

and therefore surjectivity of A. Thus, A is an isomorphism and the claimed estimate follows from (8.1) applied to $v^* = F \in V^*$. □

The term "injectivity condition" 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 where $U = V$ is a Hilbert space, coercivity of a implies both the inf–sup condition and (via contraposition) 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

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

In contrast to the conforming setting, the well-posedness of $(\mathcal{W}_h)$ cannot be deduced from the well-posedness of $(\mathcal{W})$ but needs to be proved independently. This is somewhat simpler due to the finite-dimensionality of the spaces, where injectivity of a square matrix already implies surjectivity.

Theorem 8.2. *Let U_h and V_h be finite-dimensional vector spaces with norms $\|\cdot\|_{U_h}$ and $\|\cdot\|_{V_h}$, respectively. 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 assumptions:*

(i) discrete Inf–sup condition*: there exists a* $c_1 > 0$ *such that*

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

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

$$|a_h(u_h, v_h)| \leq c_2 \|u_h\|_{U_h} \|v_h\|_{V_h},$$
$$|F_h(v_h)| \leq c_3 \|v_h\|_{V_h}$$

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

Assume further that $\dim U_h = \dim V_h$. *Then there exists a unique solution* $u_h \in U_h$ *to* ($\mathcal{W}_h$), *which satisfies*

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

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 $\mathbf{K} \in \mathbb{R}^{n \times n}$, $K_{ij} = a_h(\varphi_i, \psi_j)$. 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. The estimate follows again from the inf–sup condition. □

Note that since U_h and V_h are no longer assumed to be a subspaces of U and V, respectively, they can't simply inherit the norms of the latter, and we thus have to choose new one. These *discrete norms* will have to be specifically adapted to the discrete bilinear form a_h in order to ensure that both the inf–sup and the continuity condition are satisfied. Note also 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- and injectivity condition only amounts to requiring injectivity and surjectivity.

The error estimates for non-conforming methods are based on the following two generalization of Céa's lemma. 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 $U_* \subset U$ containing the exact solution such that

$$U(h) := U_* + U_h = \{w + w_h : w \in U_*, w_h \in U_h\}$$

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

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

(ii) $\|u\|_{U(h)} \leq c \|u\|_U$ for all $u \in U_*$.

The first result concerns non-consistent but conforming approaches and can be used to prove estimates for the error arising from numerical integration; see Theorem 7.1. Note that here we do not assume that the discrete bilinear form a_h is well-defined for the continuous solution $u \in U$.

Theorem 8.3 (first Strang lemma). *Assume that the conditions of Theorem 8.2 hold and 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*

$$|a(u, v_h)| \leq c_4 \|u\|_{U(h)} \|v_h\|_{V_h} \qquad \text{for all } u \in U, v_h \in V_h.$$

Then the solutions u *and* u_h *to* ($\mathcal{W}$) *and* ($\mathcal{W}_h$)*, respectively, satisfy*

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

Proof. Let $w_h \in U_h$ be arbitrary. By the triangle inequality and the assumption on the error norm, we have

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

For the second term, we can apply the discrete inf–sup condition to obtain

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

Using ($\mathcal{W}$) and ($\mathcal{W}_h$), by assumption (i) we can write

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

Inserting this into the last estimate and applying the assumption (ii) yields

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

The claim now follows by 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 conformity.

Theorem 8.4 (second Strang lemma). *Assume that the conditions of Theorem 8.2 hold and that there exists a constant $c_4 > 0$ independent of h such that*

$$|a_h(u, v_h)| \leq c_4 \|u\|_{U(h)} \|v_h\|_{V_h} \qquad \text{for all } u \in U(h), v_h \in V_h.$$

Then the solutions u and u_h to ($\mathcal{W}$) and ($\mathcal{W}_h$), respectively, satisfy

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

Proof. We proceed as before. 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 then imply

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

and we again conclude using the triangle inequality and taking the infimum over all $w_h \in U_h$. □

To illustrate the application of the first Strang lemma, we consider the effect of quadrature on the Galerkin approximation. For simplicity, we consider for $u, v \in H_0^1(\Omega) = U = V$ the continuous bilinear form

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

with $\alpha \in W^{1,\infty}(\Omega) \hookrightarrow C^0(\overline{\Omega})$, $\alpha_1 \geq \alpha(x) \geq \alpha_0 > 0$. Let $U_h = V_h \subset H_0^1(\Omega) = U(h)$ be constructed from triangular Lagrange elements of degree m on an affine-equivalent triangulation $\mathcal{T}_h$. The discrete bilinear form is then

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

where $w_k > 0$ and x_k are Gauß quadrature weights and nodes on each element and l_m is chosen sufficiently large that the quadrature is exact for polynomials of degree up to $2m-1$. Since ∇u_h is a vector of polynomials of degree $m-1$, this implies

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

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

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

Similarly, a_h is coercive since

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

by positivity of the weights and Poincaré's inequality (Theorem 2.6). As coercivity implies the inf–sup condition, the discrete problem is well-posed by Theorem 8.2.

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

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

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

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

Since for any $m \geq 1$,

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

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

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

A scaling argument then yields (noting that the right-hand norm involves the essential supremum over K and thus doesn't scale with $\operatorname{vol}(K)$)

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

Inserting this for $m = 1$ and using again that $\nabla u_h, \nabla v_h$ are constant on each element, we obtain

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

For the quadrature error on the right-hand side $F_h(v_h) = \sum_{k=1}^{l_m} w_k f(x_k) v_h(x_k)$ for given $f \in W^{1,\infty}(\Omega)$, we can proceed similarly (applying the Bramble–Hilbert lemma to $E_K(f v_h)$ and using the product rule and equivalence of norms on V_h, followed by a scaling argument) to obtain

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

Combining these estimates with the first Strang lemma yields (with a generic constant C independent of h and using $\|w_h\|_{H^1(\Omega)} \leq \|u - w_h\|_{H^1(\Omega)} + \|u\|_{H^1(\Omega)}$) that

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\leq Ch \|f\|_{W^{1,\infty}(\Omega)} + \inf_{w_h\in V_h} \left(C \|u - w_h\|_{H^1(\Omega)} + Ch |\alpha|_{W^{1,\infty}(\Omega)} \|w_h\|_{H^1(\Omega)} \right) \\ &\leq Ch \|f\|_{W^{1,\infty}(\Omega)} + C \inf_{w_h\in V_h} \left(C \|u - w_h\|_{H^1(\Omega)} + Ch \|u - w_h\|_{H^1(\Omega)} \right) \\ &\qquad + Ch \|u\|_{H^1(\Omega)} \\ &\leq Ch \|f\|_{W^{1,\infty}(\Omega)} + Ch^2 |u|_{H^2(\Omega)} + Ch \|u\|_{H^2(\Omega)} \\ &\leq Ch \left(\|f\|_{W^{1,\infty}(\Omega)} + \|u\|_{H^2(\Omega)} \right), \end{aligned}$$

for $h < 1$, as claimed in Theorem 7.1.

9 DISCONTINUOUS GALERKIN METHODS

Discontinuous Galerkin methods are based on nonconforming finite element spaces consisting of piecewise polynomials that are not continuous across elements. These allow handling 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 of a form to 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 that

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

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

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

denote the *inflow boundary* and

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

denote the *outflow boundary*, and assume that they are well-separated, i.e.,

$$\inf_{s \in \partial\Omega^-, t \in \partial\Omega^+} |s - t| > 0.$$

Then we define the so-called *graph space*

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

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*

$$\|v\|_W^2 = \|v\|_{L^2(\Omega)}^2 + \|\beta \cdot \nabla v\|_{L^2(\Omega)}^2 .$$

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

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

and that the following integration by parts formula holds:

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

for all $v, w \in W$.

We can now define our weak formulation: set

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

and find $u \in U$ satisfying

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

for all $v \in V = L^2(\Omega)$. 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*

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

then there exists a unique $u \in U$ satisfying ($\mathcal{W}$). Furthermore, there exists a $C > 0$ such that

$$\|u\|_W \leq C \|f\|_{L^2(\Omega)} .$$

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

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

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

To verify the inf–sup condition, we first prove coercivity with respect to the $L^2(\Omega)$ part of the graph norm. For any $u \in U \subset V$, we integrate by parts using (9.1) for $v = w = u$ to obtain

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

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 that

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

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

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

Summing the last two inequalities and taking the infimum over all $u \in U$ 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 from $a(u,v) = 0$ that $\nabla \cdot (\beta v)$ exists as a weak derivative and that $\nabla \cdot (\beta v) = \mu v$. By the product 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 $u \in U$

$$\begin{aligned} \int_{\partial\Omega} (\beta \cdot \nu)uv\,dx &= \int_\Omega (\beta \cdot \nabla v)u + (\beta \cdot \nabla u)v + (\nabla \cdot \beta)vu\,dx \\ &= a(u,v) - ((\mu - \nabla \cdot \beta)v - \beta \cdot \nabla v, u) \\ &= 0. \end{aligned} \tag{9.2}$$

Since $\partial\Omega^+$ and $\partial\Omega^-$ are well separated, there exists a smooth cut-off function $\chi \in C^\infty(\overline{\Omega})$ with $\chi(s) = 0$ for $s \in \partial\Omega^-$ and $\chi(s) = 1$ for $s \in \partial\Omega^+$. Applying (9.2) to $u = \chi v \in U$ yields

$\int_{\partial\Omega^+}(\beta\cdot\nu)v^2\,dx = 0$. Using again that $\mu v = \nabla\cdot(\beta v)$ and integrating by parts, we deduce that

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

since the remaining boundary integral over $\partial\Omega^-$ is non-positive. This shows that $v = 0$. □

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 for $k \geq 0$ and a given triangulation $\mathcal{T}_h$ of Ω *both* of the discrete spaces as

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

(no continuity across elements is assumed, hence the name). We then search for $u_h \in V_h$ satisfying

$$(\mathcal{W}_h) \qquad a_h(u_h, v_h) = (f, v_h) \qquad \text{for all } v_h\in V_h,$$

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 all faces of $\mathcal{T}_h$ is denoted by $\partial\mathcal{T}_h$ and the set of all interior faces 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 (sufficiently regular) function $v\in L^2(\Omega)$, we denote the *jump* across F as

$$[\![v]\!]_F = v|_{K_1}\nu_1 + v|_{K_2}\nu_2 \in L^2(F)^n,$$

and the *average* as

$$\{\{v\}\}_F = \tfrac{1}{2}(v|_{K_1} + v|_{K_2}) \in L^2(F).$$

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* $\nabla_h v_h \in L^2(\Omega)$ 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

$$
\begin{aligned}
(9.3) \qquad a_h(u_h, v_h) = (\mu u_h + \beta\cdot\nabla_h u_h, v_h) &- \int_{\partial\Omega^-}(\beta\cdot\nu)u_h v_h\,ds\\
&- \sum_{F\in\Gamma_h}\int_F \beta\cdot[\![u_h]\!]\,\{\{v_h\}\}\,ds.
\end{aligned}
$$

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 in the following proof of coercivity with respect to the "discrete energy norm"

$$|||u_h|||^2 := \mu_0 \|u_h\|^2_{L^2(\Omega)} + \int_{\partial\Omega} \tfrac{1}{2}|\beta\cdot\nu|u_h^2\,ds,$$

which is clearly a norm on $V_h \subset L^2(\Omega)$.

Lemma 9.2. *Under the assumption of Theorem 9.1, there exists a constant $C > 0$ independent of h such that*

$$a_h(u_h, u_h) \geq C\,|||u_h|||^2 \qquad \text{for all } u_h \in V_h.$$

Proof. We begin by applying integration by parts on each element to the first term of (9.3) for $v_h = u_h$ to obtain

$$\begin{aligned}(\mu u_h + \beta\cdot\nabla_h u_h, u_h) &= \sum_{K\in\mathcal{T}_h}\int_K \mu u_h^2 + (\beta\cdot\nabla u_h)u_h\,dx\\ &= \sum_{K\in\mathcal{T}_h}\int_K \mu u_h^2 - \tfrac{1}{2}(\nabla\cdot\beta)u_h^2\,dx + \int_{\partial K}\tfrac{1}{2}(\beta\cdot\nu)u_h^2\,ds.\end{aligned}$$

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

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K}\tfrac{1}{2}(\beta\cdot\nu)u_h^2\,ds = \sum_{F\in\Gamma_h}\int_F \tfrac{1}{2}\beta\cdot [\![u_h^2]\!]\,ds + \sum_{F\in\partial\mathcal{T}_h\setminus\Gamma_h}\int_F \tfrac{1}{2}(\beta\cdot\nu)u_h^2\,ds.$$

Using that $\nu := \nu_1 = -\nu_2$ and therefore

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

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

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K}\tfrac{1}{2}(\beta\cdot\nu)u_h^2\,ds - \int_{\partial\Omega^-}(\beta\cdot\nu)u_h^2\,ds = \sum_{F\in\Gamma_h}\int_F \beta\cdot[\![u_h]\!]\,\{\{u_h\}\}\,ds + \int_{\partial\Omega}\tfrac{1}{2}|\beta\cdot\nu|u_h^2\,ds.$$

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

$$\begin{aligned}a_h(u_h,u_h) &= \sum_{K\in\mathcal{T}_h}\int_K\left(\mu - \tfrac{1}{2}(\nabla\cdot\beta)\right)u_h^2\,dx + \int_{\partial\Omega}\tfrac{1}{2}|\beta\cdot\nu|u_h^2\,ds\\ &\geq \mu_0\|u_h\|^2_{L^2(\Omega)} + \int_{\partial\Omega}\tfrac{1}{2}|\beta\cdot\nu|u_h^2\,ds\\ &= |||u_h|||^2.\end{aligned}$$

□

We will show continuity of a on $V_h \times V_h$ (with respect to an equivalent norm) later (Lemma 9.3), from which we then obtain existence of a unique solution $u_h \in V_h$ to ($\mathcal{W}_h$).

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, we make the additional assumption that the exact solution satisfies

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

By the trace theorem (Theorem 2.5), $u|_F$ is then well-defined in the sense of $L^2(F)$ traces. We now 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 on $U(h)\times V_h$ if the triangulation is shape-regular.

Lemma 9.3. *If $\mathcal{T}_h$ is a shape-regular triangulation of $\Omega\subset\mathbb{R}^2$, then there exists a constant $C>0$ independent of h such that*

$$a_h(u,v_h) \le C\,|||u|||_*\,|||v_h||| \qquad \textit{for all } u\in U(h),\ v_h\in V_h.$$

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

$$(9.4)\qquad (\mu u + \beta\nabla_h u, v_h) + \int_{\partial\Omega^-}(\beta\cdot\nu)uv_h\,ds \le C\,|||u|||_*\,|||v_h|||,$$

with a constant $C>0$ depending only on μ. For the last term of $a_h(u,v_h)$, we first insert $1 = (2\{\{h\}\})(2\{\{h\}\})^{-1}$, where in a slight abuse of notation we consider $h:\Omega\to\mathbb{R}$ as a function mapping $x\in K$ to h_K. Since this function h is constant on each element, we obtain using the Cauchy–Schwarz inequality that

$$(9.5)\quad \sum_{F\in\Gamma_h}\int_F \beta\cdot[\![u]\!]\,\{\{v_h\}\}\,ds \le C\left(\sum_{F\in\Gamma_h}\tfrac{1}{2}\{\{h\}\}^{-1}\|[\![u]\!]\|_{L^2(F)^n}^2\right)^{\frac12}\left(\sum_{F\in\Gamma_h}2\{\{h\}\}\,\|\{\{v_h\}\}\|_{L^2(F)}^2\right)^{\frac12},$$

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

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

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

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

Combining the terms arising from the faces of each element and applying the discrete trace inequality (obtained in the usual way)[2]

$$h_K^{1/2}\,\|v_h\|_{L^2(\partial K)} \leq C\,\|v_h\|_{L^2(K)}\,, \tag{9.6}$$

we thus obtain

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

Adding (9.4) and (9.7) yields the claim. □

On the finite-dimensional subspace $V_h \subset U(h)$, the norm $|\!|\!|\cdot|\!|\!|_*$ is equivalent to $|\!|\!|\cdot|\!|\!|$, and hence together with Lemma 9.2 we now have verified the conditions necessary for applying Theorem 8.2 to deduce well-posedness of ($\mathcal{W}_h$). However, since $|\!|\!|\cdot|\!|\!|_*$ involves h_K, the constants of equivalence also depend on the mesh size h, and hence the a priori estimate is no longer uniform in h.

Corollary 9.4. *If $\mathcal{T}_h$ is a shape-regular triangulation of $\Omega\subset\mathbb{R}^2$ and*

$$\mu(x) - \tfrac12\nabla\cdot\beta(x) \geq \mu_0 > 0 \quad \textit{for almost all } x\in\Omega,$$

then there exists a unique solution $u_h\in V_h$ to ($\mathcal{W}_h$). Furthermore, there exists a constant $C>0$ such that

$$|\!|\!|u_h|\!|\!| \leq C\,\|f\|_{L^2(\Omega)}\,.$$

Before we derive error estimates, we show that our discontinuous Galerkin approximation is consistent and hence that the consistency error in the second Strang lemma vanishes.

Lemma 9.5. *A solution $u\in U_*$ to ($\mathcal{W}$) satisfies*

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

for all $v_h\in V_h$.

Proof. By definition, $u\in U_* = U\cap H^1(\Omega)$ satisfies $\nabla_h u = \nabla u$ and thus

$$(\mu u + \beta\cdot\nabla_h u, v_h) = (f,v_h) \quad \text{for all } v_h\in V_h\subset V.$$

Furthermore, due to the boundary conditions,

$$\int_{\partial\Omega^-}(\beta\cdot\nu)u v_h\,ds = 0.$$

[2] e.g., [Di Pietro & Ern 2012, Lemma 1.46]

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) yields

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

The claim then follows from a density argument. □

We thus obtain the following error estimate.

Theorem 9.6. *Assume that the solution $u\in U(h)$ to ($\mathcal{W}$) satisfies $u\in H^{k+1}(\Omega)$. Then there exists a $c>0$ independent of h such that*

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

Proof. Since $a_h:U(h)\times V_h\to\mathbb{R}$ is consistent, continuous with respect to the $|\!|\!|\cdot|\!|\!|_*$ norm, and coercive with respect to the $|\!|\!|\cdot|\!|\!|$ norm, we deduce as in the second Strang lemma that

$$|\!|\!| u-u_h |\!|\!| \le c\inf_{w_h\in V_h}|\!|\!| u-w_h |\!|\!|_*\,.$$

Assuming that u is sufficiently smooth that the local interpolant $\mathcal{I}_K u$ is well-defined, we can show by the usual arguments that

$$\begin{aligned}
\|u-\mathcal{I}_K u\|_{L^2(K)} &\le ch_K^{k+1}|u|_{H^{k+1}(K)},\\
|u-\mathcal{I}_K u|_{H^1(K)} &\le ch_K^{k}|u|_{H^{k+1}(K)},\\
\|u-\mathcal{I}_K u\|_{L^2(\partial K)} &\le ch_K^{k+1/2}|u|_{H^{k+1}(K)}.
\end{aligned}$$

Applying these bounds in turn to each term in $|\!|\!| u-\mathcal{I}_\mathcal{T} u|\!|\!|_*$ yields the desired estimate. □

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. (The penalty only canceled the interface terms arising in the integration by parts, but did not contribute further in the coercivity). A more stable alternative is *upwinding*: Take

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

for a sufficiently large penalty parameter $\eta > 0$. It can be shown[3] that this bilinear form is consistent as well, and is coercive in the norm

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

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 $h^{k+1/2}$ (which is useful in the case $k = 0$ as well).

9.4 DISCONTINUOUS GALERKIN METHODS FOR ELLIPTIC EQUATIONS

Due to their flexibility, discontinous Galerkin methods have become popular for elliptic second-order problems as well. We illustrate the approach with the simplest example, the Poisson equation $-\Delta u = f$ on $\Omega \subset \mathbb{R}^n$ with homogeneous Dirichlet conditions. The basic idea is to write the second-order equation as a system of first-order equations, for which we can proceed as before via element-wise integration by parts to obtain face integrals that can be used as penalty terms in place of the dropped continuity requirement and boundary condition on the discrete solution. For $u \in H^1(\Omega)$, we thus introduce $\sigma := \nabla u \in L^2(\Omega)^n$ so that the Poisson equation reduces to $-\nabla \cdot \sigma = f$. We now multiply these two equations with (sufficiently smooth) test functions $\tau \in C^\infty(\overline{\Omega})^n$ and $v \in C^\infty(\overline{\Omega})$, respectively, and integrate by parts separately on each element K of a triangulation $\mathcal{T}_h$ of Ω to obtain

$$(9.8) \qquad \begin{cases} \displaystyle\sum_{K\in\mathcal{T}_h} \int_K \sigma \cdot \tau \, dx + \sum_{K\in\mathcal{T}_h} \int_K u \nabla \cdot \tau \, dx - \sum_{K\in\mathcal{T}_h} \int_{\partial K} u \, (\tau \cdot \nu) \, ds = 0, \\ \displaystyle\qquad\qquad \sum_{K\in\mathcal{T}_h} \int_K \sigma \cdot \nabla v \, dx - \sum_{K\in\mathcal{T}_h} \int_{\partial K} (\sigma \cdot \nu) \, v \, ds = (f, v). \end{cases}$$

The idea is now to replace u and σ in the face integrals by a suitable approximations $\hat{u}_F$ of u and $\hat{\sigma}_F$ of ∇u (sometimes called *potential* and *diffusive flux*, respectively) and then

[3] e.g., [Di Pietro & Ern 2012, Chapter 2.3]

eliminating σ (but not $\hat{\sigma}$). Inserting $\tau = \nabla v$ in the first equation of (9.8) and integrating by parts in the second term yields, on each element, after rearranging

$$\int_K \sigma \cdot \nabla v\,dx = \int_K \nabla u \cdot \nabla v\,dx - \int_{\partial K} u\,(\nabla v \cdot \nu)\,ds + \int_{\partial K} \hat{u}_F\,(\nabla v \cdot \nu)\,ds.$$

Inserting this into the left-hand side of the second equation then yields (using the definition of the broken gradient)

$$\begin{aligned} a_h(u,v) &:= (\nabla_h u, \nabla_h v) + \sum_{K\in\mathcal{T}_h} \int_{\partial K} (\hat{u}_F - u)\,(\nabla v \cdot \nu)\,ds - \sum_{K\in\mathcal{T}_h} \int_{\partial K} (\hat{\sigma}_F \cdot \nu)\,v\,ds \\ &= (f,v). \end{aligned} \tag{9.9}$$

The next step is to rearrange the sum over element boundary integrals into a sum over faces. A straightforward computation shows that for piecewise smooth scalar-valued v and vector-valued τ,

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} v\,(\tau \cdot \nu)\,ds = \sum_{F\in\partial\mathcal{T}_h} \int_F [\![v]\!] \cdot \{\{\tau\}\}\,ds + \sum_{F\in\Gamma_h} \int_F \{\{v\}\}\,[\![\tau]\!]\,ds \tag{9.10}$$

(recalling that the jump of a scalar function is vector-valued, while that of a vector-valued is scalar; see Section 6.2). Before applying this to the terms in (9.9), however, we first discuss the choice of fluxes, each of which leads to a different discontinous Galerkin approach. A popular choice[4] is the *symmetric interior penalty* method, which corresponds to setting

$$\hat{u}_F := \{\{u\}\}_F \quad \text{for } F \in \Gamma_h, \qquad \hat{u}_F = 0 \quad \text{for } F \in \mathcal{T}_h \setminus \Gamma_h, \qquad \hat{\sigma}_F := \{\{\nabla_h u\}\}_F - \frac{\eta}{h_F}\,[\![u]\!]_F\,,$$

where h_F is the diameter of the face $F \in \partial\mathcal{T}_h$ and $\eta > 0$ has to be chosen sufficiently large. (The specific form of the second term will again become clear when discussing coercivity below.) With these choices, applying (9.10) to (9.9) and using that $\{\{\{\{w\}\}\}\} = \{\{w\}\}$ and $[\![\{\{w\}\}]\!] = [\![[\![w]\!]]\!] = 0$ for all w, we arrive at

$$a_h(u,v) = (\nabla_h u, \nabla_h v) - \sum_{F\in\partial\mathcal{T}_h} \int_F [\![u]\!] \cdot \{\{\nabla_h v\}\} + \{\{\nabla_h u\}\} \cdot [\![v]\!]\,ds + \int_F \frac{\eta}{h_F}\,[\![u]\!]\,[\![v]\!]\,ds. \tag{9.11}$$

As usual in a discontinuous Galerkin method, we now choose

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

and search for $u_h \in V_h$ satisfying

$$a_h(u_h, v_h) = (f, v_h) \qquad \text{for all } v_h \in V_h. \tag{9.12}$$

[4] Other choices are discussed in [Arnold et al. 2002].

To show well-posedness using the Banach–Nečas–Babuška theorem, we need to show continuity and coercivity of a_h with respect to appropriate norms. We again postpone continuity (in an equivalent norm) to later, and address coercivity with respect to the discrete norm

$$|||v_h|||^2 := \|\nabla_h v\|^2_{L^2(\Omega)^n} + |v_h|^2_{\Gamma_h},$$

with the *jump seminorm*

$$|v_h|^2_{\Gamma_h} := \sum_{F\in\partial\mathcal{T}_h} h_F^{-1} \, \|[\![v_h]\!]\|^2_{L^2(F)^n}\,;$$

for $F \subset \partial\Omega$ we use the convention that $u = 0$ outside of Ω. This is indeed a norm on V_h since $|||v_h||| = 0$ implies first that v_h is piecewise constant; and since the function vanishes on the boundary and the interface jumps are zero, these constants are zero.

Again we postpone continuity to later and first verify the coercivity of a_h with respect to $|||\cdot|||$.

Lemma 9.7. *For all $\eta > 0$ sufficiently large, there exists a $C > 0$ independent of h such that*

$$a_h(u_h, u_h) \geq C\, |||u_h|||^2 \qquad \text{for all } u_h \in V_h.$$

Proof. For arbitrary $u_h \in V_h$, we have using the definition of the broken gradient and the jump seminorm that

$$a_h(u_h, u_h) = \|\nabla_h u_h\|^2_{L^2(\Omega)^n} - 2 \sum_{F\in\partial\mathcal{T}_h} \int_F \{\{\nabla_h u_h\}\} \cdot [\![u_h]\!] \, ds + \eta |u_h|^2_{\Gamma_h}.$$

Since the second term has the wrong sign, we need to absorb it into the other terms. For this, we use that for any piecewise smooth v, w and every $F \in \partial\mathcal{T}_h$, the Cauchy–Schwarz inequality yields

$$\begin{aligned}\int_F \{\{\nabla_h v\}\} \cdot [\![w]\!] \, ds &= \int_F \frac{1}{2} \left(\nabla_h v|_{K_1} + \nabla_h v|_{K_2}\right) \cdot [\![w]\!] \, ds \\ &\leq \frac{h_F^{1/2}}{2} \left(\left\|\nabla_h v|_{K_1}\right\|^2_{L^2(F)^n} + \left\|\nabla_h v|_{K_2}\right\|^2_{L^2(F)^n}\right)^{\frac{1}{2}} h_F^{-1/2} \, \|[\![w]\!]\|_{L^2(F)^n}.\end{aligned}$$

Summing over all faces and using the fact that each interior face occurs twice and that for boundary faces we set $v = w = 0$ outside of Ω, we obtain

$$\sum_{F\in\partial\mathcal{T}_h} \int_F \{\{\nabla_h v\}\} \cdot [\![w]\!] \, ds \leq \left(\sum_{K\in\mathcal{T}_h} \sum_{F\subset\partial K} h_F \, \|\nabla_h v\|^2_{L^2(F)^n}\right)^{\frac{1}{2}} |w|_{\Gamma_h}. \tag{9.13}$$

For $v_h \in V_h$, we can further use the discrete trace inequality (9.6) together with $h_F \leq h_K$ for all faces F of K to arrive at

$$\begin{aligned}\sum_{F\in\partial\mathcal{T}_h} \int_F \{\{\nabla_h v_h\}\} \cdot [\![w]\!] \, ds &\leq C \left(\sum_{K\in\mathcal{T}_h} \|\nabla_h v_h\|^2_{L^2(K)^n}\right)^{\frac{1}{2}} |w|_{\Gamma_h} \\ &= C \, \|\nabla_h v_h\|_{L^2(\Omega)^n} \, |w|_{\Gamma_h}.\end{aligned} \tag{9.14}$$

Applying this estimate for $v_h = w = u_h$ together with the generalized Young inequality $ab \le \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2$ for arbitrary $\varepsilon > 0$ then yields that

$$\begin{aligned} a_h(u_h, u_h) &\ge \|\nabla_h u_h\|^2_{L^2(\Omega)^n} - 2C\,\|\nabla_h u_h\|_{L^2(\Omega)^n}\,|u_h|_{\Gamma_h} + \eta|u_h|^2_{\Gamma_h} \\ &\ge (1 - C\varepsilon)\,\|\nabla_h u_h\|^2_{L^2(\Omega)^n} + (\eta - C\varepsilon^{-1})|u_h|^2_{\Gamma_h}. \end{aligned}$$

We can now first choose $\varepsilon > 0$ sufficiently small that the first term is positive, and then $\eta > 0$ sufficiently large that the second term is positive, which implies coercivity in the desired norm. □

For error estimates, we again need to show boundedness of a_h on a space containing both discrete and exact solutions. Here we assume that the exact solution of the Poisson equation satisfies

$$u \in U_* := H^1_0(\Omega) \cap H^2(\Omega),$$

see Theorem 2.9 or Theorem 2.10, and endow $U(h) := U_* + V_h$ with the norm

$$|||w|||^2_* := |||w|||^2 + \sum_{K\in\mathcal{T}_h} h_K\,\|\nabla_h w\|^2_{L^2(\partial K)^n}.$$

With respect to this norm, a_h is bounded in u.

Lemma 9.8. *If $\mathcal{T}_h$ is a shape-regular triangulation of Ω, then there exists a constant $C > 0$ independent of h such that*

$$a_h(u, v_h) \le C\,|||u|||_*\,|||v_h||| \qquad \text{for all } u \in U(h),\ v_h \in V_h.$$

Proof. We estimate for $u \in U_*$ and $v_h \in V_h$ each term in $a_h(u, v_h)$ separately.

(i) For the first term, the Cauchy–Schwarz inequality immediately yields

$$(\nabla_h u, \nabla_h v_h) \le \|\nabla_h u\|_{L^2(\Omega)^n}\,\|\nabla_h v_h\|_{L^2(\Omega)^n}.$$

(ii) For the second term, we apply the estimate (9.14) for $v = v_h$ and $w = u$ to obtain

$$\sum_{F\in\partial\mathcal{T}_h}\int_F [\![u]\!]\cdot\{\{\nabla_h v_h\}\} \le C\,\|\nabla_h v_h\|_{L^2(\Omega)^n}\,|u|_{\Gamma_h}.$$

(iii) For the third term, we apply the estimate (9.13) for $v = u$ and $w = v_h$ to obtain

$$\sum_{F\in\partial\mathcal{T}_h}\int_F [\![v_h]\!]\cdot\{\{\nabla_h u\}\} \le \left(\sum_{K\in\mathcal{T}_h} h_K\,\|\nabla_h u\|^2_{L^2(\partial K)^n}\right)^{\frac{1}{2}} |v_h|_{\Gamma_h},$$

using again that $h_F \le h_K$ for all faces F of K.

(iv) For the last term, we again obtain from the Cauchy–Schwarz inequality that

$$\sum_{F\in\partial\mathcal{T}_h}\int_F \frac{\eta}{h_F}\,[\![u]\!]\,[\![v_h]\!]\,ds \leq \eta|u|_{\Gamma_h}|v_h|_{\Gamma_h}.$$

Since all of the terms appearing on the right-hand sides are parts of the definition of $|\!|\!|u|\!|\!|_*$ and $|\!|\!|v_h|\!|\!|$, respectively, we conclude the desired estimate. □

Note that for $u_h \in V_h$, we could have used in step (iii) the estimate (9.14) as well to avoid the extra term in the definition of $|\!|\!|u|\!|\!|_*$. From this, we have for $u_h \in V_h$ that

$$a_h(u_h, v_h) \leq C\,|\!|\!|u|\!|\!|\,|\!|\!|v_h|\!|\!|\,,$$

i.e., the continuity necessary to apply the Banach–Nečas–Babuška theorem.

Corollary 9.9. *If $\mathcal{T}_h$ is a shape-regular triangulation of Ω and η is sufficiently large, then there exists a unique solution $u_h \in V_h$ to* (9.12)*. Furthermore, there exists a constant $C > 0$ such that*

$$|\!|\!|u_h|\!|\!| \leq C\,\|f\|_{L^2(\Omega)}\,.$$

With the same arguments as in Lemma 9.5, one can show that any $u \in H^2(\Omega)$ satisfies $[\![u]\!]_F = 0$ and $[\![\nabla u]\!]_F = 0$. Hence the exact solution $u \in U_*$ satisfies (9.12), and we can apply the second Strang lemma to obtain

$$|\!|\!|u - u_h|\!|\!| \leq C \inf_{w_h\in V_h} |\!|\!|u - w_h|\!|\!|_*\,.$$

Estimating the best approximation error by the interpolation error and applying the usual estimates for each term in $|\!|\!|\cdot|\!|\!|_*$ (noting that the appearance of h_K in the gradient term compensates for the lower power h_K^{k-1} in the corresponding estimate), we obtain for a solution $u \in H^{k+1}(\Omega)$ the a priori error estimate

$$|\!|\!|u - u_h|\!|\!| \leq Ch^k|u|_{H^{k+1}(\Omega)}.$$

Due to the face term in $|\!|\!|\cdot|\!|\!|$, this estimate is optimal; a duality trick then yields a convergence rate of $O(h^{k+1})$ for the discretization error in the L^2 norm (which is useful even for $k = 0$).

9.5 IMPLEMENTATION

As in the standard Galerkin approach, the assembly of the stiffness matrix is carried out by choosing a suitable nodal basis $\varphi_1, \dots, \varphi_N$ of V_h and computing the entries $a_h(\varphi_i, \varphi_j)$ element-wise by transformation to a reference element. For discontinous Galerkin methods, there are two important differences:

1. Since the functions in V_h can be discontinuous across elements, the degrees of freedom of each element decouple from the remaining elements.
2. There are terms arising from integration over interior as well as boundary faces.

These require some modifications to the assembly procedure described in Section 7.2.

Due to the first point, we can take each basis function φ_i to have support on only one element. Our set of global basis functions is thus just the union of the sets of local basis functions on each element $K \in \mathcal{T}_h$ (extended to zero outside K), which are constructed as in Chapter 4. Note that this implies that nodes (the interpolation points for each degree of freedom) common to multiple element domains have to be treated as distinct (e.g., a node on a vertex where m elements meet corresponds to m degrees of freedom, one for each element). The dimension of V_h is thus equal to the sum of the local degrees of freedom over all elements, and thus greater than for standard finite elements.

In particular, if the global basis functions are enumerated such that the local basis functions in each element are numbered contiguously, the mass matrix $\mathbf{M}$ with elements $M_{ij} = (\varphi_i, \varphi_j)$ is then *block diagonal*, where each block corresponds to one element. For the stiffness matrix $\mathbf{K}$, the terms arising from volume integrals are similarly block diagonal, but they are coupled via the terms arising from the integrals over interior faces. It is thus convenient to separately assemble the contributions to the bilinear form a_h from volume integrals, interior face integrals and boundary face integrals:

- The *volume terms* are assembled as described in Section 7.2, making use of the simple form of the local-to-global index.
- For the *interior face terms*, one needs a list `interfaces` of interior faces, which contains for each face F the two elements K_1, K_2 sharing it, as well as the location of the face relative to each element. For each pair of basis functions from the two elements (obtained via the list `elements`), one can then (by transformation to the reference element and, if necessary, numerical quadrature) compute the corresponding integrals, recalling for the computation of jumps and averages that each local basis function is zero outside its element, and that the unit normals can be obtained from the reference element (where they are known) by transformation.
- The *boundary terms* are similarly assembled using the list `bdy_faces`, where for advection-reaction equations, one has to check on each face the sign of $\beta(x) \cdot \nu_F$ to decide whether it is part of the inflow boundary $\partial\Omega^-$ where the boundary condition has to be prescribed.

10 MIXED METHODS

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 $a(u, v) = F(v)$ for all $v \in V$ is the unique minimizer of $J(v) = \frac{1}{2}a(v, v) - F(v)$. 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 $(u, \lambda) \in V \times M$:

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

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 conditions

$$(\mathcal{S}) \qquad \begin{cases} a(u,v) + b(v,\lambda) = \langle f, v\rangle_{V^*,V} & \text{for all } v \in V, \\ \qquad\qquad b(u,\mu) = \langle g, \mu\rangle_{M^*,M} & \text{for all } \mu \in M. \end{cases}$$

In principle, we can obtain existence and uniqueness of (u, λ) by considering $(\mathcal{S})$ as a variational problem for a bilinear form $c : (V \times M) \times (V \times M) \to \mathbb{R}$ 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

$$(10.1) \qquad \begin{cases} Au + B^*\lambda = f & \text{in } V^*, \\ \qquad\quad Bu = g & \text{in } M^*. \end{cases}$$

From this, we can see the following: If B were invertible, the existence and uniqueness first of u and then of λ would follow immediately. In the (more realistic) case that B has a nontrivial null space

$$\ker B = \{x \in V : b(x,\mu) = 0 \text{ for all } \mu \in M\}$$

(e.g., constant functions in the case $Bu = \nabla \cdot u$), we have to require that A is injective on it to obtain a unique u. Existence of λ then follows if B^* is surjective. To verify these conditions, we follow the general approach of the Banach–Nečas–Babuška theorem.

Theorem 10.1 (Brezzi splitting theorem). *Assume that*

(i) $a : V \times V \to \mathbb{R}$ satisfies the conditions of Theorem 8.1 for $U = V = \ker B$ and

(ii) $b : V \times M \to \mathbb{R}$ satisfies for $\beta > 0$ the condition

$$(10.2) \qquad \inf_{\mu \in M} \sup_{v \in V} \frac{b(v,\mu)}{\|v\|_V \|\mu\|_M} \geq \beta.$$

Then there exists a unique solution $(u, \lambda) \in V \times M$ to $(\mathcal{S})$ satisfying

$$\|u\|_V + \|\lambda\|_M \leq C(\|f\|_{V^*} + \|g\|_{M^*}).$$

Condition (ii) is an inf–sup condition for B^* (since the infimum is taken over the test functions μ) and is known as the *Ladyzhenskaya–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 following the proof of Theorem 8.1, we deduce that the LBB condition implies that B^* has closed range and therefore

$$\operatorname{ran} B^* = (\ker B)^\perp = \left\{v^* \in V^* : \langle v^*, v\rangle_{V^*,V} = 0 \text{ for all } v \in \ker B\right\}.$$

In addition, B^* is injective on M with

$$\beta \|\mu\|_M \leq \|B^*\mu\|_{V^*} \tag{10.3}$$

holds for all $\mu \in M$. By the closed range theorem, B has closed range as well and hence is surjective on $\operatorname{ran} B = (\ker B^*)^\perp = (\{0\})^\perp = M^*$. Thus for any $g \in M^*$ there exists a $\tilde u_g \in V$ satisfying $B\tilde u_g = g$. Since B is not injective, $\tilde u_g$ is not unique, nor can its norm necessarily be bounded by that of g (since one can add to $\tilde u_g$ any element in $\ker B$). However, among the possible solutions, we can find (at least) one that is bounded by applying the Hahn–Banach extension theorem. Let $v^* \in (\ker B)^\perp = \operatorname{ran} B^* \subset V^*$ be given. By the above, there then exists a unique $\lambda \in M$ such that $B^*\lambda = v^*$ and $\|\lambda\|_M \leq \frac{1}{\beta}\|v^*\|_{V^*}$. Since $V \subset (V^*)^*$, we can write

$$\langle \tilde u_g, v^*\rangle_{(V^*)^*,V^*} = \left\langle B^*\lambda, \tilde u_g\right\rangle_{V^*,V} = \langle g, \lambda\rangle_{M^*,M} \leq \|g\|_{M^*}\|\lambda\|_M \leq \frac{1}{\beta}\|g\|_{M^*}\|v^*\|_{V^*}.$$

This implies that $\tilde u_g$ is bounded as a linear functional on $(\ker B)^\perp \subset V^*$, and in particular that $\left\|\tilde u_g\right\|_{((\ker B)^\perp)^*} \leq \frac{1}{\beta}\|g\|_{M^*}$. The Hahn–Banach extension theorem thus yields existence of a $u_g \in (V^*)^* = V$ with $u_g = \tilde u_g$ on $(\ker B)^\perp = \operatorname{ran} B^*$ and

$$\left\|u_g\right\|_V = \left\|\tilde u_g\right\|_{((\ker B)^\perp)^*} \leq \frac{1}{\beta}\|g\|_{M^*}. \tag{10.4}$$

In addition, $Bu_g = g$ as well, since for all $\mu \in M$, we have that $B^*\mu \in \operatorname{ran} B^* = (\ker B)^\perp$ and hence by the extension property that

$$\left\langle Bu_g, \mu\right\rangle_{M^*,M} = \left\langle B^*\mu, u_g\right\rangle_{V^*,V} = \left\langle B^*\mu, \tilde u_g\right\rangle_{V^*,V} = \langle g, \mu\rangle_{M^*,M}.$$

Due to condition (i), A is an isomorphism from $\ker B$ to $(\ker B)^*$. Considering $f - Au_g \in V^*$ as a bounded linear form on $\ker B \subset V$, we can thus find a unique $u_f \in \ker B$ satisfying

$$Au_f = f - Au_g \quad \text{in } (\ker B)^* \tag{10.5}$$

(but not necessarily in V^*!) and

$$\left\|u_f\right\|_V \leq \frac{1}{\alpha}\left\|f - Au_g\right\|_{(\ker B)^*} \leq \frac{1}{\alpha}(\|f\|_{V^*} + C\left\|u_g\right\|_V), \tag{10.6}$$

where $\alpha > 0$ and $C > 0$ are the constants in the inf–sup and continuity conditions for a, respectively, and we have used that $f \in V^*$ such that $\|f\|_{(\ker B)^*} \leq \|f\|_{V^*}$ by definition of the dual norm and the fact that $\ker B \subset V$ is endowed with the same norm as V.

Now set $u = u_f + u_g \in V$ and consider $f - Au \in V^*$, which due to (10.5) satisfies for all $v \in \ker B$ that $\langle f - Au, v\rangle_{(\ker B)^*,(\ker B)} = 0$. This implies that $f - Au \in (\ker B)^\perp$, and the surjectivity of B^* on $(\ker B)^\perp = \operatorname{ran} B^*$ yields the existence of a $\lambda \in M$ satisfying

$$B^*\lambda = f - Au \quad \text{in } V^*$$

and

$$\|\lambda\|_M \leq \frac{1}{\beta}(\|f\|_{V^*} + C\,\|u\|_V). \tag{10.7}$$

Since $u_f \in \ker B$ and hence

$$Bu = Bu_g = g \quad \text{in } M^*,$$

we have thus found $(u, \lambda) \in V \times M$ satisfying ($\mathcal{S}$), and the claimed estimate follows by combining (10.4), (10.6) and (10.7).

To show uniqueness, 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.,

$$\begin{cases} A\overline{u} + B^*\overline{\lambda} = 0 & \text{in } V^*, \\ \qquad\quad B\overline{u} = 0 & \text{in } M^*. \end{cases}$$

The second equation yields $\overline{u} \in \ker B$, and the inf–sup condition for A on $\ker B$ implies

$$\alpha\,\|\overline{u}\|_V^2 \leq a(\overline{u}, \overline{u}) = a(\overline{u}, \overline{u}) + b(\overline{u}, \overline{\lambda}) = 0.$$

Since $\overline{u} = 0$, it follows from the first equation that $B^*\overline{\lambda} = 0$ and thus from the injectivity of B^* that $\overline{\lambda} = 0$. □

10.2 GALERKIN APPROXIMATION OF SADDLE POINT PROBLEMS

For the Galerkin approximation of ($\mathcal{S}$), we again choose subspaces $V_h \subset V$ and $M_h \subset M$ and look for $(u_h, \lambda_h) \in V_h \times M_h$ satisfying

$$\begin{cases} a(u_h, v_h) + b(v_h, \lambda_h) = \langle f, v_h\rangle_{V^*,V} & \text{for all } v_h \in V_h, \\ \qquad\qquad\quad b(u_h, \mu_h) = \langle g, \mu_h\rangle_{M^*,M} & \text{for all } \mu_h \in M_h. \end{cases} \tag{$\mathcal{S}_h$}$$

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 that in Theorem 10.1. For its statement, we 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*

$$\inf_{u_h \in \ker B_h} \sup_{v_h \in \ker B_h} \frac{a(u_h, v_h)}{\|u_h\|_V \|v_h\|_V} \geq \alpha_h, \tag{10.8}$$

$$\inf_{\mu_h \in M_h} \sup_{v_h \in V_h} \frac{b(v_h, \mu_h)}{\|v_h\|_V \|\mu_h\|_M} \geq \beta_h. \tag{10.9}$$

Then there exists a unique solution $(u_h, \lambda_h) \in V_h \times M_h$ to $(\mathcal{S}_h)$ satisfying

$$\|u_h\|_V + \|\lambda_h\|_M \leq C(h)(\|f\|_{V^*} + \|g\|_{M^*}).$$

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 the discrete BNB Theorem 8.2. □

Note that in general, this is a non-conforming approach since even for $V_h \subset V$ and $M_h \subset M$, as we do not necessarily have that B_h is the restriction of B to V_h (i.e., $B(V_h) \not\subset M_h^*$) or that $\ker B_h$ is a subspace of $\ker B$. Hence, the discrete inf–sup conditions do not follow from their continuous counterparts. 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 that the LBB condition (10.2) is satisfied. Then the discrete LBB condition (10.9) is satisfied if and only if there exists a linear operator $\Pi_h : V \to V_h$ such that*

$$b(\Pi_h v, \mu_h) = b(v, \mu_h) \quad \text{for all } \mu_h \in M_h,$$

and there exists a $\gamma_h > 0$ such that

$$\|\Pi_h v\|_V \leq \gamma_h \|v\|_V \quad \text{for all } v \in V.$$

Proof. Assume that such a Π_h exists. Since $\operatorname{ran} \Pi_h \subset V_h$, we have for all $\mu_h \in M_h \subset M$ that

$$\sup_{v_h \in V_h} \frac{b(v_h, \mu_h)}{\|v_h\|_V} \geq \sup_{v \in V} \frac{b(\Pi_h v, \mu_h)}{\|\Pi_h v\|_V} \geq \sup_{v \in V} \frac{b(v, \mu_h)}{\gamma_h \|v\|_V} \geq \frac{\beta}{\gamma_h} \|\mu_h\|_M,$$

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 a continuous right inverse. Furthermore, for any $v \in V$ we can consider $Bv \in M^*$ as a linear functional on $M_h \subset M$ only. Hence for any $v \in V$, there exists a $\pi_h \in V_h$ such that $B_h(\pi_h) = Bv|_{M_h} \in M_h^*$, i.e., $b(\pi_h, \mu_h) = b(v, \mu_h)$ for all $\mu_h \in M_h \subset M$, and

$$\beta_h \|\pi_h\|_V \leq \|Bv\|_{M_h^*} \leq C \|v\|_V.$$

We thus obtain the desired operator by defining Π_h as the (linear) mapping $v \mapsto \pi_h$. □

The operator Π_h is called *Fortin projector*. From the proof, we can see that the discrete LBB condition holds with a constant independent of h if and only if the Fortin projector is uniformly bounded in h (i.e., if $\gamma_h \equiv \gamma$).

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 $C(h) > 0$ such that*

$$\|u-u_h\|_V + \|\lambda-\lambda_h\|_M \leq C(h)\left(\inf_{w_h\in V_h}\|u-w_h\|_V + \inf_{\mu_h\in M_h}\|\lambda-\mu_h\|_M\right).$$

Proof. For arbitrary $w_h \in V_h$, consider the restriction of $B(u-w_h) \in M^*$ to M_h Due to the discrete LBB condition, the operator $B_h : V_h \to M_h^*$ is surjective and has a continuous right inverse. Hence, there exists $r_h \in V_h$ satisfying $B_h r_h = B(u-w_h)|_{M_h} \in M_h^*$, i.e.,

$$b(r_h,\mu_h) = b(u-w_h,\mu_h) \qquad \text{for all } \mu_h \in M_h \subset M$$

and

$$\beta_h \|r_h\|_V \leq C\,\|u-w_h\|_V\,.$$

Furthermore, $z_h := r_h + w_h$ satisfies

$$b(z_h,\mu_h) = b(u,\mu_h) = \langle g,\mu_h\rangle_{M^*,M} = b(u_h,\mu_h) \qquad \text{for all } \mu_h \in M_h \subset M,$$

which implies that $u_h - z_h \in \ker B_h$. The discrete inf–sup condition (10.8) thus yields

$$\begin{aligned}\alpha_h\,\|u_h-z_h\|_V &\leq \sup_{v_h\in\ker B_h}\frac{a(u_h-z_h,v_h)}{\|v_h\|_V}\\ &= \sup_{v_h\in\ker B_h}\frac{a(u_h-u,v_h)+a(u-z_h,v_h)}{\|v_h\|_V}\\ &= \sup_{v_h\in\ker B_h}\frac{b(v_h,\lambda-\lambda_h)+a(u-z_h,v_h)}{\|v_h\|_V},\end{aligned} \tag{10.10}$$

by taking the difference of the first equations of ($\mathcal{S}$) and ($\mathcal{S}_h$). For any $v_h \in \ker B_h$ and $\mu_h \in M_h$, we have

$$b(v_h,\lambda_h) = 0 = b(v_h,\mu_h)$$

and hence from the continuity of a and b that

$$\alpha_h\,\|u_h-z_h\|_V \leq C(\|u-z_h\|_V + \|\lambda-\mu_h\|_M)$$

for arbitrary $\mu_h \in M_h$. Using the triangle inequality, we thus obtain

$$\begin{aligned}\|u-u_h\|_V &\leq \|u-z_h\|_V + \|z_h-u_h\|_V\\ &\leq \left(1+\frac{C}{\alpha_h}\right)\|u-z_h\|_V + \frac{C}{\alpha_h}\,\|\lambda-\mu_h\|_M\end{aligned} \tag{10.11}$$

and, by definition of z_h

$$\|u - z_h\|_V \leq \|u - w_h\|_V + \|r_h\|_V \leq \left(1 + \frac{C}{\beta_h}\right) \|u - w_h\|_V . \tag{10.12}$$

To estimate $\|\lambda - \lambda_h\|_M$, we again use that for all $w_h \in V_h$ and $\mu_h \in M_h$,

$$a(u - u_h, w_h) = b(w_h, \lambda - \lambda_h) = b(w_h, \lambda - \mu_h) + b(w_h, \mu_h - \lambda_h).$$

After rearrangement, the discrete LBB condition and the continuity of a and b thus yield

$$\beta_h \|\lambda_h - \mu_h\|_M \leq C(\|u - u_h\|_V + \|\lambda - \mu_h\|_M).$$

Applying the triangle inequality again, we obtain

$$\begin{aligned} \|\lambda - \lambda_h\|_M &\leq \|\lambda - \mu_h\|_M + \|\lambda_h - \mu_h\|_M \\ &\leq \left(1 + \frac{C}{\beta_h}\right) \|\lambda - \mu_h\|_M + \frac{C}{\beta_h} \|u - u_h\|_V . \end{aligned} \tag{10.13}$$

Combining (10.11), (10.12), and (10.13) and taking the infimum over all $w_h \in V_h$ and $\mu_h \in M_h$ yields the claimed estimate. □

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 obtain an independent estimate for u.

Corollary 10.5. *If* $\ker B_h \subset \ker B$, *then there exists a constant* $C(h) > 0$ *such that*

$$\|u - u_h\|_V \leq C(h) \inf_{w_h \in V_h} \|u - w_h\|_V .$$

Proof. The assumption in particular implies that $b(v_h, \lambda - \lambda_h) = 0$ for all $v_h \in \ker B_h$, and hence (10.10) yields

$$\alpha_h \|u_h - z_h\|_V \leq C \|u - z_h\|_V .$$

Continuing as above, we obtain the claimed estimate. □

10.3 MIXED METHODS 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

[1] see, e.g., [Braess 2007, Chapter III.6], [Ern & Guermond 2004, Chapter 4]

a very simple example. Consider the Poisson equation $-\Delta u = f$ on $\Omega \subset \mathbb{R}^n$ with homogeneous Dirichlet conditions. If we again introduce $\sigma = \nabla u \in L^2(\Omega)^n$, we can write this equation as

$$\begin{cases} \nabla u - \sigma = 0, \\ -\nabla \cdot \sigma = f. \end{cases} \tag{10.14}$$

This system can be formulated in variational form in two different ways, called *primal* and *dual* approach, respectively.

Primal mixed method The primal approach consists in (formally) integrating by parts in the second equation of (10.14) and looking for $(\sigma, u) \in L^2(\Omega)^n \times H_0^1(\Omega)$ satisfying

$$\begin{cases} (\sigma, \tau) - (\tau, \nabla u) = 0 & \text{for all } \tau \in L^2(\Omega)^n, \\ \qquad -(\sigma, \nabla v) = -(f, v) & \text{for all } v \in H_0^1(\Omega). \end{cases} \tag{10.15}$$

This fits into the abstract framework of Section 10.1 by setting $V := L^2(\Omega)^n$, $M := H_0^1(\Omega)$,

$$a(\sigma, \tau) = (\sigma, \tau), \qquad b(\sigma, v) = -(\sigma, \nabla v).$$

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 $v \in H_0^1(\Omega) = M$ in

$$\sup_{\tau \in V} \frac{b(\tau, v)}{\|\tau\|_V} = \sup_{\tau \in V} \frac{-(\tau, \nabla v)}{\|\tau\|_{L^2(\Omega)^n}} \geq \frac{(\nabla v, \nabla v)}{\|\nabla v\|_{L^2(\Omega)^n}} = |v|_{H^1(\Omega)} \geq c_\Omega^{-1} \|v\|_M$$

using the Poincaré inequality (Theorem 2.6), i.e., the LBB condition with $\beta = c_\Omega^{-1}$. Theorem 10.1 thus yields the existence and uniqueness of the solution (σ, u) to (10.15).

To obtain a stable mixed finite element method, we take a shape-regular affine triangulation $\mathcal{T}_h$ of Ω and set for $k \geq 1$

$$\begin{aligned} V_h &:= \left\{ \tau_h \in L^2(\Omega)^n : \tau_h|_K \in P_{k-1}(K)^n \text{ for all } K \in \mathcal{T}_h \right\}, \\ M_h &:= \left\{ v_h \in C^0(\Omega) : v_h|_K \in P_k(K) \text{ for all } K \in \mathcal{T}_h \right\} \cap M. \end{aligned}$$

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 $\nabla M_h \subset V_h$, e.g., the gradient of any piecewise affine continuous function is piecewise constant. Hence, the $L^2(\Omega)^n$ projection from V on V_h (which is continuous with norm $\gamma_h = 1$) 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 with constant $\beta_h = \beta$ independent of h, and we obtain existence of and (from Theorem 10.4 combined with the usual interpolation error estimates) a priori estimates for the mixed finite element discretization of (10.15) (which coincide with those from Section 6.1).

Dual mixed method Instead of integrating by parts in the second equation, we can formally integrate by parts in the first equation of (10.14). To make this well-defined, we set

$$H^{\operatorname{div}}(\Omega) := \left\{\tau \in L^2(\Omega)^n : \operatorname{div}\tau \in L^2(\Omega)\right\},$$

endowed with the graph norm

$$\|\tau\|^2_{H^{\operatorname{div}}(\Omega)} := \|\tau\|^2_{L^2(\Omega)^n} + \|\operatorname{div}\tau\|^2_{L^2(\Omega)}.$$

Since $C^\infty(\overline{\Omega})^n$ is dense in $L^2(\Omega)^n \supset H^{\operatorname{div}}(\Omega)$, one can show that $\tau \in H^{\operatorname{div}}(\Omega)$ has a well-defined *normal trace* $(\tau|_{\partial\Omega} \cdot \nu) \in H^{-1/2}(\partial\Omega)$, and that for any $\tau \in H^{\operatorname{div}}(\Omega)$ and $w \in H^1(\Omega)$ the integration by parts formula

$$\int_\Omega (\operatorname{div}\tau) w\, dx + \int_\Omega \tau \cdot \nabla w\, dx = \int_{\partial\Omega} (\tau \cdot \nu) w\, dx \tag{10.16}$$

holds.[2] Similarly to Theorem 2.4, one can show that for a partition $\{\Omega_j\}_{j\in J}$ of Ω,

$$\left\{\tau \in L^2(\Omega)^n : \tau|_{\Omega_j} \in H^1(\Omega_j) \text{ and } \tau|_{\Omega_j} \cdot \nu = \tau|_{\Omega_i} \cdot \nu \text{ on all } \overline{\Omega}_j \cap \overline{\Omega}_i \neq \emptyset\right\} \subset H^{\operatorname{div}}(\Omega)$$

holds, i.e., piecewise differentiable functions with continuous normal traces across elements are in $H^{\operatorname{div}}(\Omega)$. This will be important for constructing conforming approximations of $H^{\operatorname{div}}(\Omega)$.

After integrating by parts in (10.14) and using that $u|_{\partial\Omega} = 0$, we are therefore looking for $(\sigma, u) \in H^{\operatorname{div}}(\Omega) \times L^2(\Omega)$ satisfying

$$\begin{cases} (\sigma, \tau) + (\operatorname{div}\tau, u) = 0 & \text{for all } \tau \in H^{\operatorname{div}}(\Omega), \\ (\operatorname{div}\sigma, v) = -(f, v) & \text{for all } v \in L^2(\Omega). \end{cases} \tag{10.17}$$

(Note that in contrast to the standard – and primal – formulation, the Dirichlet condition appears here as the natural boundary condition.) This formulation fits into the abstract framework of Section 10.1 by setting $V := H^{\operatorname{div}}(\Omega)$, $M := L^2(\Omega)$,

$$a(\sigma, \tau) = (\sigma, \tau), \qquad b(\sigma, v) = (\operatorname{div}\sigma, v).$$

Boundedness of a and b follows directly from the Cauchy–Schwarz inequality. Now we note that

$$\ker B = \left\{\tau \in H^{\operatorname{div}}(\Omega) : (\operatorname{div}\tau, v) = 0 \text{ for all } v \in L^2(\Omega)\right\}.$$

Since $\operatorname{div}\tau \in L^2(\Omega)$ and thus $\|\operatorname{div}\tau\|^2_{L^2(\Omega)} = 0$ for all $\tau \in \ker B \subset H^{\operatorname{div}}(\Omega)$, this implies

$$a(\tau, \tau) = \|\tau\|^2_{L^2(\Omega)^n} = \|\tau\|^2_{H^{\operatorname{div}}(\Omega)} \qquad \text{for all } v \in \ker B,$$

yielding coercivity of a with constant $\alpha = 1$. For verification of the LBB condition, we make use of the following lemma showing surjectivity of B on M. For simplicity, we assume from here on that Ω either has a C^1 boundary or is convex.

[2] e.g., [Boffi, Brezzi & Fortin 2013, Lemma 2.1.1]

Lemma 10.6. *For any $f \in L^2(\Omega)$, there exists a function $\tau \in H^1(\Omega)^n$ with $\operatorname{div}\tau = f$ and $\|\tau\|_{H^1(\Omega)^n} \le C\|f\|_{L^2(\Omega)}$.*

Proof. Due to the regularity of Ω, we can apply Theorem 2.9 or Theorem 2.10 to obtain for given $f \in L^2(\Omega)$ a solution $u \in H^2(\Omega) \cap H^1_0(\Omega)$ to the Poisson equation

$$(\nabla u, \nabla v) = (f, v) \qquad \text{for all } v \in H^1_0(\Omega)$$

satisfying $\|u\|_{H^2(\Omega)} \le C\|f\|_{L^2(\Omega)}$. Now set $\tau := -\nabla u \in H^1(\Omega)^n$ and observe that

$$(f, v) = -(\tau, \nabla v) \qquad \text{for all } v \in H^1_0(\Omega),$$

and thus $f = \operatorname{div}\tau$ by definition of the weak derivative. The a priori bound on τ then follows from the fact that $\|\nabla u\|_{H^1(\Omega)^n} \le \|u\|_{H^2(\Omega)}$. □

Using this lemma and the inclusion $H^1(\Omega)^n \subset H^{\text{div}}(\Omega)$, we immediately obtain for any $v \in M$ and corresponding τ_v with $\operatorname{div}\tau_v = v$ that

$$\sup_{\tau\in V} \frac{b(\tau, v)}{\|\tau\|_V} = \sup_{\tau\in V} \frac{(\operatorname{div}\tau, v)}{\|\tau\|_{H^{\text{div}}(\Omega)}} \ge \frac{(\operatorname{div}\tau_v, v)}{\|\tau_v\|_{H^{\text{div}}(\Omega)}} \ge \frac{(v, v)}{C\|v\|_{L^2(\Omega)}} = \frac{1}{C}\|v\|_{L^2(\Omega)},$$

which verifies the LBB condition for $\beta = C^{-1}$. From Theorem 10.1 we thus obtain existence of a unique solution $(\sigma, u) \in V \times M$ to (10.17) as well as the estimate

$$\|\sigma\|_{H^{\text{div}}(\Omega)} + \|u\|_{L^2(\Omega)} \le C\|f\|_{L^2(\Omega)}.$$

Although this initially yields only a solution $u \in L^2(\Omega)$, one can then use the first equation of (10.17) to show that u has a weak derivative and (using integration by parts) satisfies the boundary conditions; i.e., $u \in H^1_0(\Omega)$ as expected.

We now construct conforming finite element discretizations of V and M. Let $\mathcal{T}_h$ be a shape-regular affine triangulation of $\Omega \subset \mathbb{R}^n$. For $M = L^2(\Omega)$, we again take piecewise (discontinuous) polynomials of degree $k \ge 0$, i.e.,

$$M_h = \left\{v_h \in L^2(\Omega) : v_h|_K \in P_k(K) \text{ for all } K \in \mathcal{T}_h\right\}.$$

For $V = H^{\text{div}}(\Omega)$, we construct a space V_h of piecewise polynomials on the same triangulation that satisfy the two key properties of V: Functions $\tau_h \in V_h$ have continuous normal traces across elements, and the divergence is surjective from V_h to M_h. One possible choice is

$$V_h = \left\{\tau_h \in H^{\text{div}}(\Omega) : \tau_h|_K \in RT_k(K) \text{ for all } K \in \mathcal{T}_h\right\},$$

with

$$\begin{aligned} RT_k(K) &= P_k(K)^n + xP_k(K) := \{p_1 + p_2\,x : p_1 \in P_k(K)^n, p_2 \in P_k(K)\} \\ &= P_k(K)^n \oplus xP_k^0(K), \end{aligned}$$

where

$$P_k^0(K) = \left\{ \sum_{|\alpha|=k} c_\alpha x^\alpha : c_\alpha \in \mathbb{R} \right\}$$

is the space of *homogeneous* polynomials of degree k (which is chosen in order to have a unique representation). This construction yields the following properties, which guarantee a conforming $H^{\mathrm{div}}(\Omega)$ discretization.

Lemma 10.7. *For $\tau_h \in RT_k(K)$, we have*

(i) $\operatorname{div} \tau_h \in P_k(K)$ *and*

(ii) $\tau_h|_F \cdot \nu_F \in P_k(F)$ *for every* $F \subset \partial K$.

The verification is a straightforward computation (recalling that $x \cdot \nu_F(x)$ is constant for every $x \in F$). It remains to specify the degrees of freedom, of which we need

$$\dim RT_k(K) = \begin{cases} (k+1)(k+3) & \text{for } n = 2, \\ \frac{1}{2}(k+1)(k+2)(k+4) & \text{for } n = 3. \end{cases}$$

In order to achieve a $H^{\mathrm{div}}(\Omega)$-conforming discretization, we take

$$N_{i,j}(\tau) = \int_{F_i} (\tau \cdot \nu_i) q_{ij} \, ds,$$

where the q_{ij} are a basis of $P_k(F_i)$, $i = 1, \dots, n+1$, and if $k \geq 1$,

$$N_{0,j}(\tau) = \int_K \tau \cdot q_j \, dx,$$

where the q_j are a basis of $P_{k-1}(K)^n$. To show that $(K, RT_k(K), \{N_{ij}\}_{i,j})$ defines a finite element – called the *Raviart–Thomas element* – we need to determine whether these conditions form a basis of $RT_k(K)^*$, which we can do via Lemma 4.3.

Lemma 10.8. *If $\tau_h \in RT_k(K)$ satisfies $N_{i,j}(\tau_h) = 0$ for all i, j, then $\tau_h = 0$.*

Proof. First, observe that $N_{i,j}(\tau_h) = 0$ for some i and all j implies that

$$\int_{F_i} (\tau_h \cdot \nu_i) q_k \, ds = 0 \quad \text{for all } q_k \in P_k(F_i),$$

and since $\tau_h|_{F_i} \cdot \nu_i \in P_k(F_i)$ by Lemma 10.7 (ii), $\tau_h|_F \cdot \nu_F = 0$ on each face F of K. Similarly, we have that

$$\int_K \tau_h \cdot \tilde{q}_k \, dx = 0 \quad \text{for all } \tilde{q}_k \in P_{k-1}(K)^n, \tag{10.18}$$

and hence for all $q_k \in P_k(K)$ that

$$\int_K \operatorname{div} \tau_h q_k \, dx = -\int_K \tau_h \nabla q_k \, dx + \int_{\partial K} \tau_h \cdot \nu q_k \, ds = 0$$

since $\nabla q_k = 0$ for $k = 0$ and $\nabla q_k \in P_{k-1}(K)^n$ for $k \geq 1$. As $\operatorname{div} \tau_h \in P_k(K)$ by Lemma 10.7 (i), this yields $\operatorname{div} \tau_h = 0$ on K.

By construction, $\tau_h = p_1 + xp_2$ for some $p_1 \in P_k(K)^n$ and $p_2 \in P_k^0(K)$. First, it is straightforward to verify that a homogeneous polynomial $p \in P_k^0(K)$ satisfies $x \cdot \nabla p = kp$ (this is known as *Euler's theorem for homogeneous functions*). Hence by the product rule,

$$0 = \operatorname{div}(\tau_h) = \operatorname{div} p_1 + (n+k)p_2.$$

Since $\operatorname{div} p_1$ for $p_1 \in P_k(K)^n$ is a polynomial of degree at most $k-1$ and p_2 is a homogeneous polynomial of degree k, this identity can only hold on K if $p_2 = 0$. Hence, $\operatorname{div} p_1 = 0$ as well.

For the remainder of the proof, we assume, without loss of generality, that K is the reference unit simplex spanned by the unit vectors in $\mathbb{R}^n$. Consider now for $1 \leq i \leq n$ the face F_i aligned with the coordinate plane $\{x \in \mathbb{R}^n : x_i = 0\}$, which has unit normal $\nu_i = -e_i$. Then

$$0 = \tau_h \cdot \nu_i = p_1 \cdot (-e_i) = -[p_1]_i,$$

and hence $[p_1]_i$ is a polynomial of degree k that vanishes for all x with $x_i = 0$. By Lemma 4.4, there thus exists a $\psi_i \in P_{k-1}(K)$ such that $[p_1]_i = x_i \psi_i$ for all $1 \leq i \leq n$. We thus obtain a $\tilde{q}_k = (\psi_1, \dots, \psi_n)^T \in P_{k-1}(K)^n$, which we can insert into (10.18) to deduce

$$\sum_{i=1}^n \int_K x_i |\psi_i|^2 \, dx = 0.$$

Since we are on the unit simplex, all terms are non-negative and thus have to vanish separately. This implies that $\psi_i = 0$ for $i = 1, \dots, n$ and thus $\tau_h = p_1 = 0$. □

Our next task is to construct interpolants in V_h for functions in V. This is complicated by the fact that functions in $H^{\text{div}}(K)$ have normal traces on $H^{-1/2}(\partial K)$, which cannot be localized to single faces $F \subset \partial K$. We therefore proceed as follows. For $\tau \in H^1(K)^n$ – which does have well-defined normal traces in $L^2(F)$ by Theorem 2.5 – we define the local *Raviart–Thomas projection* $\Pi_K \tau \in RT_k(K)$ by

$$\begin{aligned} &\int_F (\Pi_K \tau \cdot \nu - \tau \cdot \nu) q_k \, ds = 0 && \text{for all } q_k \in P_k(F), F \subset \partial K, \\ &\int_K (\Pi_K \tau - \tau) \cdot q_k \, dx = 0 && \text{for all } q_k \in P_{k-1}(K)^n \text{ if } k \geq 1. \end{aligned}$$

From Lemma 10.8, we already know that the projection conditions imply the uniqueness (and hence existence) of $\Pi_K \tau$. The next lemma shows that these conditions are chosen precisely in order to use the Raviart–Thomas projector Π_K in the construction of a Fortin projector. (Since Π_K is not continuous on $H^{\text{div}}(\Omega)$, it cannot be used directly.)

Lemma 10.9. *For any $\tau \in H^1(K)^n$,*

$$\int_K \operatorname{div}(\Pi_K\tau)q_k\,dx = \int_K (\operatorname{div}\tau)q_k\,dx \qquad \text{for all } q_k \in P_k(K).$$

Proof. Using integration by parts and the definition of the Raviart–Thomas projector, we have for any $q_k \in P_k(K)$ that

$$\int_K \operatorname{div}(\Pi_K\tau - \tau)q_k\,dx = \int_{\partial K} (\Pi_K\tau\cdot\nu - \tau\cdot\nu)q_k\,ds - \int_K (\Pi_K\tau - \tau)\cdot\nabla q_k\,dx = 0,$$

since $\nabla q_k = 0$ for $k = 0$ and $\nabla q_k \in P_{k-1}(K)^n$ for $k \geq 1$. □

This also yields local projection error estimates.

Lemma 10.10. *For any $\tau \in H^1(K)^n$,*

$$\begin{aligned}\|\Pi_K\tau - \tau\|_{L^2(K)^n} &\leq Ch_K|\tau|_{H^1(K)^n},\\ \|\operatorname{div}(\Pi_K\tau - \tau)\|_{L^2(K)} &\leq C|\tau|_{H^1(K)^n}.\end{aligned}$$

In addition, if $\tau \in H^2(K)^n$,

$$\|\operatorname{div}(\Pi_K\tau - \tau)\|_{L^2(K)} \leq Ch_K|\tau|_{H^2(K)^n}.$$

Proof. Since the projection conditions define a basis of $RT_k(K)^*$, we can write

$$\Pi_K\tau = \sum_{i=0}^{n+1}\sum_{j=0}^{d(i)} N_{i,j}(\tau)\psi_{i,j},$$

where $\{\psi_{i,j}\}_{i,j}$ is the corresponding nodal basis of $RT_k(K)$. The trace theorem and Hölder's inequality imply that for every $q_k \in P_k(F)$, the mapping $\tau \mapsto \int_F \tau\cdot\nu q_k\,ds$ is continuous on $H^1(K)^n$. We argue similarly for the degrees of freedom on K. Furthermore, from Lemma 10.9 and the fact that $\operatorname{div}(\Pi_K\tau) \in P_k(K)$ by Lemma 10.7 (i), we obtain

$$\begin{aligned}\|\operatorname{div}(\Pi_K\tau)\|^2_{L^2(K)} &= \int_K \operatorname{div}(\Pi_K\tau)\operatorname{div}(\Pi_K\tau)\,dx = \int_K (\operatorname{div}\tau)\operatorname{div}(\Pi_K\tau)\,dx\\ &\leq \|\operatorname{div}\tau\|_{L^2(K)}\,\|\operatorname{div}(\Pi_K\tau)\|_{L^2(K)}\,.\end{aligned}$$

The projection errors thus define bounded linear functionals on $H^1(K)^n$. The estimates then follow from the Bramble–Hilbert lemma and suitable scaling arguments.[3] □

[3] Since the local coordinate x appears explicitly in the definition of $RT_k(K)$, Raviart–Thomas elements are not affine-equivalent. One thus has to use the *Piola transform*: If K is generated from $\hat{K}$ by the affine transformation $\hat{x} \mapsto A_K\hat{x} + b_K$ and $\hat{p} \in RT_k(\hat{K})$, then $p = \det(A_K)^{-1}A_K\hat{p} \in RT_k(K)$. Furthermore, the transformed elements are interpolation equivalent; see [Raviart & Thomas 1977] and [Nédélec 1980].

The global Raviart–Thomas projector $\Pi_{\mathcal{T}}$ for $\tau \in H^1(\Omega)^n$ is now defined via $(\Pi_{\mathcal{T}}\tau)_K = \Pi_K \tau|_K$ for all $K \in \mathcal{T}_h$. This projector is bounded in the $H^{\operatorname{div}}(\Omega)$ norm by Lemma 10.10. Similarly, we obtain from the definition of M_h and Lemma 10.9 that $(\operatorname{div}(\Pi_{\mathcal{T}}\tau), v_h) = (\operatorname{div}\tau, v_h)$ for all $v_h \in M_h$. It remains to argue that $\Pi_{\mathcal{T}}\tau \in V_h$. Since $\Pi_{\mathcal{T}}\tau$ is a piecewise polynomial, it suffices to show that the normal trace is continuous across elements. Let K_1 and K_2 be two elements sharing a face F. Then $\tau \in H^1(\Omega)^n$ has a well-defined normal trace $\tau \cdot \nu \in L^2(F)$ and thus by construction,

$$\int_F (\Pi_{K_1}\tau) \cdot \nu\, q_k\, ds = \int_F \tau \cdot \nu\, q_k\, ds = \int_F (\Pi_{K_2}\tau) \cdot \nu\, q_k\, ds \quad \text{for all } q_k \in P_k(F).$$

Since $(\Pi_K \tau) \cdot \nu \in P_k(F)$ by Lemma 10.7 (ii), we obtain as in the proof of Lemma 10.8 that $(\Pi_{K_1}\tau - \Pi_{K_2}\tau) \cdot \nu = 0$ on F.

We are now in a position to apply the abstract saddle point framework to the mixed finite element discretization of (10.17): Find $(\sigma_h, u_h) \in V_h \times M_h$ satisfying

$$\begin{cases} (\sigma_h, \tau_h) + (\operatorname{div}\tau_h, u_h) = 0 & \text{for all } \tau_h \in V_h, \\ \qquad\qquad (\operatorname{div}\sigma_h, v_h) = -(f, v_h) & \text{for all } v_h \in M_h. \end{cases} \tag{10.19}$$

Since $V_h \subset V$ and $M_h \subset M$, the bilinear forms $a : V_h \times V_h \to \mathbb{R}$ and $b : V_h \times M_h \to \mathbb{R}$ are continuous. Furthermore, for $\tau_h \in V_h$ we have $\operatorname{div}\tau_h \in M_h$ and hence the coercivity of a on $\ker B_h \subset V_h$ follows exactly as in the continuous case. For the discrete LBB condition, we proceed as in the proof of the Fortin criterion: For given $v_h \in M_h \subset L^2(\Omega)$, let $\tau_{v_h} \in H^1(\Omega)^n$ be the function given by Lemma 10.6. Then $\Pi_{\mathcal{T}}\tau_{v_h} \in V_h \subset V$ and thus

$$\sup_{\tau_h \in V_h} \frac{b(\tau_h, v_h)}{\|\tau_h\|_V} \geq \frac{(\operatorname{div}(\Pi_{\mathcal{T}}\tau_{v_h}), v_h)}{\left\|\Pi_{\mathcal{T}}\tau_{v_h}\right\|_{H^{\operatorname{div}}(\Omega)}} \geq \frac{(\operatorname{div}\tau_{v_h}, v_h)}{C\left\|\tau_{v_h}\right\|_{H^1(\Omega)^n}} \geq \frac{(v_h, v_h)}{C\,\|v_h\|_{L^2(\Omega)}} = \frac{1}{C}\,\|v_h\|_M$$

by the properties of the Raviart–Thomas projector and Lemma 10.6. The conditions of the discrete Brezzi splitting theorem (Theorem 10.2) are thus satisfied, and we deduce well-posedness of (10.19).

Theorem 10.11. *For given $f \in L^2(\Omega)$, there exists a unique solution $(\sigma_h, u_h) \in V_h \times M_h$ to (10.19) satisfying*

$$\|\sigma_h\|_{H^{\operatorname{div}}(\Omega)} + \|u_h\|_{L^2(\Omega)} \leq C\,\|f\|_{L^2(\Omega)}.$$

Using Theorem 10.4 to bound the discretization error by the projection error and applying Lemma 10.10 yields a priori error estimates.

Theorem 10.12. *Assume the exact solution $(\sigma, u) \in H^{\operatorname{div}}(\Omega) \times L^2(\Omega)$ to (10.17) satisfies $u \in H^3(\Omega)$. Then the solution $(\sigma_h, u_h) \in V_h \times M_h$ satisfies*

$$\|\sigma - \sigma_h\|_{H^{\operatorname{div}}(\Omega)} + \|u - u_h\|_{L^2(\Omega)} \leq Ch\,\|u\|_{H^3(\Omega)}.$$

Part IV

TIME-DEPENDENT PROBLEMS

11 VARIATIONAL THEORY OF PARABOLIC PDES

In this chapter, we study time-dependent partial differential 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 at equilibrium, the *heat equation*

$$\begin{cases} \partial_t u(t,x) - \Delta u(t,x) = f(t,x), \\ \qquad\qquad\qquad u(0,x) = u_0(x), \end{cases}$$

describes the evolution in time of u starting from the given initial temperature distribution u_0 (called *initial condition* in this context). 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 consisting of functions depending on x only:

$$u : (0,T) \to V, \qquad t \mapsto u(t) \in V.$$

Similarly to the real-valued case, we define the following function spaces:

- *Hölder spaces*: For $k \geq 0$, 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
$$\|u\|_{C^k(0,T;V)} := \sum_{j=0}^{k} \sup_{t\in[0,T]} \left\| d_t^j u(t) \right\|_V$$

- *Lebesgue spaces* (also called *Bochner spaces*):[1] For $1 \leq p \leq \infty$, 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 a function in $L^p(0,T)$. This 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 via the integration-by-parts formula (2.1) with *scalar* test functions) in $L^p(0,T;V)$, we say that $u \in W^{1,p}(0,T;V)$. This 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 $1 < p, q < \infty$ and two reflexive Banach spaces V_0, V_1 with continuous embedding $V_0 \hookrightarrow V_1$, we set

$$W^{1,p,q}(0,T;V_0,V_1) := \left\{v \in L^p(0,T;V_0) : d_t v \in L^q(0,T;V_1)\right\}.$$

  This is a Banach space if equipped with the norm

$$\|u\|_{W(0,T;V_0,V_1)} := \|u\|_{L^p(0,T;V_0)} + \|d_t u\|_{L^q(0,T;V_1)}.$$

Of particular importance is the case $q = p/(p-1)$ (i.e., $1/p + 1/q = 1$) and $V_1 = V_0^*$, since in this case $L^p(0,T;V)^*$ can be identified with $L^q(0,T;V^*)$;[2] this is relevant because we later want to test $d_j u(t)$ with $v \in L^p(0,T;V)$. We can then transfer (via mollifiers)[3] the usual calculus rules to $W^p(0,T;V) := W^{1,p,q}(0,T;V,V^*)$.

Similarly to the Rellich–Kondrachov theorem, we can now ask whether we can use the integrability of $d_t u$ to obtain more regularity for u itself and, in particular, to deduce that u is continuous in time. This requires an additional assumption linking V and V^*. 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

$$V \hookrightarrow H \cong H^* \hookrightarrow V^*$$

with dense embeddings. We call (V,H,V^*) *Gelfand* or *evolution triple.*

Theorem 11.1. *Let $1 < p < \infty$ and (V,H,V^*) be a Gelfand triple. Then the embedding*

$$W^p(0,T;V) \hookrightarrow C(0,T;H)$$

is continuous.

[1] For a rigorous definition, see [Wloka 1987, § 24]

[2] see, e.g., [Edwards 1965, Theorem 8.20.3]

[3] 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)]

This result guarantees that functions in $W^p(0,T;V)$ have well-defined traces $u(0), u(T) \in H$, which is important to make sense of the initial condition $u(0) = u_0$.

We also need the following integration by parts formula (where now the test function is Banach-space valued).

Lemma 11.2. *Let (V,H,V^*) be a Gelfand triple. Then for every $u,v \in W^p(0,T;V)$,*

$$\frac{d}{dt}\langle u(t),v(t)\rangle_H = \langle d_t u(t),v(t)\rangle_{V^*,V} + \langle d_t v(t),u(t)\rangle_{V^*,V} \quad \text{for a.e. } t \in (0,T),$$

and hence

$$\int_0^T \langle d_t u(t),v(t)\rangle_{V^*,V}\,dt = \langle u(T),v(T)\rangle_H - \langle u(0),v(0)\rangle_H - \int_0^T \langle d_t v(t),u(t)\rangle_{V^*,V}\,dt.$$

In the following, we focus for simplicity only the case $p = q = 2$, for which we set $W(0,T;V) := W^2(0,T;V)$.

11.2 WEAK SOLUTION OF PARABOLIC PDES

We can now formulate our parabolic evolution problem. Let for almost every $t \in (0,T)$ a bilinear form $a(t;\cdot,\cdot) : V \times V \to \mathbb{R}$ and a linear form $f \in L^2(0,T;V^*)$ as well as a $u_0 \in H$ be given. The problem in strong form (in time) is then to find $u \in W(0,T;V)$ such that

$$\tag{11.1} \begin{cases} \langle d_t u(t),v\rangle_{V^*,V} + a(t;u(t),v) = \langle f(t),v\rangle_{V^*,V} \text{ for all } v \in V \text{ and a.e. } t \in (0,T),\\ \qquad\qquad u(0) = u_0. \end{cases}$$

(For, e.g., the heat equation, we have $V = H_0^1(\Omega) \hookrightarrow L^2(\Omega) = H$ and $a(t;u,v) = (\nabla u, \nabla v)$.) Just as in the stationary case, we now formulate this in fully variational or weak form. For simplicity, assume $u_0 = 0$ (the inhomogeneous case can be treated in the same fashion as inhomogeneous Dirichlet conditions) and consider the Banach spaces

$$X = \{w \in W(0,T;V) : w(0) = 0\}, \qquad Y = L^2(0,T;V),$$

such that $Y^* = L^2(0,T;V^*)$. Setting

$$b : X \times Y \to \mathbb{R}, \qquad b(u,y) = \int_0^T \langle d_t u(t), y(t)\rangle_{V^*,V} + a(t;u(t),y(t))\,dt$$

and

$$\langle f,y\rangle_{Y^*,Y} = \int_0^T \langle f(t),y(t)\rangle_{V^*,V}\,dt,$$

we look for $u \in X$ such that

$$b(u, y) = \langle f, y\rangle_{Y^*,Y} \qquad \text{for all } y \in Y. \tag{11.2}$$

The equivalence to (11.1) follows from considering $y(t) = \varphi(t)v$ for arbitrary $\varphi \in C_0^\infty(0, T)$ and $v \in V$ and using the fundamental theorem of the calculus of variations.[4]

Well-posedness of (11.1) can then be shown using the Banach–Nečas–Babuška theorem.

Theorem 11.3. *Assume that the bilinear form $a(t; \cdot, \cdot) : V \times V \to \mathbb{R}$ satisfies the following properties:*

(i) The mapping $t \mapsto a(t; u, v)$ is measurable for all $u, v \in V$.

(ii) There exists $M > 0$ such that $|a(t; u, v)| \leq M \|u\|_V \|v\|_V$ for almost every $t \in (0, T)$ and all $u, v \in V$.

(iii) There exists $\alpha > 0$ such that $a(t; u, u) \geq \alpha \|u\|_V^2$ for almost every $t \in (0, T)$ and all $u \in V$.

Then (11.2) has a unique solution $u \in W(0, T; V)$ satisfying

$$\|u\|_{W(0,T;V)} \leq C \|f\|_{Y^*}.$$

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

$$A(t) : V \to V^*, \qquad \langle A(t)u, v\rangle_{V^*,V} := a(t; u, v) \quad \text{for all } u, v \in V.$$

Continuity of a implies that for almost every $t \in (0, T)$, the operator $A(t)$ is bounded with constant M. Similarly, coercivity of a and the Lax–Milgram theorem shows that $A(t)$ is an isomorphism, hence $A(t)^{-1} : V^* \to V$ is bounded as well with constant α^{-1}. Therefore, for almost every $t \in (0, T)$ and all $v^* \in V^*$

$$\begin{aligned} \left\langle v^*, A(t)^{-1}v^*\right\rangle_{V^*,V} &= \left\langle A(t)A(t)^{-1}v^*, A(t)^{-1}v^*\right\rangle_{V^*,V} \geq \alpha \left\|A(t)^{-1}v^*\right\|_V^2 \\ &\geq \frac{\alpha}{M^2} \|v^*\|_{V^*}^2. \end{aligned} \tag{11.3}$$

For arbitrary $u \in X$ and $\mu > 0$, set $z = A(t)^{-1}d_t u + \mu u$. By the triangle inequality, the uniform continuity of $A(t)^{-1}$, and the definition of the norms in X and Y, we have that

$$\|z\|_Y^2 \leq 2\alpha^{-2} \int_0^T \|d_t u(t)\|_{V^*}^2 \, dt + 2\mu^2 \int_0^T \|u(t)\|_V^2 \, dt \leq c \|u\|_X^2,$$

[4] see, e.g., [Ern & Guermond 2021, Lemma 65.4].

and thus in particular that $z \in Y$. Moreover, using (11.3), integration by parts, and continuity of $A(t)$ and $A(t)^{-1}$, respectively, we can estimate term by term in

$$\begin{aligned} b(u,z) &= \int_0^T \left\langle d_t u(t) + A(t)u(t), A(t)^{-1} d_t u(t) + \mu u(t) \right\rangle_{V^*,V} dt \\ &\geq \frac{\alpha}{M^2} \int_0^T \|d_t u(t)\|_{V^*}^2 \, dt + \frac{\mu}{2} \|u(T)\|_H^2 - \frac{M}{\alpha} \int_0^T \|u(t)\|_V \|d_t u(t)\|_{V^*} \, dt \\ &\quad + \mu\alpha \int_0^T \|u(t)\|_V^2 \, dt \\ &\geq \frac{\alpha}{2M^2} \int_0^T \|d_t u(t)\|_{V^*}^2 \, dt + \left(\mu\alpha - \tfrac{M^4}{2\alpha^3}\right) \int_0^T \|u(t)\|_V^2 \, dt, \end{aligned}$$

using the generalized Young's inequality with $\varepsilon = \alpha/M^2$.

Taking $\mu = M^4\alpha^{-4}$, the term in parenthesis is positive, which yields (for generic constants $c > 0$)

$$b(u,z) \geq c \|u\|_X^2 \geq c \|u\|_X \|z\|_Y .$$

This implies the inf–sup condition via

$$\inf_{u\in X} \sup_{y\in Y} \frac{b(u,y)}{\|u\|_X \|y\|_Y} \geq \inf_{u\in X} \frac{b(u,z)}{\|u\|_X \|z\|_Y} \geq c.$$

It remains to show that the injectivity condition holds. Assume $y \in Y$ is such that $b(u,y) = 0$ for all $u \in X$. For any $\varphi \in C_0^\infty(0,T)$ and $v \in V$, we have $\varphi v \in X$. Due to the definition of the weak time derivative and $b(\varphi v, y) = 0$, we thus obtain that

$$\begin{aligned} \int_0^T \langle d_t y(t), v\rangle_{V^*,V} \varphi(t) \, dt &= -\int_0^T \langle d_t \varphi(t) v, y(t)\rangle_{V^*,V} \, dt = \int_0^T a(t; \varphi(t)v, y(t)) \, dt \\ &= \int_0^T \langle A(t)^* y(t), v\rangle_{V^*,V} \varphi(t) \, dt, \end{aligned}$$

and hence (by density of $C_0^\infty(0,T)$ in $L^2(0,T)$) that $d_t y(t) = A(t)^* y(t)$ for almost all $t \in (0,T)$. In particular, we deduce that $d_t y \in L^2(0,T;V^*)$ and therefore $y \in W(0,T;V)$.

Since $d_t y = A^* y$ in Y^* and $tv \in X \hookrightarrow Y$ for any $v \in V$, we obtain using Lemma 11.2 that

$$\begin{aligned} 0 &= \int_0^T \langle -d_t y(t), tv\rangle_{V^*,V} + \langle A(t)^* y(t), tv\rangle_{V^*,V} \, dt \\ &= -\langle y(T), Tv\rangle_H + \int_0^T \langle d_t(tv), y(t)\rangle_{V^*,V} + a(t; tv, y(t)) \, dt \\ &= -T \langle y(T), v\rangle_H . \end{aligned}$$

By density of V in H, this implies that $y(T) = 0$. Similary, $y \in W(0,T;V)$ and the first part of Lemma 11.2 yields

$$\begin{aligned} 0 &= \int_0^T -\langle d_t y(t), y(t)\rangle_{V^*,V} + \langle A(t)^* y(t), y\rangle_{V^*,V}\, dt \\ &\geq \int_0^T -\frac{d}{dt}\left(\frac{1}{2}\|y(t)\|_H^2\right) + \alpha\,\|y(t)\|_V^2\, dt \\ &= \frac{1}{2}\|y(0)\|_H^2 + \alpha\,\|y\|_Y^2 \end{aligned}$$

and hence $y = 0$. We can thus apply the Banach–Nečas–Babuška theorem, and the claim follows. □

12 GALERKIN APPROACH FOR PARABOLIC PROBLEMS

To obtain a finite-dimensional approximation of (11.1), we need to discretize in both space and time: either separately (combining finite elements in space with a time stepping method for ordinary differential equations) or all-at-once (using a Galerkin approach with suitable discrete test spaces). Only a brief overview over the different approaches is given here.

12.1 TIME STEPPING METHODS

These approaches can be further discriminated based on the order of operations:

Method of lines This method starts with a discretization in space to obtain a (very large) 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 $\{\varphi_j\}_{j=1}^{N_h}$ of V_h, we approximate the unknown solution as $u_h(t,x) = \sum_{j=1}^{N_h} U_j(t)\varphi_j(x)$. 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 $U(t) = (U_1(t), \dots U_{N_h}(t))^T$:

$$\begin{cases} \mathbf{M}\dfrac{d}{dt}U(t) + \mathbf{K}(t)U(t) = \mathbf{M}F(t), \\ \hfill U(0) = U_0, \end{cases}$$

where U_0 and $F(t)$ are the coefficients vectors 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 & 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

  $$d_t u(t+\tau) \approx \frac{u(t+\tau)-u(t)}{\tau}$$

  for $\tau > 0$ at time $t+\tau$ to obtain for given $u(t)$ and unknown $u(t+\tau) \in V$ the *stationary* partial differential equation

  $$\langle u(t+\tau), v\rangle_H + \tau\, a(t+\tau; u(t+\tau), v) = \langle u(t), v\rangle_H + \tau\, \langle f(t+\tau), v\rangle_{V^*,V}$$

  for all $v \in V$.

- The *Crank–Nicolson scheme* uses the central difference quotient

  $$d_t u(t+\tfrac{\tau}{2}) \approx \frac{u(t+\tau)-u(t)}{\tau}$$

  for $\tau > 0$ at time $t+\frac{\tau}{2}$ to obtain

  $$\begin{aligned}\langle u(t+\tau), v\rangle_H + \tfrac{\tau}{2}\, a(t+\tfrac{\tau}{2}; u(t+\tau), v) = \langle u(t), v\rangle_H - \tfrac{\tau}{2}\, a(t+\tfrac{\tau}{2}; u(t), v)\\ + \tau\left\langle f(t+\tfrac{\tau}{2}), v\right\rangle_{V^*,V}\end{aligned}$$

  for all $v \in V$.

Starting with $t = 0$, these are then approximated and solved in turn for $u(t_m)$, $t_m := m\tau$, using a finite element discretization in space. This approach is discussed in detail in [Thomée 2006, Chapters 7–9]. The advantage of Rothe's method is that at each time step, a different spatial discretization can be used.

12.2 GALERKIN METHODS

Proceeding as in the stationary case, we can apply a Galerkin approximation to (11.2) by replacing X and Y with finite-dimensional spaces X_h and Y_h. Again, we can further discriminate between conforming and non-conforming approaches.

Conforming Galerkin methods

In a conforming approach, we choose $X_h \subset X$ and $Y_h \subset Y$ and seek $u_h \in X_h$ 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 $y_h \in Y_h$. We now 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, $1 \leq m \leq N$, 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 for $r \geq 1$

$$\begin{aligned} X_h &= \left\{ w_h \in C(0,T;V) : w_h|_{[t_{m-1},t_m]} \in P_r(t_{m-1}, t_m; V_m),\ 1 \leq m \leq N,\ w_h(0) = u_0 \right\}, \\ Y_h &= \left\{ y_h \in L^2(0,T;V) : y_h|_{(t_{m-1},t_m]} \in P_{r-1}(t_{m-1}, t_m; V_m),\ 1 \leq m \leq N \right\}. \end{aligned}$$

Since this is a conforming approximation, we can deduce well-posedness of the corresponding discrete problem in the usual fashion (noting that $d_t u_h \in Y_h$ for $u_h \in X_h$). (Since functions in X – and hence in X_h – are continuous in time by Theorem 11.1, this approach is often called *continuous Galerkin* or $cG(r)$ method.)

This approach is closely related to Rothe's method. Consider the case $r = 1$ (i.e., piecewise linear in time) and, for simplicity, a time-independent bilinear form. We also assume that we choose the same space discretization at each time step, i.e., that $V_1 = \cdots = V_N = V_h$. Since functions in X_h are continuous at $t = t_m$ for all $0 \leq m \leq N$ and linear on each intervall $[t_{m-1}, t_m]$, we can write

$$u_h(t) = \frac{t_m - t}{t_m - t_{m-1}} u_h(t_{m-1}) + \frac{t - t_{m-1}}{t_m - t_{m-1}} u_h(t_m), \qquad t \in [t_{m-1}, t_m],$$

with coefficients $u_h(t_{m-1}), u_h(t_m) \in V_h$. (For $t_0 = 0$, we fix $u_h(t_0) = u_0$.) Similarly, functions in Y_h are constant in time and thus

$$y_h(t) \equiv y_h(t_m) =: v_h \in V_h, \qquad t \in (t_{m-1}, t_m].$$

Inserting this into (12.1) and setting $\tau_m := t_m - t_{m-1}$ yields for all $v_h \in V_h$ that

$$\langle u_h(t_m) - u_h(t_{m-1}), v_h \rangle_{V^*,V} + \frac{\tau_m}{2} a(u_h(t_{m-1}) + u_h(t_m), v_h) = \int_{t_{m-1}}^{t_m} \langle f(t), v_h \rangle_{V^*,V}\, dt,$$

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 $y_h \in Y_h$).[1] For this method, one can show error estimates of the form[2]

$$\|u_h(t_m) - u(t_m)\|_{L^2(\Omega)} \leq C(h^s \|u_0\|_{H^s(\Omega)} + \tau^2 \|u_0\|_{H^4(\Omega)}),$$

for $f = 0$ and $u_0 \neq 0$, where s depends on the accuracy of the spatial discretization, and $\tau = \max_{1 \leq m \leq N} \tau_m$.

[1] If the discrete spaces are different for each time interval, we need to use the H-projection of u_{m-1} on V_m.

[2] [Thomée 2006, Theorem 7.8]

Discontinuous Galerkin methods Instead of enforcing continuity of the discrete solution u_h through the definition of X_h, we can also use $X_h = Y_h$ and modify the bilinear form. Let $J_m := (t_{m-1}, t_m]$ denote the half-open interval between two time steps of length $\tau_m = t_m - t_{m-1}$. Then we set for $r \geq 0$

$$X_h = Y_h = \left\{ y_h \in L^2(0,T;V) : y_h|_{J_m} \in P_r(t_{m-1}, t_m; V_m),\ 1 \leq m \leq N \right\} \subset Y,$$

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 for $u_h \in X_h$

$$u_m := u_h(t_m) = \lim_{\varepsilon \to 0} u_h(t_m - \varepsilon), \qquad u_m^+ := \lim_{\varepsilon \to 0} u_h(t_m + \varepsilon)$$

and

$$[\![u_h]\!]_m = u_m^+ - u_m.$$

Similarly to the stationary case, we now define the discrete bilinear form

$$b_h(u_h, y_h) = \sum_{m=1}^{N} \int_{J_m} \langle d_t u_h(t), y_h(t) \rangle_V + a(t; u_h(t), y_h(t))\, dt + \sum_{m=1}^{N} \left\langle [\![u_h]\!]_{m-1}, y_{m-1}^+ \right\rangle_H$$

(which can be derived by integration by parts on each interval J_m and rearranging the jump terms). As $0 \notin J_1$, we will need to specify $u_h(0) = u_0$ separately, which we do by setting $[\![u_h]\!]_0 := u_0^+ - u_0$. Note that this makes b_h *affine* instead of bilinear unless $u_0 = 0$. (In other words, we should actually split the jump term for $m = 0$ into the part involving u_0^+, which remains part of b_h, and the part involving u_0, which should be part of the right-hand side. However, we stick with the above formulation for the sake of presentation.)

We then search for $u_h \in X_h$ satisfying

$$b_h(u_h, y_h) = \langle f, y_h \rangle_{Y^*,Y} \qquad \text{for all } y_h \in X_h. \tag{12.2}$$

Since the exact solution $u \in X$ is continuous and satisfies $u(0) = u_0$, we have

$$b_h(u, y_h) = b(u, y_h) = \langle f, y_h \rangle_{Y^*,Y} \qquad \text{for all } y_h \in X_h,$$

and hence this is a consistent approximation. To prove well-posedness of the discrete problem, we again define a discrete "jump-norm"

$$|\!|\!|u_h|\!|\!|^2 = \sum_{m=1}^{N} \int_{J_m} \|d_t u_h(t)\|_H^2 + \|u_h(t)\|_V^2 \, dt + \sum_{m=1}^{N} \left\| [\![u_h]\!]_m \right\|_H^2.$$

We can then proceed as in the proof of Theorem 8.2.

Theorem 12.1. *Under the assumptions of Theorem 11.3, there exists a unique solution $u_h \in X_h$ to (12.2), and*

$$|\!|\!|u_h|\!|\!| \leq C \left(\|f\|_{Y^*}^2 + \|u_0\|_H^2 \right).$$

Proof. Continuity of b_h with respect to $|||\cdot|||$ follows from the definition. It remains to show injectivity of $B_h : X_h \to Y_h^*$, $u_h \mapsto b_h(u_h, \cdot)$, (which suffices for bijectivity since $X_h = Y_h$ are finite-dimensional). Instead of verifying the inf–sup condition, we do this directly. Let $u_h \in X_h$ satisfy $b_h(u_h, y_h) = 0$ for all $y_h \in X_h$ with $u_0 = 0$. Since functions in Y_h can be discontinuous at the time points t_m, we can insert $y_h = \mathbb{1}_{J_m} u_h \in Y_h$ for each $1 \leq m \leq N$, where $\mathbb{1}_{J_m}(t) = 1$ if $t \in J_m$ and zero else. We start with $J_1 = (t_0, t_1]$. Since $\mathbb{1}_{J_1}$ is constant on J_1 and zero outside J_1, we have using $u_0 = 0$ that

$$\begin{aligned}
0 &= b_h(u_h, \mathbb{1}_{J_1} u_h) \\
&= \int_{J_1} \langle d_t u_h(t), u_h(t)\rangle_{V^*,V} + a(t; u_h(t), u_h(t))\, dt + \left\langle u_0^+ - u_0, u_0^+ \right\rangle_H \\
&\geq \frac{1}{2}\|u_1\|_H^2 - \frac{1}{2}\left\|u_0^+\right\|_H^2 + \alpha \int_{J_1} \|u_h(t)\|_V^2\, dt + \left\|u_0^+\right\|_H^2 \\
&\geq \frac{1}{2}\|u_1\|_H^2 + \alpha \int_{J_1} \|u_h(t)\|_V^2\, dt.
\end{aligned}$$

Hence, $u_h|_{J_1} = 0$ and $u_1 = 0$, and we can proceed in a similar way for $J_2, J_3, \dots, J_N$ to deduce that $u_h = 0$. The estimate then follows from bijectivity using the closed range theorem. □

Before we address a priori error estimates, we discuss how to formulate discontinuous Galerkin methods as time stepping methods. For simplicity, we again assume that the bilinear form a is time-independent and that $V_1 = \dots = V_N = V_h$. First consider the case $r = 0$, i.e., piecewise constant functions in time. Then $d_t(u_h|_{J_m}) = 0$ and $u_h|_{J_m} \equiv u_m = u_{m-1}^+ \in V_h$. Using as test functions $y_h = \mathbb{1}_{J_m} v_h \in Y_h$ for arbitrary $v_h \in V_h$ and $m = 1, \dots, N$, we obtain

$$\langle u_m, v_h\rangle_H + \tau_m\, a(u_m, v_h) = \langle u_{m-1}, v_h\rangle_H + \int_{J_m} \langle f(t), v_h\rangle_{V^*,V}\, dt$$

for all $v_h \in V_h$, which is a variant of the implicit Euler scheme. For $r = 1$ (piecewise linear functions), we make the ansatz

$$u_h|_{J_m}(t) = u_m^0 + \frac{t - t_{m-1}}{\tau_m} u_m^1 \in X_h$$

for coefficients $u_m^0, u_m^1 \in V_h$ (such that $u_{m-1}^+ = u_m^0$ and $u_m = u_m^0 + u_m^1$). Again, we choose for each J_m test functions which are zero outside J_m; specifically, we take $\mathbb{1}_{J_m}(t) v_h$ and $\mathbb{1}_{J_m}(t)\frac{t-t_{m-1}}{\tau_m} w_h$ for arbitrary $v_h, w_h \in V_h$. Inserting these in turn into the bilinear form and computing the integrals yields the coupled system

$$\begin{aligned}
&\left\langle u_m^0, v_h\right\rangle_H + \tau_m\, a(u_m^0, v_h) + \left\langle u_m^1, v_h\right\rangle_H + \frac{\tau_m}{2}\, a(u_m^1, v_h) \\
&\qquad\qquad = \langle u_{m-1}, v_h\rangle_H + \int_{J_m} \langle f(t), v_h\rangle_{V^*,V}\, dt \quad \text{for all } v_h \in V_h, \\
&\frac{\tau_m}{2}\, a(u_m^0, w_h) + \frac{1}{2}\left\langle u_m^1, w_h\right\rangle_H + \frac{\tau_m}{3}\, a(u_m^1, w_h) \\
&\qquad\qquad = \frac{1}{\tau_m}\int_{J_m} (t - t_{m-1})\, \langle f(t), w_h\rangle_{V^*,V}\, dt \quad \text{for all } w_h \in V_h.
\end{aligned}$$

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. Similarly, discontinuous Galerkin methods for $r \geq 2$ lead to $(r+1)$-stage implicit Runge–Kutta time-stepping schemes.

12.3 A PRIORI ERROR ESTIMATES FOR DISCONTINUOUS GALERKIN METHODS

To derive a priori error estimates for discontinuous Galerkin approximations, we will first show a discrete stability result. 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., $\langle Au, v\rangle_{V^*,V} = a(u,v)$ for all $u, v \in V$. We also assume for the sake of presentation that $f \in L^2(0,T;H)$ and that the discrete solution u_h is sufficiently regular that $Au_h(t) \in H$.

Theorem 12.2. *For given $f \in L^2(0,T;H)$ and $u_0 \in H$, the solution u_h of* (12.2) *satisfies*

$$\sum_{m=1}^{N} \left(\int_{J_m} \|d_t u_h(t)\|_H^2 + \|Au_h(t)\|_H^2 \, dt + \tau_m^{-1} \left\| [\![u_h]\!]_{m-1} \right\|_H^2 \right) \leq C \left(\|f\|_{L^2(0,T;H)}^2 + \|u_0\|_H^2 \right).$$

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 $\|Au_h(t)\|_H$, we set $y_h = \mathbb{1}_{J_m} Au_h$ for $1 \leq m \leq N$ to obtain

$$\int_{J_m} \langle d_t u_h(t), Au_h(t)\rangle_H + \|Au_h(t)\|_H^2 \, dt + \left\langle [\![u_h]\!]_{m-1}, (Au_h)_{m-1}^+ \right\rangle_H = \int_{J_m} \langle f(t), Au_h(t)\rangle_H \, dt.$$

Due to the bilinearity and symmetry of a, we have

$$\begin{aligned}\int_{J_m} \langle d_t u_h(t), Au_h(t)\rangle_H \, dt = \int_{J_m} a(u_h(t), d_t u_h(t)) \, dt &= \int_{J_m} \frac{d}{dt}\left(\frac{1}{2} a(u_h(t), u_h(t))\right) dt \\ &= \frac{1}{2} a(u_m, u_m) - \frac{1}{2} a(u_{m-1}^+, u_{m-1}^+).\end{aligned}$$

Similarly, since A is time-independent,

$$\begin{aligned}\left\langle [\![u_h]\!]_{m-1}, (Au_h)_{m-1}^+ \right\rangle_H &= a([\![u_h]\!]_{m-1}, u_{m-1}^+) \\ &= \frac{1}{2} a([\![u_h]\!]_{m-1}, u_{m-1}^+ + u_{m-1} + [\![u_h]\!]_{m-1}) \\ &= \frac{1}{2} a(u_{m-1}^+, u_{m-1}^+) - \frac{1}{2} a(u_{m-1}, u_{m-1}) + \frac{1}{2} a([\![u_h]\!]_{m-1}, [\![u_h]\!]_{m-1}).\end{aligned}$$

Inserting these into the bilinear form $b_h(u_h, y_h)$ yields

$$a(\llbracket u_h \rrbracket_{m-1}, \llbracket u_h \rrbracket_{m-1}) + a(u_m, u_m) - a(u_{m-1}, u_{m-1}) + 2\int_{J_m} \|Au_h(t)\|_H^2 \,dt \\ = 2\int_{J_m} \langle f, Au_h(t)\rangle_H \,dt.$$

Summing over all $1 \leq m \leq N$ yields

$$\sum_{m=1}^N a(\llbracket u_h \rrbracket_{m-1}, \llbracket u_h \rrbracket_{m-1}) + \sum_{m=1}^N \int_{J_m} 2\,\|Au_h(t)\|_H^2 \,dt \\ \leq \sum_{m=1}^N \int_{J_m} 2\,\langle f(t), Au_h(t)\rangle_H \,dt + a(u_0, u_0).$$

For $2 \leq m \leq N$, we can simply use coercivity of a to eliminate the jump terms and apply Young's inequality to $\langle f(t), Au_h(t)\rangle_H$ 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

$$a(u_0, u_0^+) = \left\langle u_0, Au_0^+\right\rangle_H \leq \frac{\varepsilon}{2}\left\|Au_0^+\right\|_H^2 + \frac{1}{2\varepsilon}\|u_0\|_H^2.$$

Since $t \mapsto \|Au_h(t)\|_H^2$ is a polynomial in t of degree up to $2r$ on J_1, we have the estimate

$$\tau_1\left\|Au_0^+\right\|_H^2 \leq C\int_{t_0}^{t_1} \|Au_h(t)\|_H^2 \,dt.$$

Choosing $\varepsilon > 0$ small enough such that $\varepsilon C\tau_1^{-1} < 1$ yields

$$\sum_{m=1}^N \int_{J_m} \|Au_h(t)\|_H^2 \,dt \leq C\left(\int_0^T \|f(t)\|_H^2 \,dt + \|u_0\|_H^2\right). \tag{12.3}$$

Step 2. For the bound on $d_t u_h$, we use the inverse estimate

$$\int_{J_m} \|y_h(t)\|_H^2 \,dt \leq C\tau_m^{-1}\int_{J_m} (t - t_{m-1})\,\|y_h(t)\|_H^2 \,dt$$

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 $y_h = \mathbb{1}_{J_m}(t - t_{m-1})d_t u_h$

for $1 \le m \le N$. Since $y_{m-1}^+ = 0$, we have using the Cauchy–Schwarz inequality that

$$\begin{aligned}\int_{J_m} (t - t_{m-1}) \, \|d_t u_h(t)\|_H^2 \, dt &= \int_{J_m} (t - t_{m-1}) \, \langle f(t) - Au_h(t), d_t u_h(t)\rangle_H \, dt \\ &\le \left(\int_{J_m} (t - t_{m-1}) \, \|f(t) - Au_h(t)\|_H^2 \, dt\right)^{\frac{1}{2}} \\ &\qquad \cdot \left(\int_{J_m} (t - t_{m-1}) \, \|d_t u_h(t)\|_H^2 \, dt\right)^{\frac{1}{2}}.\end{aligned}$$

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

$$\sum_{m=1}^{N} \int_{J_m} \|d_t u_h(t)\|_H^2 \, dt \le C\left(\int_0^T \|f(t)\|_H^2 \, dt + \|u_0\|_H^2\right). \tag{12.4}$$

Step 3. It remains to estimate the jump terms. For this, we set $y_h = \mathbb{1}_{J_m} [\![u_h]\!]_{m-1}$ for $1 \le m \le N$. This yields

$$\begin{aligned}\left\|[\![u_h]\!]_{m-1}\right\|_H^2 &= \int_{J_m} \left\langle f(t) - Au_h(t), [\![u_h]\!]_{m-1}\right\rangle_H - \left\langle d_t u_h(t), [\![u_h]\!]_{m-1}\right\rangle_H \, dt \\ &\le \frac{\tau_m}{2} \int_{J_m} \|f(t) - Au_h(t) - d_t u_h(t)\|_H^2 \, dt + \frac{1}{2\tau_m} \int_{J_m} \left\|[\![u_h]\!]_{m-1}\right\|_H^2 \, dt,\end{aligned}$$

where we have used the generalized Young's inequality. Since $[\![u_h]\!]_{m-1}$ is constant in time, we have

$$\int_{J_m} \left\|[\![u_h]\!]_{m-1}\right\|_H^2 \, dt = \tau_m \left\|[\![u_h]\!]_{m-1}\right\|_H^2.$$

From (12.3) and (12.4), we thus obtain

$$\sum_{m=1}^{N} \tau_m^{-1} \left\|[\![u_h]\!]_{m-1}\right\|_H^2 \le C\left(\int_0^T \|f(t)\|_H^2 \, dt + \|u_0\|_H^2\right),$$

which completes the proof. □

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

$$X_r = \left\{y_r \in L^2(0,T;V) : y_r|_{J_m} \in P_r(t_{m-1}, t_m; V), \ 1 \le m \le N\right\}$$

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

$$\begin{cases} \pi_r u(t_m) = u(t_m), & \\ \int_{J_m} (u(t) - \pi_r u(t))\varphi(t)\,dt = 0 & \text{for all } \varphi \in P_{r-1}(J_m; V), \end{cases}$$

for all $1 \leq m \leq N$. (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 $t \in J_m$, $1 \leq m \leq N$:

$$\|u(t) - \pi_r u(t)\|_H \leq C\tau_m^{r+1} \int_{J_m} \left\|d_t^{r+1} u(\tau)\right\|_H \,d\tau. \tag{12.5}$$

Similarly, we assume that for each $t \in [0, T]$ the spatial interpolation error in V_h satisfies the estimate

$$\|u(t) - \mathcal{I}_h u(t)\|_H + h\,\|u(t) - \mathcal{I}_h u(t)\|_V \leq Ch^{s+1}\,\|u(t)\|_{H^{s+1}(\Omega)}\,.$$

(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 $s \geq 1$; see 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

$$b_h(y_h, z_h) = 0 \qquad \text{with} \quad z_N = \varphi.$$

Integrating by parts on each interval J_m and rearranging the jump terms, we can express the adjoint equation as

$$\begin{aligned} \sum_{m=1}^{N} \int_{J_m} -\langle y_h(t), d_t z_h(t)\rangle_H + a(y_h(t), z_h(t))\,dt \qquad& \\ + \sum_{m=1}^{N-1} \left\langle y_m, [\![z_h]\!]_m \right\rangle_H + \langle y_N, z_N\rangle_H = \langle y_N, \varphi\rangle_H\,.& \end{aligned} \tag{12.6}$$

This can be interpreted as a backwards in time equation with "initial value" $z_h(t_N) = \varphi$. Making the substitution $t \mapsto t_N - t$, we can apply Theorem 12.2 to obtain

$$\sum_{m=1}^{N} \int_{J_m} \|d_t z_h(t)\|_H^2 + \|A^* z_h(t)\|_H^2\,dt + \sum_{m=1}^{N} \left\|[\![z_h]\!]_{m-1}\right\|_H^2 \leq C\,\|\varphi\|_H^2\,, \tag{12.7}$$

where A is again the operator corresponding to the bilinear form a and we have used that $\tau_m < 1$ for N sufficiently large.

Now everything is in place to show the following a priori estimate for the discrete solution at each time step.[3]

[3] It is possible – though more involved – to show error estimates for arbitrary $t \in [0, T]$; see, e.g., [Thomée 2006, Theorem 12.2].

Theorem 12.3. *For $r = 0$, the solutions $u \in X$ to (11.2) and $u_h \in X_h$ to (12.2) satisfy*

$$\|u(t_m) - u_m\|_H \leq C \max_{1\leq n\leq m} \left(h^{s+1} \sup_{t\in J_n} \|u(t)\|_{H^{s+1}(\Omega)} + \tau_n \int_{J_n} \|d_t u\|_H \, dt \right)$$

for all $1 \leq m \leq N$.

Proof. We write the error $e(t)$ at almost every $t \in (0, T]$ as

$$\begin{aligned} e(t) := u(t) - u_h(t) &= (u(t) - \mathcal{I}_h \pi_r u(t)) + (\mathcal{I}_h \pi_r u(t) - u_h(t)) \\ &=: e_1(t) + e_2(t). \end{aligned}$$

(Note that pointwise a.e., $e(t) \in V$ since $V_h \subset V$, but as a function in time, only $e_2 \in X_h$ is in a meaningful function space.) For $t = t_m$, we have $\pi_r u(t_m) = u(t_m)$ by construction, and hence

$$\|e_1(t_m)\|_H = \|\mathcal{I}_h u(t_m) - u(t_m)\|_H \leq C h^{s+1} \|u(t_m)\|_{H^{s+1}(\Omega)} . \tag{12.8}$$

To bound $e_2(t_m)$, we use the duality trick. For arbitrary $\varphi \in H$, let z_h denote the solution of (12.6) with $N = m$. Since we have a consistent approximation, we can insert u and u_h in (12.2) for $y_h \in Y_h \subset Y$ and subtract to deduce that

$$0 = b_h(e, y_h) = b_h(e_1, y_h) + b_h(e_2, y_h) \qquad \text{for all } y_h \in X_h.$$

From this and $d_t(z_h|_{J_n}) = 0$, we obtain with $y_h = e_2 \in X_h$ that z_h satisfies

$$\begin{aligned} \langle e_2(t_m), \varphi\rangle_H &= b_h(e_2, z_h) = -b_h(e_1, z_h) \\ &= -\sum_{n=1}^{m} \int_{J_n} a(e_1(t), z_h(t))\, dt - \sum_{n=1}^{m-1} \left\langle e_1(t_n), [\![z_h]\!]_n \right\rangle_H - \langle e_1(t_m), \varphi\rangle_H . \end{aligned}$$

Introducing $\langle A e_1, z_h(t)\rangle_H = a(e_1, z_h)$ as above, using the Cauchy–Schwarz inequality, and estimating e_1 by its maximum pointwise in time yields

$$|\langle e_2(t_m), \varphi\rangle_H| \leq \left(\sup_{t\leq t_m} \|e_1(t)\|_H \right) \left(\sum_{n=1}^{m} \int_{J_n} \|A^* z_h(t)\|_H \, dt + \sum_{n=1}^{m-1} \left\| [\![z_h]\!]_n \right\|_H + \|\varphi\|_H \right).$$

From the dual definition of the norm in H and estimate (12.7), we obtain

$$\|e_2(t_m)\|_H \leq C \max_{1\leq n\leq m} \sup_{t\in J_n} \|e_1(t)\|_H . \tag{12.9}$$

It remains to bound $e_1(t)$ for arbitrary $t \in J_n$, which we do by estimating

$$\begin{aligned} \|e_1(t)\|_H &= \|u(t) - \mathcal{I}_h \pi_r u(t)\|_H \\ &\leq \|u(t) - \pi_r u(t)\|_H + \|\pi_r u(t) - \mathcal{I}_h \pi_r u(t)\|_H \\ &\leq C\tau_n \int_{J_n} \|d_t u(\tau)\|_H \, d\tau + C h^{s+1} \|u(t)\|_{H^{s+1}(\Omega)} , \end{aligned} \tag{12.10}$$

where we have used that the spatial approximation properties are independent of time and that $\pi_r u(t)$ has the same spatial regularity as $u(t)$. Combining (12.8), (12.9) and (12.10) 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 $\int_{J_n} \langle d_t z_h(t), u(t) - \pi_r u(t)\rangle_H \, dt$ vanishes by definition of π_r) to obtain[4]

Theorem 12.4. *For $r = 1$, the solutions $u \in X$ to (11.2) and $u_h \in X_h$ to (12.2) satisfy*

$$\|u(t_m) - u_m\|_H \leq C \max_{1\leq n\leq m} \left(h^{s+1} \sup_{t\in J_n} \|u(t)\|_{H^{s+1}(\Omega)} + \tau_n^3 \int_{J_n} \left\|d_t^2 u(t)\right\|_{H^2(\Omega)} dt \right)$$

for all $1 \leq m \leq N$.

The general case (including time-dependent bilinear form a and different discrete spaces V_m) can be found in [Chrysafinos & Walkington 2006].

[4] e.g., [Thomée 2006, Theorem 12.7]

BIBLIOGRAPHY

R. A. Adams & J. J. F. Fournier (2003), *Sobolev Spaces*, 2nd ed., Academic Press, Amsterdam.

D. Arnold, F. Brezzi, B. Cockburn & L. Marini (2002), Unified analysis of discontinuous Galerkin methods for elliptic problems, *SIAM Journal on Numerical Analysis* 39(5), 1749–1779, doi: 10.1137/s0036142901384162.

W. Bangerth, R. Hartmann & G. Kanschat (2013), deal.II Differential Equations Analysis Library, Technical Reference, url: http://www.dealii.org/.

D. Boffi, F. Brezzi & M. Fortin (2013), *Mixed and Finite Element Methods and Applications*, vol. 44, Springer Series in Computational Mathematics, Springer, New York, doi: 10.1007/978-3-642-36519-5.

D. Braess (2007), *Finite Elements*, 3rd ed., Cambridge University Press, Cambridge, doi: 10.1017/cbo9780511618635.

J. H. Bramble & S. R. Hilbert (1970), Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation. *SIAM J. Numer. Anal.* 7, 112–124, doi: 10.1137/0707006.

S. C. Brenner & L. R. Scott (2008), *The Mathematical Theory of Finite Element Methods*, 3rd ed., vol. 15, Texts in Applied Mathematics, Springer, New York, doi: 10.1007/978-0-387-75934-0.

K. Chrysafinos & N. J. Walkington (2006), Error estimates for the discontinuous Galerkin methods for parabolic equations, *SIAM J. Numer. Anal.* 44(1), 349–366 (electronic), doi: 10.1137/030602289.

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 & Applied Mathematics (SIAM), Philadelphia, PA, doi: 10.1137/1.9780898719208.

R. Courant (1943), Variational methods for the solution of problems of equilibrium and vibrations, *Bull. Amer. Math. Soc.* 49, 1–23, doi: 10.1090/s0002-9904-1943-07818-4.

D. A. Di Pietro & A. Ern (2012), *Mathematical Aspects of Discontinuous Galerkin Methods*, vol. 69, Mathématiques et Applications, Springer, New York, doi: 10.1007/978-3-642-22980-0.

R. E. Edwards (1965), *Functional Analysis. Theory and Applications*, Holt, Rinehart & Winston, New York.

A. Ern & J.-L. Guermond (2004), *Theory and Practice of Finite Elements*, vol. 159, Applied Mathematical Sciences, Springer, New York, doi: 10.1007/978-1-4757-4355-5.

A. Ern & J.-L. Guermond (2021), *Finite Elements III, First-Order and Time-Dependent PDEs*, vol. 74, Texts in Applied Mathematics, Springer, doi: 10.1007/978-3-030-57348-5.

L. C. Evans (2010), *Partial Differential Equations*, 2nd ed., vol. 19, Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, doi: 10.1090/gsm/019.

P. Grisvard (2011), *Elliptic Problems in Nonsmooth Domains*, Classics in Applied Mathematics 69, Society for Industrial & Applied Mathematics, doi: 10.1137/1.9781611972030.

O. A. Ladyzhenskaya & 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.

A. Logg, K.-A. Mardal, G. N. Wells, et al. (2012), *Automated Solution of Differential Equations by the Finite Element Method*, Springer, doi: 10.1007/978-3-642-23099-8,

J.-C. Nédélec (1980), Mixed finite elements in $\mathbb{R}^3$, *Numerische Mathematik* 35(3), 315–341, doi: 10.1007/bf01396415.

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

P. Raviart & J. Thomas (1977), A mixed finite element method for 2-nd order elliptic problems, in: *Mathematical Aspects of Finite Element Methods*, ed. by I. Galligani & E. Magenes, vol. 606, Lecture Notes in Mathematics, Springer, Berlin, 292–315, doi: 10.1007/bfb0064470.

M. Renardy & R. C. Rogers (2004), *An Introduction to Partial Differential Equations*, 2nd ed., vol. 13, Texts in Applied Mathematics, Springer, New York, doi: 10.1007/b97427.

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, doi: 10.1090/surv/049.

G. Strang (1972), Variational crimes in the finite element method, in: *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972)*, Academic Press, New York, 689–710, doi: 10.1016/b978-0-12-068650-6.50030-7.

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, Berlin, doi: 10.1007/3-540-33122-0.

G. M. Troianiello (1987), *Elliptic Differential Equations and Obstacle Problems*, The University Series in Mathematics, Plenum Press, New York, doi: 10.1007/978-1-4899-3614-1.

R. Verfürth (2013), *A Posteriori Error Estimation Techniques for Finite Element Methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, doi: 10.1093/acprof:oso/9780199679423.001.0001.

J. Wloka (1987), *Partial Differential Equations*, Translated from the German by C. B. Thomas and M. J. Thomas, Cambridge University Press, Cambridge, doi: 10.1017/cbo9781139171755.

E. Zeidler (1995a), *Applied Functional Analysis, Applications to mathematical physics*, vol. 108, Applied Mathematical Sciences, Springer, New York, doi: 10.1007/978-1-4612-0815-0.

E. Zeidler (1995b), *Applied Functional Analysis, Main principles and their applications*, vol. 109, Applied Mathematical Sciences, Springer, New York, doi: 10.1007/978-1-4612-0821-1.

M. Zlámal (1968), On the finite element method, *Numer. Math.* 12, 394–409, doi: 10.1007/bf02161362.