Lecture Notes in Advanced Computation in Fluid Dynamics

Johan Hoffman

Contents

Chapter 1

Scalar boundary value problems

The boundary value problem in one variable is an ordinary differential equation where boundary conditions are specified at each end of an interval. Contrary to the initial value problem, the independent variable does not represent time, but could instead be thought of as a spatial coordinate.

Instead of time stepping, we here present finite element methods to compute approximate solutions to the boundary value problem. In the framework of Sobolev spaces we prove that the finite element approximation is optimal, and we derive a posteriori error estimates that can be used in adaptive algorithms.

1.1 The boundary value problem

The boundary value problem

Consider the following model *boundary value problem*, for which we seek a function $u \in C^2([0,1])$, such that

$$-u''(x) = f(x), \quad x \in (0,1), \tag{1.1}$$
$$u(0) = u(1) = 0, \tag{1.2}$$

given a source term $f(x) \in C^0([0,1])$, and *boundary conditions* at the endpoints of the unit interval $I = [0,1]$. A boundary value problem is an ordinary differential equation just as the initial value problem, but with the difference that boundary values are prescribed instead of an initial value.

For the particular boundary value problem (1.1) we need to prescribe two boundary conditions, because to obtain a unique solution we must

determine two unknowns constants $C_0, C_1 \in \mathbb{R}$, since

$$\tilde{u}(x) = u(x) + C_1 x + C_0$$

is also a solution to the equation (1.1).

The type of boundary conditions (1.2), where the value of the solution is prescribed, is referred to as *Dirichlet boundary conditions*, whereas a condition on the derivative of the solution is a *Neumann boundary condition*.

Approximation of the boundary value problem

We seek an approximate solution to (1.1)-(1.2) in the form of a continuous piecewise polynomial,

$$U \in V_h = \{v \in V_h^{(q)} : v(0) = v(1) = 0\},$$

such that the error $e = u - U$ is small in some suitable norm $\|\cdot\|$.

The error cannot be evaluated without the exact solution, but we can compute the residual of the equation,

$$R(U(x)) = U''(x) + f(x),$$

with $R(u(x)) = 0$ for the exact solution $u(x)$. The residual measures how well an approximation satisfies the boundary value problem (1.1)-(1.2).

We recall three methods to compute $U \in V_h$ with a minimal residual: (i) the collocation method, which computes an approximation for which the residual is zero in a set of nodes x_i,

$$R(U(x_i)) = 0, \quad i = 1, ..., n,$$

(ii) the least squares method, where we seek the approximation with the minimal residual measured in the L_2-norm,

$$\min_{U \in V_h} \|R(U)\|,$$

and (iii) Galerkin's method, where we seek the approximation for which the residual is orthogonal the subspace V_h,

$$(R(U), v) = 0, \quad \forall v \in V_h. \tag{1.3}$$

By using finite difference stencils to approximate the derivatives at each node (i) becomes a finite difference method. With an approximation space consisting of piecewise polynomials, (ii) is a *least squares finite element method*, and (iii) a *Galerkin finite element method*. With an approximation space of sine waves, we refer to (iii) as a *spectral method*.

1.2 Galerkin finite element method

Galerkin finite element method

The finite element method (FEM) based on (1.3) takes the form: find the *trial function* $U \in V_h$, such that

$$\int_0^1 -U''(x)v(x)\,dx = \int_0^1 f(x)v(x)\,dx, \tag{1.4}$$

for all *test functions* $v \in V_h$. For (1.4) to be well defined, we must be able to represent the second order derivative U'', which is not obvious in the case of low order polynomials, for example, if $V_h \subset V_h^{(1)}$.

To relax this constraint, we can use partial integration to move one derivative from the trial function U to the test function v, so that

$$\int_0^1 -U''(x)v(x)\,dx = \int_0^1 U'(x)v'(x)\,dx, \tag{1.5}$$

since $v \in V_h$ and hence satisfies the boundary conditions (1.2) so that the boundary terms vanishes.

The Galerkin finite element method now reads: find $U \in V_h$, such that

$$\int_0^1 U'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx, \tag{1.6}$$

for all $v \in V_h$.

The discrete problem

Let V_h be the space of continuous piecewise linear functions over a mesh $\mathcal{T}_h$, with n internal nodes x_i, that satisfies the boundary conditions (1.2),

$$V_h = \{v \in V_h^{(1)} : v(0) = v(1) = 0\}.$$

With $\{\phi\}_{i=1}^n$ the nodal basis in V_h, we can write any function $v \in V_h$ in the form

$$v(x) = \sum_{i=1}^n v_i\phi_i(x), \tag{1.7}$$

with $v_i = v(x_i)$, and we thus search for an approximate solution

$$U(x) = \sum_{j=1}^n U_j\phi_j(x), \tag{1.8}$$

with $U_j = U(x_j)$. If we insert (1.7) and (1.8) into (1.6), we get

$$\sum_{j=1}^{n} U_j \int_0^1 \phi_j'(x)\phi_i'(x)\,dx = \int_0^1 f(x)\phi_i(x)\,dx, \quad i = 1, ..., n, \tag{1.9}$$

which corresponds to the matrix equation

$$Sx = b, \tag{1.10}$$

with $s_{ij} = (\phi_j', \phi_i')$, $x_j = U_j$, and $b_i = (f, \phi_i)$.

The *stiffness matrix* S is sparse, since $s_{ij} = 0$ for $|i - j| > 1$, and we compute the entries of S from the definition of the basis functions (**??**), starting with the diagonal entries,

$$\begin{aligned} s_{ii} &= (\phi_i', \phi_i') = \int_0^1 (\phi_i')^2(x)\,dx = \int_{x_{i-1}}^{x_i} (\lambda_{i,1}')^2(x)\,dx + \int_{x_i}^{x_{i+1}} (\lambda_{i+1,0}')^2(x)\,dx \\ &= \int_{x_{i-1}}^{x_i} \left(\frac{1}{h_i}\right)^2 dx + \int_{x_i}^{x_{i+1}} \left(\frac{1}{h_{i+1}}\right)^2 dx = \frac{1}{h_i} + \frac{1}{h_{i+1}}, \end{aligned}$$

and similarly we compute the off-diagonal entries,

$$s_{ii+1} = (\phi_i', \phi_{i+1}') = \int_0^1 \phi_i'(x)\phi_{i+1}'(x)\,dx = \int_{x_i}^{x_{i+1}} \frac{-1}{h_{i+1}}\frac{1}{h_{i+1}}\,dx = -\frac{1}{h_{i+1}},$$

and

$$s_{ii-1} = (\phi_i', \phi_{i-1}') = \int_0^1 \phi_i'(x)\phi_{i-1}'(x)\,dx = ... = -\frac{1}{h_i}.$$

In practise we use an assembly algorithm to compute the stiffness matrix S and *load vector* b, analogous to Algorithm **??** for the mass matrix. The traditional names of mass and stiffness matrices, and load vectors, comes from the area of structural mechanics, where much of the early development of finite element methods took place.

1.3 Analysis of the finite element method

The variational problem

Galerkin's method is based on the *variational formulation*, or *weak form*, of the boundary value problem, where we search for a solution in a vector space V, for which the variational form is well defined: find $u \in V$,

$$\int_0^1 u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx, \quad \forall v \in V. \tag{1.11}$$

If $u \in C^2([0,1])$ and $f \in C^0([0,1])$, then u is also the solution to the boundary value problem in strong form (1.1)-(1.2), which follows from partial integration and the continuity of the resulting integrand $f + u''$.

Sobolev spaces

To construct an appropriate function space V for which (1.11) is well defined, we need to extend L^2 spaces to include also derivatives. This leads us to *Sobolev spaces*, Hilbert spaces of functions with square integrable *weak derivatives*. For an integrable function $f \in L^1([0,1])$, we say that $g \in L^1([0,1])$ is the weak derivative of f, if

$$\int_0^1 f(x)\varphi'(x)\,dx = -\int_0^1 g(x)\varphi(x)\,dx,$$

for all $\varphi \in C^\infty([0,1])$ with $\varphi(0) = \varphi(1) = 0$. In what follows, derivatives in variational formulations are interpreted as weak derivatives, unless regular (strong) derivatives are known to exist.

The H^1 Hilbert spaces

We first introduce the Sobolev space

$$H^1([0,1]) = \{v : \|v\|_{H^1([0,1])} < \infty\},$$

which is a Hilbert space with norm

$$\|v\|_{H^1([0,1])} = (\|v\|^2 + \|v'\|^2)^{1/2},$$

and inner product

$$(v,w)_{H^1([0,1])} = (v,w) + (v',w'),$$

for which

$$(v,v)_{H^1([0,1])} = \|v\|^2_{H^1([0,1])}.$$

Next we define the Sobolev space of functions that also satisfy the boundary conditions of the variational problem,

$$H^1_0([0,1]) = \{v \in H^1(0,1) : v(0) = v(1) = 0\},$$

with the same norm and inner product as $H^1([0,1])$.

The variational form (1.11) is now well defined for $V = H_0^1([0,1])$, since by Cauchy-Schwarz inequality,

$$\int_0^1 u'(x)v'(x)\,dx \leq \|u'\|\|v'\| \leq \|u\|_{H^1([0,1])}\|v\|_{H^1([0,1])} < \infty, \tag{1.12}$$

and

$$\int_0^1 f(x)v(x)\,dx \leq \|f\|_{H^{-1}([0,1])}\|v\|_{H^1([0,1])} < \infty, \tag{1.13}$$

for $f \in H^{-1}([0,1])$, where we define the negative order Sobolev space by

$$H^{-1}([0,1]) = \{v : \|v\|_{H^{-1}([0,1])} < \infty\},$$

with the negative norm

$$\|v\|_{H^{-1}([0,1])} = \sup_{w \in H_0^1([0,1])} \frac{(v,w)}{\|w\|_{H^1([0,1])}} \tag{1.14}$$

from which the inequality (1.13) follows.

Optimality of Galerkin's method

For $u \in V = H_0^1([0,1])$ the solution to the variational problem (1.11), and $U \in V_h \subset V$ the solution to the Galerkin finite element method (1.6), we can prove that U is the best possible approximation in V_h, with respect to the *energy norm*,

$$\|w\|_E = \|w'\| = \left(\int_0^1 |w'(x)|^2\,dx\right)^{1/2}.$$

Hence $U \in V_h$ represents a projection of $u \in V$ onto V_h, with respect to an inner product defined on V,

$$(v,w)_E = (v',w') = \int_0^1 v'(x)w'(x)\,dx,$$

with $\|v\|_E^2 = (v,v)_E$. The fact that $(\cdot,\cdot)_E$ is an inner product and $\|\cdot\|_E$ a norm in $H_0^1([0,1])$, follows from *Poincaré-Friedrich inequality*,

$$\|v\|^2 \leq C\|v'\|, \quad \forall v \in H_0^1(0,1),$$

with $C > 0$. By subtracting (1.6) from (1.11), we expose a *Galerkin orthogonality* property,

$$(u - U, v)_E = 0, \quad \forall v \in V_h,$$

which we can use to express the optimality of the approximation U.

Theorem 1. *The Galerkin solution $U \in V_h$ is the optimal approximation of u with respect to the energy norm,*

$$\|u - U\|_E \leq \|u - v\|_E, \quad \forall v \in V_h.$$

Proof. For any $v \in V_h$,

$$\begin{aligned} \|u - U\|_E^2 &= (u - U, u - u_h)_E = (u - U, u - v)_E + (u - U, v - u_h)_E \\ &= (u - U, u - v)_E \leq \|u - U\|_E \|u - v\|_E. \end{aligned}$$

□

An a priori error estimate follows from the observation that the Galerkin error can be bounded by an interpolation error by choosing $v = \pi_h u \in V_h$, which in turn is bounded by a derivative of the exact solution,

$$\|u - U\|_E \leq \|u - \pi_h u\|_E = \|u' - (\pi_h u)'\| \leq C\|hu''\|.$$

A posteriori error estimation

Consider the following adjoint problem in variational form: find $\varphi \in V$,

$$(v', \varphi') = (v, \psi), \quad v \in V, \tag{1.15}$$

with $\psi \in V$, and $V = H_0^1([0,1])$. By Riesz representation theorem and (1.15), any continuous linear functional $F \in V^*$ of the solution to (1.11) can be expressed as

$$F(u) = (u, \psi) = (u', \varphi') = (f, \varphi),$$

where the Riesz representer $\psi \in V$ is the data to the adjoint problem (1.15). Hence the solution to the adjoint problem acts as a Green's function, by which the functional F of any solution to (1.11) can be computed from the source term f in (1.11).

The Galerkin approximation error with respect to the functional is

$$F(u) - F(U) = (u - U, \psi) = (u' - U', \varphi') = r(U, \varphi) = (R(U), \varphi),$$

with the *weak residual*

$$r(U, \varphi) = (f, \varphi) - (U', \varphi'),$$

and the *strong residual*

$$R(U(x)) = f(x) + U''(x). \tag{1.16}$$

We can derive a posteriori error estimates based on both the weak and strong residuals, to be used, for example, in an adaptive algorithm.

$$\begin{aligned} |F(u) - F(U)| &= r(U,\varphi) = \sum_{i=1}^{n+1} \int_{I_i} (f\varphi - U'\varphi')\, dx, \\ |F(u) - F(U)| &= (R(U),\varphi) = (R(U), \varphi - \pi_h\varphi) \\ &= \sum_{i=1}^{n+1} \int_{I_i} R(U)(\varphi - \pi_h\varphi)\, dx \leq \sum_{i=1}^{n+1} C_i \|h_i^2 R(U)\| \|\varphi''\|. \end{aligned}$$

The equation (1.5) suggests a method to approximate the term $U''(x)$ in the strong residual (1.16) for $U \in V_h = V_h^{(1)}$, as the function $w \in V_h$,

$$\int_0^1 w(x)v(x)\, dx = \int_0^1 U'(x)v'(x)\, dx, \quad \forall v \in V_h.$$

1.4 Exercises

Problem 1. *Derive the variational formulation and the finite element method for the boundary value problem*

$$-(a(x)u'(x))' + c(x)u(x) = f(x), \quad x \in (0,1), \tag{1.17}$$
$$u(0) = u(1) = 0, \tag{1.18}$$

with $a(x) > 0$, and $c(x) \geq 0$.

Problem 2. *Derive the variational formulation and the finite element method for the boundary value problem*

$$\begin{aligned} -(a(x)u'(x))' &= f(x), \quad x \in (0,1), && (1.19) \\ u(0) &= 0, && (1.20) \\ -a(x)u'(1) &= 1, && (1.21) \end{aligned}$$

with $a(x) > 0$.

Problem 3. *Show that the inequality (1.13) follows from (1.14).*

Problem 4. *Show that for the boundary value problem (1.1)-(1.2), $(\cdot,\cdot)_E$ is an inner product and $\|\cdot\|_E$ a norm in $H_0^1([0,1])$.*

Problem 5. *Show that for the boundary value problem (1.1)-(1.2), the energy norm $\|\cdot\|_E$ and $\|\cdot\|_{H^1([0,1]}$ are equivalent. That is, show that*

$$C_1\|f\|_E \leq \|f\|_{H^1([0,1])} \leq C_2\|f\|_E,$$

for some constants $C_1, C_2 > 0$.

Chapter 2

Function spaces in $\mathbb{R}^n$

In this chapter we develop the mathematical framework needed to address partial differential equations to $\mathbb{R}^n$. Not only are the differential and integral operators in multiple dimensions richer than in one dimension, but the geometry of the domain and the evaluation of functions on the boundary is more intricate.

We define function spaces for continuous functions and integrable functions, specifically we introduce Sobolev spaces for integrable functions and their weak derivatives.

2.1 Differential operators in $\mathbb{R}^n$

The domain in $\mathbb{R}^n$

In this chapter we prepare to study partial differential equations defined on a domain $\Omega \subset \mathbb{R}^n$. If the *diameter* $\operatorname{diam}(\Omega) < \infty$, we say that the domain is *bounded*, with

$$\operatorname{diam}(\Omega) = \sup_{x,y\in\Omega} \|x - y\|.$$

The domain $\Omega \subset \mathbb{R}^n$ is an *open* set if for every $x \in \Omega$ it exists an *open ball*

$$B(x, r) = \{y \in \mathbb{R}^n : \|x - y\| < r\}$$

such that $B(x, r) \subset \Omega$, and Ω is a *closed* set if its complement Ω^c is open.

The *boundary* of Ω is defined as $\partial\Omega = \overline{\Omega} \cap \overline{\Omega^c}$, where the set $\overline{\Omega}$ is the *closure* of Ω in $\mathbb{R}^n$, which consists of all $x \in \mathbb{R}^n$ such that for any $\epsilon > 0$, there exists a $y \in \Omega$ so that $\|x-y\| < \epsilon$. A closed set contains its boundary, whereas an open set does not.

Partial differentiation

$D_j = \partial/\partial x_j$ is the differential operator of partial differentiation, and

$$D^\alpha = D_1^{\alpha_1} \cdots D_n^{\alpha_n}$$

is a differential operator of order $|\alpha| = \alpha_1 + ... + \alpha_n$, where $\alpha = (\alpha_1, ..., \alpha_n)^T$ is a *multi-index*, with $\alpha_i \geq 0$ for all $i = 1, ..., n$.

Gradient and Jacobian

The *gradient* of a scalar function $f \in C^1(\Omega)$ is denoted by

$$\operatorname{grad} f = \nabla f = (D_1 f, ..., D_n f)^T,$$

or in index notation $D_i f$, with the *nabla* operator

$$\nabla = (D_1, ..., D_n)^T.$$

Further, the *directional derivative* $\nabla_v f$, of f in the direction of the vector field $v : \mathbb{R}^n \to \mathbb{R}^n$, is defined as

$$\nabla_v f = (v \cdot \nabla) f = v_j D_j f.$$

For the $C^1(\Omega)$ vector field $F : \mathbb{R}^n \to \mathbb{R}^m$, we define the *Jacobian* J,

$$J = F' = \nabla F = \begin{bmatrix} D_1 F_1 & \cdots & D_n F_1 \\ \vdots & \ddots & \vdots \\ D_1 F_m & \cdots & D_n F_m \end{bmatrix} = \begin{bmatrix} (\nabla F_1)^T \\ \vdots \\ (\nabla F_m)^T \end{bmatrix} = D_j F_i,$$

with directional derivative

$$\nabla_v F = (v \cdot \nabla) F = Jv = v_j D_j F_i.$$

Divergence and rotation

For a $C^1(\Omega)$ vector field $F : \mathbb{R}^n \to \mathbb{R}^n$, we define the *divergence*

$$\operatorname{div} F = \nabla \cdot F = D_1 F_1 + ... + D_n F_n = \frac{\partial F_i}{\partial x_i},$$

and, for $n = 3$, the *rotation*,

$$\operatorname{rot} F = \operatorname{curl} F = \nabla \times F,$$

where

$$\nabla \times F = \nabla \wedge F = (D_2F_3 - D_3F_2, D_3F_1 - D_1F_3, D_1F_2 - D_2F_1).$$

The divergence can be interpreted in terms of *Gauss theorem*, which states that the volume integral of the divergence of F in $\Omega \subset \mathbb{R}^n$, is equal to a surface integral over $\partial\Omega$ of F projected in the direction of the unit outward normal n of $\partial\Omega$,

$$\int_\Omega \nabla \cdot F \, dx = \int_{\partial\Omega} F \cdot n \, ds,$$

with the surface integral defined by a suitable parameterization of $\partial\Omega$.

Similarly, the rotation can be viewed through the *Kelvin-Stokes theorem*, which relates the surface integral of the rotation over a surface Σ, to the curve integral over its bounding curve $\partial\Sigma$,

$$\int_\Sigma \nabla \times F \cdot ds = \int_{\partial\Sigma} F \cdot dr,$$

with the integrals defined by the parameterizations of Σ and $\partial\Sigma$.

Laplacian and Hessian

For a scalar function $f \in C^2(\mathbb{R}^n)$, we define the *Laplacian*

$$\Delta f = \nabla^2 f = \nabla^T \nabla f = \nabla \cdot \nabla f = D_1^2 f + ... + D_n^2 f = D_i^2 f,$$

and the *Hessian*,

$$H = f'' = \nabla \nabla^T f = \begin{bmatrix} D_1^2 f & \cdots & D_1 D_n f \\ \vdots & \ddots & \vdots \\ D_n D_1 f & \cdots & D_n^2 f \end{bmatrix} = D_i D_j f.$$

The *vector Laplacian* of a $C^2(\Omega)$ vector field $F : \mathbb{R}^n \to \mathbb{R}^n$, is defined as

$$\Delta F = \nabla^2 F = (\Delta F_1, ..., \Delta F_n)^T,$$

and for $n = 3$,

$$\Delta F = \nabla(\nabla \cdot F) - \nabla \times (\nabla \times F).$$

2.2 Function spaces

Spaces of continuous functions

For $\Omega \subset \mathbb{R}^n$, we define the set of functions with k continuous derivatives,

$$C^k(\Omega) = \{\phi : D^\alpha \phi \in C(\Omega),\ |\alpha| \leq k\},$$

with $C(\Omega) = C^0(\Omega)$ and $C^\infty(\Omega) = \cap_{k=0}^\infty C^k(\Omega)$. The subset $C_0^k(\Omega)$ consists of the functions $\phi \in C^k(\Omega)$ that have *compact support* in Ω, that is, the support

$$\text{supp}(\phi) = \{x \in \Omega : \phi(x) \neq 0\},$$

is closed and bounded.

The $L^p(\Omega)$ Banach spaces

For $\Omega \subset \mathbb{R}^n$ an open set and p a positive real number, we denote by $L^p(\Omega)$ the class of all Lebesgue measurable functions u defined on Ω, such that

$$\int_\Omega |u(x)|^p \, dx < \infty,$$

where we identify functions that are equal almost everywhere in Ω.

$L^p(\Omega)$ is a Banach space for $1 \leq p < \infty$, with the norm

$$\|u\|_{L^p(\Omega)} = \left(\int_\Omega |u(x)|^p \, dx \right)^{1/p}.$$

In the case of a vector valued function $u : \mathbb{R}^n \to \mathbb{R}^m$, we replace the integrand in the definitions by the l^p norm, and for a matrix function $u : \mathbb{R}^n \to \mathbb{R}^{m \times k}$, a generalized Frobenius norm

$$\sum_i^m \sum_j^k |u_{ij}(x)|^p.$$

$L^\infty(\Omega)$ is a Banach space with the norm

$$\|u\|_{L^\infty(\Omega)} = \text{ess sup}_{x \in \Omega} |u(x)|,$$

where for vector and matrix valued functions, the maximum is taken over the components.

The Minkowski inequality holds in $L^p(\Omega)$, for all $1 \leq p \leq \infty$.

Theorem 2 (Minkowski's inequality for $L^p(\Omega)$)**.** *For* $1 \leq p \leq \infty$,

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p, \quad \forall f, g \in L^p(\Omega).$$

Duality and the Hilbert space $L^2(\Omega)$

$L^2(\Omega)$ is a Hilbert space with the inner product

$$(u, v) = (u, v)_{L^2(\Omega)} = \int_\Omega u(x)v(x)\,dx, \tag{2.1}$$

which induces the $L^2(\Omega)$ norm. For vector valued functions the integrand is replaced by the l_2 inner product, and for matrix functions by the Frobenius inner product. In what follows, we let $\|\cdot\| = \|\cdot\|_{L^2(\Omega)}$.

The Hölder inequality holds for L^p spaces, and analogous to l^p spaces the integral (2.1) defines a duality pairing for L^p spaces, where $L^q(\Omega)$ is isometrically isomorphic to the dual space $(L^p(\Omega))'$, for $1 \leq p < \infty$ and q the dual index of p.

Theorem 3 (Hölder's inequality for $L^p(\Omega)$)**.** *Let $1 \leq p, q \leq \infty$ and $1/p + 1/q = 1$. If $f \in L^p(\Omega)$ and $g \in L^q(\Omega)$ then $fg \in L^1(\Omega)$, and*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

Specifically, we have that

$$\|fg\|_1 \leq \|f\|_1 \|g\|_\infty,$$

for $f \in L^1(\Omega)$ and $g \in L^\infty(\Omega)$.

The dual space $(L^2(\Omega))'$ is isometrically isomorphic to $L^2(\Omega)$ itself, with the duality pairing equal to the L^2 inner product.

The weak derivative

With $\Omega \subset \mathbb{R}^n$ an open domain, $L^1_{loc}(\Omega)$ is the space of functions that are absolute integrable over all closed and bounded subsets $K \subset \Omega$. Hence for $1 \leq p \leq \infty$, all $f \in L^p(\Omega)$ are also in $L^1_{loc}(\Omega)$, since by Hölder's inequality,

$$\int_K |f(x)|\,dx = \int_\Omega |f(x)\chi_K(x)|\,dx \leq \|f\|_p \left(\int_K dx\right)^{1/q} = \|f\|_p |K|^{1/q} < \infty,$$

with $1/p + 1/q = 1$, and the indicator function

$$\chi_K(x) = \begin{cases} 1, & x \in K, \\ 0, & x \notin K. \end{cases}$$

Alternatively, we can characterize $L^1_{loc}(\Omega)$ as the space of functions which are absolute integrable when multiplied by smooth test functions,

$$L^1_{loc}(\Omega) = \{u : \int_\Omega |u\phi|\, dx < \infty, \forall \phi \in C_0^\infty\}.$$

For $u \in L^1_{loc}(\Omega)$, if there exists a function $v_\alpha \in L^1_{loc}(\Omega)$, such that

$$\int_\Omega u(x) D^\alpha \phi(x)\, dx = (-1)^{|\alpha|} \int_\Omega v_\alpha(x)\phi(x)\, dx, \quad \forall \phi \in C_0^\infty,$$

then v_α is the *weak partial derivative* of u, unique almost everywhere. We will use the notation $D^\alpha u = v_\alpha$ also for the weak derivative, which coincides with the classical derivative when it exists.

Mollifiers and regularization

Consider a nonnegative function $J : \mathbb{R}^n \to \mathbb{R}$ which belongs to $C_0^\infty(\mathbb{R}^n)$, such that $J(x) = 0$ for $\|x\| \geq 1$, and

$$\int_{\mathbb{R}^n} J(x)\, dx = 1,$$

for example,

$$J(x) = \begin{cases} k \exp(-1/(1 - \|x\|^2)), & \|x\| < 1, \\ 0, & \|x\| \geq 0, \end{cases}$$

with $k > 0$ a normalization constant. We can then define the *mollifier*

$$J_\epsilon(x) = \epsilon^{-n} J(x/\epsilon),$$

with the properties that $J_\epsilon \in C_0^\infty(\mathbb{R}^n)$, $J_\epsilon(x) = 0$ for $\|x\| \geq \epsilon$, and

$$\int_{\mathbb{R}^n} J_\epsilon(x)\, dx = 1.$$

For a locally integrable function $u \in L^1_{loc}(\mathbb{R}^n)$ the convolution with a mollifier is smooth $J_\epsilon * u \in C^\infty(\mathbb{R}^n)$, since

$$D^\alpha(J_\epsilon * u) = D^\alpha \int_{\mathbb{R}^n} J_\epsilon(x - y) u(y)\, dy = \int_{\|x-y\|<\epsilon} D^\alpha_x J_\epsilon(x - y) u(y)\, dy,$$

and if u has compact support in Ω then $J_\epsilon * u \in C_0^\infty(\mathbb{R}^n)$, provided that

$$\epsilon < \inf_{x \in \mathrm{supp}(u)} \|x - \partial\Omega\|.$$

Further, it can be proven that if $u \in L^p(\Omega)$ for $1 \leq p < \infty$, then $J_\epsilon * u \in L^p(\Omega)$, $\|J_\epsilon * u\|_{L^p(\Omega)} \leq \|u\|_{L^p(\Omega)}$ and

$$\lim_{\epsilon \to 0^+} \|J_\epsilon * u - u\|_{L^p(\Omega)} = 0.$$

Hence mollifiers provide a method to approximate an integrable function by a smooth function to an arbitrary precision, which we refer to as a *regularization* of the function.

The Sobolev spaces $W^{k,p}$ and $W_0^{k,p}$

To construct appropriate vector spaces for partial differential equations, we extend the L^p spaces with derivatives. We first define the *Sobolev norms*,

$$\|u\|_{k,p} = \left(\sum_{0 \leq |\alpha| \leq k} \|D^\alpha u\|_{L^p(\Omega)}^p \right)^{1/p},$$

for $1 \leq p < \infty$, and

$$\|u\|_{k,\infty} = \max_{0 \leq |\alpha| \leq k} \|D^\alpha u\|_{L^\infty(\Omega)},$$

where $D^\alpha u$ refers to weak derivatives. Equipped with the Sobolev norms, we then define the *Sobolev spaces*,

$$W^{k,p}(\Omega) = \{u \in L^p(\Omega) : D^\alpha u \in L^p(\Omega),\, 0 \leq |\alpha| \leq k\},$$

for each positive integer k and $1 \leq p \leq \infty$, with $W^{0,p}(\Omega) = L^p(\Omega)$.

By the *Sobolev embedding theorem*, each element in the Sobolev space $W^{k,p}(\Omega)$ is the limit of a sequence of functions in $C^k(\Omega)$ with respect to the norm $\|\cdot\|_{k,p}$, and we define the Sobolev space $W_0^{k,p}(\Omega)$ as the closure of $C_0^\infty(\Omega)$ in $W^{k,p}(\Omega)$. In other words, any function in $W^{k,p}(\Omega)$ or $W_0^{k,p}(\Omega)$ can be approximated by a continuous function with k continuous derivatives, to an arbitrary precision with respect to the Sobolev norm.

The Hilbert spaces H^k and H_0^k

The Sobolev spaces $H^k(\Omega) = W^{k,2}(\Omega)$ and $H_0^k(\Omega) = W_0^{k,2}(\Omega)$ are Hilbert spaces with the inner product and associated norm

$$(u, v)_k = \sum_{0 \leq \alpha \leq k} (D^\alpha u, D^\alpha v), \quad \|u\|_k = (u, u)_k^{1/2},$$

for which Cauchy-Schwarz inequality is satisfied,

$$|(u,v)_k| \leq \|u\|_k \|v\|_k.$$

We denote by $H^{-k}(\Omega)$ the dual space of $H_0^k(\Omega)$, with the norm

$$\|u\|_{-k} = \sup_{v \in H_0^k(\Omega)} \frac{|(u,v)|}{\|v\|_k} = \sup_{v \in H_0^k(\Omega): \|v\|_k = 1} |(u,v)|,$$

satisfying a generalized Hölder inequality for $u \in H^{-k}(\Omega)$ and $v \in H_0^k(\Omega)$,

$$|(u,v)| \leq \|u\|_{-k} \|v\|_k.$$

The spaces relate to each other as $H^k(\Omega) \subset L^2(\Omega) \subset H^{-k}(\Omega)$.

Boundary traces

For functions in Sobolev spaces we cannot prescribe boundary conditions for each $x \in \partial\Omega$ as for continuous functions. Instead we use a *trace operator*

$$T : H^1(\Omega) \to L^2(\partial\Omega),$$

which exists by the Sobolev embedding theorem, and for which we have the following trace inequality

$$\|Tu\|_{L^2(\partial\Omega)} \leq C\|u\|_1.$$

Hence a function in $H^1(\Omega)$ can be restricted to the boundary $\partial\Omega$ as a function in $L^2(\partial\Omega)$ by the trace operator. To simplify notation, we will not write out the trace operator explicitly, but instead write

$$\|u\|_{L^2(\partial\Omega)} \leq C\|u\|_1.$$

The results above rely on the boundary to satisfy a Lipschitz regularity condition, which we will assume valid from now on. We say that $\partial\Omega$ is a *Lipschitz boundary* if $\partial\Omega$ locally can be represented as the graph of a Lipschitz continuous function. With this assumption we also have the following theorem.

Theorem 4 (Poincaré-Friedrich's inequality)**.** *For all $u \in H^1(\Omega)$ there exists a constant $C > 0$, such that*

$$\|u\|_{L^2(\Omega)}^2 \leq C(\|u\|_{L^2(\partial\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2).$$

Partial integration in $\mathbb{R}^n$

For a scalar function $f : \mathbb{R}^n \to \mathbb{R}$, and a vector function $F : \mathbb{R}^n \to \mathbb{R}^n$, we have the following generalization of partial integration over a domain $\Omega \subset \mathbb{R}^n$, referred to as *Green's theorem*,

$$(\nabla f, F) = -(f, \nabla \cdot F) + \langle f, F \cdot n \rangle,$$

with n the unit outward normal vector for the boundary $\partial\Omega$, and where we use the notation

$$\langle v, w \rangle = (v, w)_{L^2(\partial\Omega)} \tag{2.2}$$

for the boundary integral. With $F = \nabla g$ and $g : \mathbb{R}^n \to \mathbb{R}$ a scalar function,

$$(\nabla f, \nabla g) = -(f, \Delta g) + \langle f, \nabla g \cdot n \rangle.$$

Chapter 3

Boundary value problems in $\mathbb{R}^n$

In this chapter we make the transition to partial differential equations, in the form of boundary value problems in $\mathbb{R}^n$. While the theory and methods from the one dimensional boundary value problem to a large part extend to multiple dimensions, there are some important differences.

For linear elliptic boundary value problems we give a complete theory of existence and uniqueness of solutions, but for nonlinear boundary value problems no general mathematical theory is available.

3.1 Approximation of Poisson's equation

The Poisson equation

We now consider the *Poisson equation* for a function $u : \mathbb{R}^n \to \mathbb{R}$,

$$-\Delta u = f, \quad \text{in } \Omega, \tag{3.1}$$

with the domain $\Omega \subset \mathbb{R}^n$, and $f : \mathbb{R}^n \to \mathbb{R}$ given data. The Poisson equation is used in a number of scientific fields, from electrostatics to aerodynamics. Specifically, the Poisson equation is used to model potential fields, such as the gravitational potental or the electric potential. If $f = 0$, we refer to (3.1) as the *Laplace equation*.

For the equation to have a unique solution we need to specify boundary conditions. We may prescribe *Dirichlet boundary conditions*,

$$u|_{\partial\Omega} = g_D, \tag{3.2}$$

Neumann boundary conditions,

$$\nabla u \cdot n|_{\partial\Omega} = g_N, \tag{3.3}$$

or a linear combination of the two, referred to as a *Robin boundary condition*,

$$\nabla u \cdot n|_{\partial\Omega} = \alpha(u|_{\partial\Omega} - g_D) + g_N, \tag{3.4}$$

with $\alpha(x)$ a given weight function.

The existence of a unique solution to the Poisson equation depends on the boundary conditions and the given data f.

Homogeneous Dirichlet boundary conditions

With homogeneous Dirichlet boundary conditions, we have the problem

$$\begin{aligned} -\Delta u &= f, \qquad \text{in } \Omega, \\ u &= 0, \qquad \text{on } \partial\Omega. \end{aligned} \tag{3.5}$$

We will use a Galerkin method to compute approximate solutions to Poisson's equation, and thus the proper statement of the problem is in weak form, that is, a variational formulation. The weak form is obtained by taking the inner product of the equation (3.1) with a test function v, and then using Green's theorem to move one derivative from the solution u to the test function v.

To make the problem statement precise, let the trial and test functions belong to a certain function space V. With $V = H_0^1(\Omega)$ we obtain the following variational formulation: find $u \in V$, such that

$$(\nabla u, \nabla v) = (f, v), \quad \forall v \in V, \tag{3.6}$$

since the boundary term vanishes as the test function is an element of the vector space $H_0^1(\Omega)$.

Homogeneous Neumann boundary conditions

Now consider the Poisson equation with homogeneous Neumann boundary conditions,

$$\begin{aligned} -\Delta u &= f, \qquad \text{in } \Omega, \\ \nabla u \cdot n &= 0, \qquad \text{on } \partial\Omega. \end{aligned} \tag{3.7}$$

With $V = H^1(\Omega)$ we have the following variational formulation: find $u \in V$, such that

$$(\nabla u, \nabla v) = (f, v), \quad \forall v \in V, \tag{3.8}$$

since the boundary term vanishes due to the homogeneous Neumann boundary condition. Hence the variational forms (3.6) and (3.8) are similar, with the only difference being the choice of test and trial space V.

However, it turns out that the variational problem (3.8) has no unique solution, since for any solution $u \in V$, also $u + C$ is a solution, with $C \in \mathbb{R}$ any constant. To ensure a unique solution, we need an extra condition for the solution which determines the arbitrary constant, for example, we may change the trial space to

$$V = \{u \in H^1(\Omega) : \int_\Omega u(x)\, dx = 0\}. \tag{3.9}$$

Non-homogeneous boundary conditions

The Poisson equation with non-homogeneous Dirichlet and Neumann boundary conditions takes the following form,

$$\begin{aligned} -\Delta u &= f, && \text{in } \Omega, \\ u &= g_D, && \text{on } \Gamma_D, \\ \nabla u \cdot n &= g_N, && \text{on } \Gamma_N, \end{aligned}$$

with $\partial\Omega = \Gamma_D \cup \Gamma_N$.

To enforce the non-homogeneous Dirichlet condition, we need to extend the Dirichlet boundary data to the rest of the domain. This is done by introducing an arbitrary function $G_D \in H^1(\Omega)$ with the property that $G_D|_{\partial\Gamma_D} = g_D$. By Green's theorem, we then get the following variational formulation: find $\bar{u} = u - G_D \in V$, with $V = H_0^1(\Omega)$, such that

$$(\nabla \bar{u}, \nabla v) = (f, v) - (\nabla G_D, \nabla v) + (g_N, v)_{L^2(\Gamma_N)}, \tag{3.10}$$

for all $v \in V$, from which we obtain the weak solution $u = \bar{u} + G_D$.

We note the difference that the Dirichlet boundary condition is enforced through the trial space, denoted an *essential boundary condition*, whereas the Neumann boundary condition is enforced through the variational form, referred to as a *natural boundary condition*.

Robin boundary conditions

With Robin boundary conditions the Poisson equation takes the form

$$\begin{aligned} -\Delta u &= f, && \text{in } \Omega, \\ \nabla u \cdot n &= \alpha(u - g_D) + g_N, && \text{on } \partial\Omega, \end{aligned} \tag{3.11}$$

with the variational formulation: find $u \in V$, such that

$$(\nabla u, \nabla v) = (f, v) + (\alpha(u - g_D), v)_{L^2(\partial\Omega)} + (g_N, v)_{L^2(\partial\Omega)}, \tag{3.12}$$

for all $v \in V$, with $V = H^1(\Omega)$.

Note that for $\alpha = 0$ we recover the variational form for Poisson's equation with a Neumann boundary condition, whereas $\alpha^{-1} = 0$ corresponds to a Dirichlet boundary condition.

Green's functions and boundary integral methods

Consider the Poisson equation with non-homogeneous Dirichlet boundary conditions,

$$\begin{aligned} -\Delta u &= f, && \text{in } \Omega, \\ u &= g_D, && \text{on } \partial\Omega, \end{aligned}$$

with continuous data f, g_D. The corresponding *Green's function* is defined as the solution to the equation

$$\begin{aligned} -\Delta G_x &= \delta_x, && \text{in } \Omega, \\ G_x &= 0, && \text{on } \partial\Omega. \end{aligned}$$

If $u \in C^2(\Omega)$ is the solution to Poisson's equation, for any point $x \in \Omega$,

$$u(x) = -\int_{\partial\Omega} g_D(y) \nabla G_x(y) \cdot n \, dy + \int_\Omega f(y) G_x(y) \, dy,$$

assuming that we can construct the Green's function. Such representation formulas can be used to develop *boundary integral methods* to compute approximate solutions to Poisson's equation. For Laplace equation, with $f = 0$, the representation formula only involves a boundary integral, and thus the solution at any $x \in \Omega$, even for exterior unbounded domains, can be approximated without any need to discretize the domain Ω, only its boundary $\partial\Omega$.

While Green's functions only exist for certain partial differential equations for specific domains Ω, when available they can be very powerful tools for approximation.

Galerkin methods

To formulate a Galerkin method for the Poisson equation we replace the Hilbert space V by a finite dimensional subspace $V_h \subset V$ in the variational formulation of the equation. We hence seek $U \in V_h$, such that

$$(\nabla U, \nabla v) = (f, v), \quad \forall v \in V_h. \tag{3.13}$$

A finite element method is a Galerkin method for which V_h is constructed from a set of finite elements defined over a mesh. As opposed to boundary integral methods, finite element methods can be applied to a general partial differential equation defined on any domain.

For periodic domains, spectral methods can be formulated where trigonometric Fourier expansions are used to construct V_h.

3.2 Elliptic partial differential equations

The abstract problem

We can express a general linear partial differential equation as the abstract problem,

$$A(u) = f, \quad \text{in } \Omega, \tag{3.14}$$

with boundary conditions,

$$B(u) = g, \quad \text{on } \partial\Omega. \tag{3.15}$$

For a Hilbert space V consisting of functions with finite norm $\|\cdot\|_V$, we formulate the corresponding variational problem: find $u \in V$, such that

$$a(u, v) = L(v), \quad \forall v \in V, \tag{3.16}$$

with $a : V \times V \to \mathbb{R}$ a bilinear form and $L : V \to \mathbb{R}$ a linear form. Any linear partial differential equation can be expressed in weak form as (3.16), and for the variational problem to make sense, both the bilinear form and the linear form must be bounded,

$$\begin{aligned} a(u, v) &\leq C_1\|u\|_V\|v\|_V, \quad \forall u, v \in V, \\ L(v) &\leq C_2\|v\|_V, \quad \forall v \in V. \end{aligned}$$

We say that the bilinear form is *coercive*, or *elliptic*, if

$$a(v, v) \geq \alpha\|v\|_V^2, \quad \forall v \in V,$$

and we will refer to the variational problem (3.16) as elliptic if the bilinear form is elliptic and bounded, and the linear form is bounded.

Uniqueness of solution

A solution $u \in V$ to an elliptic variational problem (3.16) is unique, since if we assume that $\tilde{u} \in V$ is also a solution, then

$$a(u - \tilde{u}, v) = 0, \quad \forall v \in V, \tag{3.17}$$

by the linearity of the bilinear form, so that

$$\|u - \tilde{u}\|_V^2 \leq \alpha^{-1} a(u - \tilde{u}, u - \tilde{u}) = 0,$$

by the ellipticity of the bilinear form, and (3.17) since $u - \tilde{u} \in V$.

Existence of solution

If the bilinear form is symmetric it defines an inner product on V, and then by Riesz representation theorem there exists a unique $u \in V$ that satisfies the variational problem (3.16).

For the general case when the bilinear form may not be symmetric, we can use the Banach fixed point theorem to prove existence of a solution.

Theorem 5 (Lax-Milgram theorem)**.** *The variational problem (3.16) has a unique solution $u \in V$, if the bilinear form is elliptic and bounded, and the linear form is bounded. That is, there exist constants $\alpha > 0$, $C_1, C_2 < \infty$, such that for $u, v \in V$,*

$$\begin{aligned} &(i) && a(v,v) \geq \alpha \|v\|_V^2, \\ &(ii) && a(u,v) \leq C_1 \|u\|_V \|v\|_V, \\ &(iii) && L(v) \leq C_2 \|v\|_V. \end{aligned}$$

Proof. Let $A[u] \in V'$ be defined by $A[u](v) = a(u, v)$, so that we can express the variational problem (3.16) in V' as $A[u] = L$, or equivalently, by Riesz representation theorem,

$$\phi_{A[u]} = \phi_L,$$

with $\phi_{A[u]}, \phi_L \in V$ the corresponding Riesz representers, defined by

$$A[u](v) = (\phi_{A[u]}, v)_V, \quad L(v) = (\phi_L, v)_V, \quad \forall v \in V.$$

The map $\Phi : u \mapsto \phi_{A[u]}$ is a bounded linear map on V, by the linearity and boundedness of the bilinear form, since for $u, w \in V$ and $c \in \mathbb{R}$,

$$\begin{aligned} (\Phi(cu + w), v)_V &= (\phi_{A[cu+w]}, v)_V = a(cu + w, v) = ca(u, v) + a(w, v) \\ &= (c\Phi(u), v)_V + (\Phi(w), v)_V = (c\Phi(u) + \Phi(w), v)_V, \end{aligned}$$

for all $v \in V$, specifically for $v = \Phi(cu + w) - (c\Phi(u) + \Phi(w)) \in V$, and

$$\|\Phi(u)\| = \sup_{v \in V : \|v\|_V = 1} (\Phi(u), v)_V = \sup_{v \in V : \|v\|_V = 1} a(u, v) \leq C_1 \|u\|_V.$$

To prove the existence of a solution to the variational problem (3.16) we formulate the equivalent fixed point $T : V \to V$,

$$T(u) = u - \epsilon(\Phi(u) - \phi_L), \tag{3.18}$$

which is a contraction for a sufficiently small $\epsilon > 0$, since

$$\begin{aligned}
\|T(u) - T(w)\|_V^2 &= \|u - w - \epsilon(\Phi(u) - \Phi(w))\|_V^2 \\
&= (u - w - \epsilon(\Phi(u) - \Phi(w)), u - w - \epsilon(\Phi(u) - \Phi(w)))_V \\
&= \|u - w\|_V^2 - 2\epsilon(u - w, \Phi(u - w))_V + \epsilon^2\|\Phi(u) - \Phi(w)\|_V^2 \\
&= \|u - w\|_V^2 - 2\epsilon a(u - w, u - w) + \epsilon^2\|\Phi(u - w)\|_V^2 \\
&\leq \|u - w\|_V^2 - 2\epsilon\alpha\|u - w\|_V^2 + \epsilon^2 C_1\|u - w\|_V^2 \\
&= (1 - 2\alpha\epsilon + C_1\epsilon^2)\|u - w\|_V^2,
\end{aligned}$$

and hence by the Banach fixed point theorem the fixpoint (3.18) exists, which is equivalent to the existence of a solution to $\Phi(u) - \phi_L$ and thus to the variational problem (3.16). Uniqueness follows by (3.17). □

Existence and uniqueness for Poisson's equation

We now use the Lax-Milgram theorem to prove that (3.6) has a unique solution in $V = H_0^1(\Omega)$. First we show that the bilinear form is elliptic, using the Poincaré-Friedrich inequality,

$$\|v\|_1^2 = \|v\|^2 + \|\nabla v\|^2 \leq (1 + C)\|\nabla v\|^2,$$

so that the bilinear form is elliptic with $\alpha = 1/(1+C)$. By Cauchy-Schwarz inequality, the bilinear form is also continuous with $C_1 = 1$,

$$a(u, v) = (\nabla u, \nabla v) \leq \|\nabla u\|\|\nabla v\| \leq \|u\|_1\|v\|_1.$$

For $f \in L^2(\Omega)$ the linear form is continuous with $C_2 = \|f\|$,

$$L(v) = (f, v) \leq \|f\|\|v\| \leq \|f\|\|v\|_1,$$

and for $f \in H^{-1}(\Omega)$ with $C_2 = \|f\|_{-1}$,

$$L(v) = (f, v) \leq \|f\|_{-1}\|v\|_1.$$

Hence all conditions (i)-(iii) are satisfied and we conclude that the variational problem (3.6) has a unique solution $u \in V$.

Linear reaction-diffusion equation

Now consider the following linear *reaction-diffusion equation*,

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

The corresponding bilinear form,

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

is elliptic in $H_0^1(\Omega)$ with $\alpha = 1$, since $\|v\|_1^2 = a(v, v)$, and continuous with $C_1 = 1$ by Cauchy-Schwarz inequality,

$$a(u, v) = (\nabla u, \nabla v) + (u, v) \leq \|\nabla u\|\|\nabla v\| + \|u\|\|v\| \leq \|u\|_1\|v\|_1.$$

With $f \in H^{-1}(\Omega)$, the linear form is continuous with $C_2 = \|f\|_{-1}$, and hence the variational problem has a unique solution.

Regularity of solutions

We refer to the solution to the variational problem as a *weak solution*. Whether this weak solution is also a *strong solution* in the sense that the underlying partial differential equation is satisfied for all $x \in \Omega$, is the question of *regularity*.

The regularity of a weak solution depends on the regularity of the data and the domain of the problem. Specifically, a weak solution to an elliptic variational problem with smooth data and a smooth domain, is also smooth and thus satisfies the underlying partial differential equation pointwise.

Energy norm and stability of solutions

Partial differential equations rarely admit closed form solutions, but we can still infer some characteristics of the solutions from the weak form (3.16). For an elliptic variational problem, a symmetric bilinear form defines an inner product $(\cdot, \cdot)_E = a(\cdot, \cdot)$ on the Hilbert space V, with an associated *energy norm*

$$\|\cdot\|_E = a(\cdot, \cdot)^{1/2},$$

which is equivalent to the norm $(\cdot, \cdot)_V$, since

$$\alpha\|\cdot\|_V^2 \leq (\cdot, \cdot)_E \leq C_1\|\cdot\|_V^2.$$

For the energy norm we can derive the following stability estimate for the solution $u \in V$ to the variational problem (3.16),

$$\|u\|_E^2 = a(u,u) = L(u) \leq C_2\|u\|_V \leq (C_2/\alpha)\|u\|_E,$$

so that

$$\|u\|_E \leq (C_2/\alpha).$$

This illustrates a strategy for deriving stability estimates for general variational problems, to choose the test function equal to the solution of the variational problem. For the Poisson problem (3.6), we can derive the following stability estimate

$$\|\nabla u\| \leq C\|f\|,$$

which follows from

$$\|\nabla u\|^2 = (f,u) \leq \|f\|\|u\| \leq C\|f\|\|\nabla u\| \leq \frac{C^2}{2}\|f\|^2 + \frac{1}{2}\|\nabla u\|^2,$$

where we used Cauchy-Schwarz inequality, Poincaré-Friedrich inequality, and the following version of Young's inequality.

Theorem 6 (Young's inequality)**.** *For $a, b \geq 0$ and $\epsilon > 0$,*

$$ab \leq \frac{1}{2\epsilon}a^2 + \frac{\epsilon}{2}b^2.$$

Proof. $0 \leq (a - \epsilon b)^2 = a^2 + \epsilon^2 b^2 - 2ab\epsilon$ □

Optimality of Galerkin's method and Cea's lemma

In a Galerkin finite element method we seek an approximation $U \in V_h$,

$$a(U,v) = L(v), \quad \forall v \in V_h, \tag{3.19}$$

with $V_h \subset V$ a finite dimensional subspace, which in the case of a finite element method is a piecewise polynomial space. For an elliptic problem, existence and uniqueness of a solution follows from Lax-Milgram's theorem.

Since $V_h \subset V$, the weak form (3.16) is satisfied also for $v \in V_h$, and by subtracting (3.19) from (3.16) we obtain the Galerkin orthogonality property,

$$a(u - U, v) = 0, \quad \forall v \in V_h.$$

For an elliptic problem with symmetric bilinear form we can show that the Galerkin approximation is optimal in the energy norm, since

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

and hence

$$\|u-U\|_E \leq \|u-v\|_E, \quad \forall v \in V_h.$$

For an elliptic non-symmetric bilinear form, we can prove *Cea's lemma*,

$$\|u-U\|_V \leq \frac{C_1}{\alpha}\|u-v\|_V, \quad \forall v \in V, \tag{3.20}$$

which follows from

$$\begin{aligned} \|u-U\|_V^2 &\leq (1/\alpha)a(u-U,u-U) = (1/\alpha)a(u-U,u-v) \\ &\leq (C_1/\alpha)\|u-U\|_V \|u-v\|_V. \end{aligned}$$

A priori error estimation and interpolation

For a Galerkin finite element method the approximation space V_h consists of piecewise polynomial functions defined over a mesh that approximates the domain $\Omega \subset \mathbb{R}^n$.

Cea's lemma (3.20) provides an estimate of the Galerkin error in terms of an arbitrary function $v \in V_h$, which we can choose to be an interpolant of the exact solution $v = \mathcal{I}^h u$, with

$$\mathcal{I}^h : V \to V_h$$

an interpolation operator, from which we obtain the *a priori* error estimate

$$\|u-U\|_V \leq (C_1/\alpha)\|u-\mathcal{I}^h u\|_V,$$

only in terms of the exact solution to the variational problem.

For a simplicial mesh $\mathcal{T}^h = \{K\}$ with local mesh size $h = h|_K$, we define the local interpolation operator

$$\mathcal{I}_K u(x) = \sum_{i=0}^{n_q-1} \sigma_i(u)\lambda_i(x), \quad x \in K,$$

for $\lambda_i(x)$ the local Lagrange shape function on element K, and $\sigma_i(u) = u(x_i)$ the local degree of freedom. The global interpolation operator $\mathcal{I}^h$ is then defined by the local interpolation operator as

$$\mathcal{I}^h|_K = \mathcal{I}_K^h,$$

with possibly global continuity enforced by global coupling of certain local degrees of freedom. The interpolation error can be estimated as

$$\left(\sum_K \|v - \mathcal{I}^h v\|^p_{W^{s,p}(K)}\right)^{1/p} \leq Ch^{k-s}|v|_{W^{k,p}(\Omega)}, \quad \forall v \in W^{s,p}(\Omega),$$

where

$$|v|_{W^{k,p}(\Omega)} = \sum_{|\alpha|=k} \|D^\alpha u\|^p_{L^p(\Omega)},$$

is a *Sobolev semi-norm*. We also have the following *discrete trace inequality*

$$\|v\|_{L^p(\partial K)} \leq C\left(h_K^{-\frac{1}{p}}\|v\|_{L^p(K)} + h_K^{1-\frac{1}{p}}\|\nabla v\|_{L^p(K)}\right),$$

by which we can estimate the trace of any function $v \in W^{1,p}(K)$ on an element boundary ∂K, in terms of its value and gradient in the interior of the element K.

A posteriori error estimation

In contrast to an *a priori* error estimate which is expressed in terms of the unknown exact solution $u \in V$, an *a posteriori* error estimate is bounded in terms of a computed approximate solution $U \in V_h$. We define a bounded linear functional

$$M(\cdot) = (\cdot, \psi),$$

with ψ the Riesz representer of the functional $M \in V'$, guaranteed to exist by the Riesz representation theorem. To estimate the error with respect to $M(\cdot)$, we introduce an adjoint problem: find $\varphi \in V$, such that

$$a(v, \varphi) = M(v), \quad \forall v \in V. \tag{3.21}$$

An *a posteriori* error representation then follows from (3.16) and (3.21),

$$M(u) - M(U) = a(u, \varphi) - a(U, \varphi) = L(\varphi) - a(U, \varphi) = r(U, \varphi), \tag{3.22}$$

with the weak residual functional $r(U, \cdot) = L(\cdot) - a(U, \cdot) \in V'$, acting on the adjoint solution $\varphi \in V$,

$$r(U, \varphi) = L(\varphi) - a(U, \varphi).$$

Adaptive methods

With $U \in V_h$ a finite element approximation computed over a mesh $\mathcal{T}^h$, we can split the integral over the elements K in $\mathcal{T}^h$, so that the a posteriori error representation (3.22) is expressed as

$$M(u) - M(U) = r(U, \varphi) = \sum_{K \in \mathcal{T}^h} r(U, \varphi)|_K = \sum_{K \in \mathcal{T}^h} \mathcal{E}_K,$$

with the local error indicator

$$\mathcal{E}_K = r(U, \varphi)|_K,$$

defined for each element K. To approximate the error indicator we can compute an approximation $\Phi \approx \varphi$ to the adjoint problem (3.21), so that

$$\mathcal{E}_K \approx r(U, \Phi)|_K.$$

Such local error indicators can be used in an adaptive algorithm to identify which elements K that contribute most to the global error, to be selected for mesh refinement in an *h-adaptive method*, or for an increased order of the local polynomial approximation space P_K in a *p-adaptive method*. A combination of the two is referred to as an *hp-adaptive method*.

3.3 Uncertainty quantification

Propagation of uncertainties

The error estimates that we have derived for the approximation error only take into consideration errors from numerical discretization, assuming that the equation and the boundary conditions are exact. In practise this is often not the case, since model parameters, source terms or boundary conditions may rely on data which is uncertain.

Uncertainty may stem from a lack of knowledge, so called *epistemic uncertainties*, or from known statistical variations, referred to as *aleatoric uncertainties*. For aleatoric uncertainties we can employ a stochastic method such as Monte Carlo simulation, whereas for epistemic uncertainties we need to make assumptions regarding the probability distribution of the underlying uncertainties, carry out sensitivity studies with respect to the uncertain parameters, or account for worst case scenarios.

We will here focus on the *forward* propagation of uncertainties, that is, to quantify the effect of input model uncertainties on the model output of interest. The *inverse* problem of estimation and reduction of model uncertainties will be addressed in a later chapter.

Monte Carlo simulation

Consider the model problem

$$\begin{aligned} -\nabla \cdot (a(x,\xi)\nabla u) &= f(x), \quad x \in \Omega_x, \\ u|\partial\Omega_x &= 0, \end{aligned} \tag{3.23}$$

where the parameter $a : \Omega_x \times \Omega_\xi \to \mathbb{R}$ is a random field, with $\Omega_x \subset \mathbb{R}^n$ a spatial domain and Ω_ξ a probability sample space. That is, for each $x \in \Omega_x$ the function $a(x,\cdot) : \Omega_\xi \to \mathbb{R}$ is a random variable, and each observation is a deterministic function $a(\cdot,\xi) : \Omega_x \to \mathbb{R}$.

A Monte Carlo method for the solution of the model problem (3.23) would be to draw a random sample of N observations $\{a(\cdot,\xi_i)\}_{i=1}^N$, to use as parameters in N deterministic partial differential equations which we solve by standard numerical methods. This gives an ensemble of approximate solutions $\{U_i\}_{i=1}^N$ and corresponding sample functional outputs of interest $\{M(U_i)\}_{i=1}^N$, for which we compute statistics and estimate their distribution.

We have thereby estimated the propagated uncertainty in the input parameter to the uncertainty in the output of interest, by Monte Carlo simulation based on random sampling of the uncertain parameter. Monte Carlo simulation is a *non-intrusive method* in the sense that standard numerical methods, algorithms and software can be used as for the corresponding deterministic problem. To implement Monte Carlo simulation we only need to add the preprocessing step of random sampling of uncertain parameters, and the postprocessing step of estimating the probability distribution of the output of interest, for example, by constructing a histogram.

Multilevel Monte Carlo Method

In a Monte Carlo simulation it can be very expensive to solve one partial differential equation for each random sample. Therefore various methods have been developed to minimize the number of random samples needed to reach a certain level of accuracy in the approximation of a random variable X, corresponding to some output of interest.

One such method is the *Multilevel Monte Carlo Method* (MLMC) which is based on a hierarchy of mesh resolutions $l = 0, 1, ..., L$ and associated approximations $X^l \approx X$, using that we can express the expected value

$$E[X^L] = E[X^0] + \sum_{l=1}^{L} E[X^l - X^{l-1}],$$

and that the variance $Var[X^l - X^{l-1}]$ is smaller for finer meshes This implies that fewer random samples are needed on the finer meshes compared to the

coarse meshes in the hierarchy, which can mean a significantly cheaper method compared to standard Monte Carlo simulation.

Stochastic Galerkin method

An alternative to Monte Carlo simulation is the *Stochastic Galerkin method* which for the model problem (3.23) is based on a combination of Galerkin projection in $H_0^1(\Omega_x)$ and L^2 projection in the probability space $L^2(\Omega_\xi)$, defined as the space of all centered random variables with bounded variance.

We define the Hilbert space

$$V = L^2(\Omega_\xi; H_0^1(\Omega_x)) = \{v : \Omega_\xi \to H_0^1(\Omega_x) : E[\|v\|_1^2] < \infty\},$$

with inner product and norm

$$(u, v)_V = E[(\nabla u, \nabla v)], \quad \|u\|_V = (u, u)_V^{1/2}.$$

If for all $x \in \Omega_x$ and almost surely (in the sense of probabilities),

$$0 < a_{min} \leq a(x, \xi) \leq a_{max} < \infty,$$

then the bilinear form

$$a(u, v) = E[(a\nabla u, \nabla v)]$$

is coercive and bounded for all $u, v \in V$, and the linear form

$$L(v) = E[(f, v)]$$

is bounded for all $v \in V$. The Lax-Milgram theorem guarantees that the following stochastic variational problem has a unique solution: find $u \in V$ such that

$$a(u, v) = L(v), \quad \forall v \in V.$$

The stochastic Galerkin method is then to find $U \in V_h$ such that

$$a(U, v) = L(v), \quad \forall v \in V_h,$$

where $V_h \subset V$ is a finite dimensional subspace, constructed from finite dimensional subspaces of $H_0^1(\Omega_x)$ and $L^2(\Omega_\xi)$. The dimension of V_h is $N_x N_\xi$, with N_x and N_ξ the dimensions of the finite dimensional deterministic and stochastic subspaces, respectively. The stochastic Galerkin approximation exists uniquely by the Lax-Milgram theorem, and is optimal with respect to the norm in V due to a Galerkin orthogonality property. From the approximation $U \in V_h$ we can then compute the output functional of interest and various statistics.

The stochastic Galerkin method is an *intrusive method*, in the sense that the construction of the approximation space V_h in general requires modifications of the algorithms and software used for the deterministic problem.

Polynomial chaos expansions

Let $\{\psi_i(\xi)\}_{i=0}^{\{\infty}$ be an orthonormal basis for $L^2(\Omega_\xi)$, so that $E[\psi_i\psi_j] = \delta_{ij}$. Then any random variable $\eta \in L^2(\Omega_\xi)$ can be expressed as

$$\eta(\xi) = E[\eta] + \sum_{k=1}^{\infty} \eta_k \psi_k(\xi)$$

where $\eta_k = E[\eta\psi_k]$. Specifically, there exists orthonormal bases $\{\psi_i(\xi)\}_{i=0}^{\infty}$ in the form of orthogonal polynomials of random variables, with corresponding *polynomial chaos expansions*.

It follows that any random field $v \in V = L^2(\Omega_\xi; H_0^1(\Omega_x))$ can be expressed as

$$v(x,\xi) = E[v(x,\cdot)] + \sum_{k=1}^{\infty} v_k(x)\psi_k(\xi)$$

with expansion coefficients $v_k(x) = E[v(x,\cdot)\psi_k]$, and hence in the context of a stochastic Galerkin method a finite dimensional subspace $V_h \subset V$ can be constructed from a truncated polynomial chaos expansion together with a finite element basis.

Stochastic collocation method

If we choose a nodal basis $\{\psi_k\}_{k=0}^{N_\xi}$ for a set of nodes $\{\xi_k\}_{k=1}^{N_\xi} \subset \Omega_\xi$, the polynomial chaos expansion of a random field becomes

$$v(x,\xi) = E[v(x,\cdot)] + \sum_{k=1}^{N_\xi} v(x,\xi_k)\psi_k(\xi).$$

If we use such a nodal basis the stochastic Galerkin method reduces to a *stochastic collocation method*, corresponding to N_ξ decoupled deterministic problems to be solved, similar to Monte Carlo methods, but now the nodes $\{\xi_k\}_{k=1}^{N_\xi} \subset \Omega_\xi$ are not randomly sampled.

3.4 The discrete system

Finite element approximation spaces

Analogous to the boundary value problem in one dimension, we formulate a Galerkin finite element method to seek approximate solutions in the form of piecewise polynomial functions. In a continuous Galerkin finite element

method we use an approximation space of continuous piecewise polynomials, whereas in a discontinuous Galerkin finite element method we use a space of discontinuous piecewise polynomials, where the local shape functions are globally connected only through the variational form.

For example, consider $V_h^{(1)}(\Omega)$, the space of continuous piecewise linear functions over a tetrahedral mesh $\mathcal{T}^h$ in $\mathbb{R}^3$. For each tetrahedral element K, the local polynomial space P_K is expressed by the linear Lagrange nodal basis $\{\lambda_i\}_{i=0}^3$, and the set of local degrees of freedom Σ_K corresponds to function evaluation at the vertices of the tetrahedron K. The continuity of a function in $V_h^{(1)}$ is enforced by identifying all degrees of freedom associated with the same global vertex N_j in the mesh, to build a global basis $\{\phi_j\}$ associated to the set of global vertices $\mathcal{N} = \{N_j\}$.

The assembly algorithm

For a simplicial mesh $\mathcal{T}^h$, the global approximation space of continuous piecewise polynomial functions V_h is spanned by the global nodal basis $\{\phi_j\}$, where each basis function ϕ_j is associated to a global vertex N_j. Hence with Dirichlet boundary conditions the finite element approximation $U \in V_h$ can be expressed as

$$U(x) = \sum_{N_j \in \mathcal{N}_I} U(N_j)\phi_j(x) + \sum_{N_j \in \mathcal{N}_D} U(N_j)\phi_j(x),$$

with $\mathcal{N}_I$ all internal vertices in the mesh and $\mathcal{N}_D$ all vertices on the Dirichlet boundary, and where $U(N_j)$ is the node which corresponds to function evaluation at the vertex N_j.

The finite element method takes the form of a matrix problem

$$Ax = b, \tag{3.24}$$

where $a_{ij} = a(\phi_j, \phi_i)$, $x_j = U(N_j)$ and $b_i = L(\phi_i)$. To compute the Galerkin finite element approximation, we thus have to construct the matrix A and vector b, and then solve the resulting matrix problem (3.24) to obtain the nodal values $U(N_j)$.

The boundary nodal values are given by the interpolated Dirichlet boundary condition $U(N_j) = \mathcal{I}^h g_D(N_j)$, for all $N_j \in \mathcal{N}_D$. Hence, if we order the nodes so that the boundary nodes have the highest indices, the matrix problem (3.24) has a block structure,

$$\left[\begin{array}{c|c} A_{II} & A_{ID} \\ \hline 0_{DI} & I_{DD} \end{array}\right] \left[\begin{array}{c} x_I \\ \hline x_D \end{array}\right] = \left[\begin{array}{c} b_I \\ \hline b_D \end{array}\right],$$

where A_{II} is a square $n_I \times n_I$ matrix, with n_I the number of internal nodes, A_{ID} an $n_I \times n_D$ matrix, with n_D the number of boundary nodes, I_{DD} an $n_D \times n_D$ identity matrix, 0_{DI} an $n_D \times n_I$ zero matrix, and b_D is an n_D vector with components $(b_D)_j = \mathcal{I}^h g_D(N_j)$.

We can then choose to solve the equation (3.24) directly, or to eliminate the boundary nodes to obtain the reduced $n_I \times n_I$ system

$$A_{II} x_I = b_I - A_{ID} b_D,$$

which is a discrete analog of (3.10).

The matrix and vector are constructed by an assembly algorithm, which loops over all elements K in the mesh to compute the local element matrices $A^K = (a_{ij}^K)$, with

$$a_{i,j}^K = a(\lambda_j, \lambda_i)|_K,$$

and the local element vector

$$b_i^K = L(\lambda_i)|_K,$$

with $a(\cdot,\cdot)|_K$ and $L(\cdot)|_K$ the bilinear and linear forms restricted to element K, and with $\{\lambda_i\}_{i=1}^{n_q - 1}$ the element shape functions, for example, local Lagrange basis functions over K. The integrals are often approximated by quadrature over a reference element $\hat{K}$, based on a map $F_K : \hat{K} \to K$.

To add the local element matrix and element vector to the global matrix and vector, we use an index map

$$loc2glob : i_K \to i_A,$$

which maps the index of each local degree of freedom $i \in i_K$, to the corresponding index in the global matrix $loc2glob(i) \in i_A$.

In the case of Dirichlet boundary conditions, the rows in the matrix corresponding to boundary nodes $N_j \in \mathcal{N}_D$ are replaced by a row with one on the diagonal and with all other components zero. To enforce the Dirichlet boundary condition, each corresponding vector component is then set to the interpolated Dirichlet boundary value $b_j = \mathcal{I}^h g_D(N_j)$.

Matrix-free methods

The assembly algorithm can be a significant part of the computational cost, sometimes even dominating. Specifically, the insertion of element matrices into into the global matrix A can be expensive in a distributed parallel computation due to the cost of communication and data transfer.

Algorithm 1: Assembly of matrix $A = (a_{i,j})$ and vector $b = (b_i)$

for $K \in \mathcal{T}^h$ **do**
  for $i = 0, 1..., n_q - 1$ **do**
    $b_i^K = L(\lambda_i)|_K$ ▷ `compute element vector`
    $b_{loc2glob(i)}$ += b_i^K ▷ `add to global vector`
  end
  for $i = 0, 1..., n_q - 1$ **do**
    for $j = 0, 1, ..., n_q - 1$ **do**
      $a_{i,j}^K = a(\lambda_j, \lambda_i)|_K$ ▷ `compute element matrix`
      $a_{loc2glob(i),loc2glob(j)}$ += $a_{i,j}^K$ ▷ `add to global matrix`
    end
  end
end

Under such conditions a matrix-free method may be preferable, where global matrices are never assembled, instead the matrix-vector product Ax is implemented directly based on the local element matrices.

By the linearity of the matrix-vector product, we can assemble the local matrix-vector products $A^K x$ for each element K independently, and add each local contribution to the global vector Ax.

Algorithm 2: Assembly of matrix-vector product $Ax = (p_i)$

for $K \in \mathcal{T}^h$ **do**
  for $i = 0, 1..., n_q - 1$ **do**
    for $j = 0, 1, ..., n_q - 1$ **do**
      $a_{i,j}^K = a(\lambda_j, \lambda_i)|_K$ ▷ `compute element matrix`
      $x_j = x_{loc2glob(j)}$ ▷ `get global solution vector`
      $p_{loc2glob(i)}$ += $a_{i,j}^K \, x_j$ ▷ `add to global vector`
    end
  end
end

Domain decomposition methods

It may sometimes be beneficial to break down the discretization of (3.14) into a set of subproblems defined over a decomposition of the domain,

$$\Omega = \Omega_1 \cup ... \cup \Omega_N,$$

where each subdomain Ω_i has an associated mesh partition $\mathcal{T}_i^h$ with an approximation space $V_{h,i} \subset H^1(\Omega_i)$. Specifically, if the subproblems can be solved independently, the domain decomposition method can be used to design a parallel algorithm to solve (3.19).

Each subproblem takes the form: Find $U_i^{(k)} \in V_{h,i}$,

$$a_i(U_i^{(k)}, v) = L_i(v), \quad \forall v \in V_{h,i},$$

where the boundary conditions over the internal boundaries $\Gamma_{i,j}$ are given by the boundary operators $B_{i,j} : V_{h,i} \to \Gamma_{i,j}$, as

$$B_{i,j}(U_i^{(k)}) = B_{j,i}(U_j^{(k-1)}),$$

with k indicating a step in an iterative solution of the subproblems.

With overlapping domains and the boundary operators prescribing Dirichlet boundary conditions, we get a *Parallel Schwarz method*, but different choices of domain overlap and boundary conditions results in a range of different domain decomposition methods with their own characteristics.

Multigrid methods

The error in a Galerkin finite element method can be decomposed into different frequencies, with the highest frequency corresponding to a wavelength of the same order as the mesh size h. It is observed that when a simple stationary iterative method is applied to solve the discrete problem (3.19), the highest frequencies of the error are most efficiently reduced, and we thus refer to a this operation as *smoothing*.

The idea of a *multigrid method* is to apply smoothing on a hierarchy of meshes of different resolutions, to efficiently reduce all frequencies of the error. The following basic two-level algorithm can be applied recursively for a mesh hierarchy, where we denote by $\mathcal{T}^h$ the fine mesh and by $\mathcal{T}^H$ the coarse mesh, with associated approximation spaces V_h and V_H.

First consider $U_h \in V_h$, which satisfies

$$a(U_h, v) = L(v), \quad \forall v \in V_h,$$

which we approximate by smoothing using a stationary iterative method to get $\tilde{U}_h \approx U_h$. We then compute the residual

$$r_h = L(v_h) - a(\tilde{U}_h, v), \quad \forall v \in V_h,$$

which we project to $\mathcal{T}^H$ by a restriction operator $\mathcal{I}_h^H$, for example a weighted average of neighbouring nodes, to get

$$r_H = \mathcal{I}_h^H r_h.$$

An approximation of the error $e_H \in V_H$ then satisfies

$$a(e_H, v) = r_H, \quad \forall v \in V_H,$$

which we can approximate by a smoothing operation to get $\tilde{e}_H \approx e_H$, which we use to correct the approximation on the fine mesh $\mathcal{T}^h$, as

$$U_h = \tilde{U}_h + \mathcal{I}_H^h \tilde{e}_H,$$

where $\mathcal{I}_H^h$ typically is a linear interpolation operator.

3.5 Nonlinear boundary value problems

Nonlinear boundary value problems

Now consider the boundary value problem

$$A(u) = f, \tag{3.25}$$

in Ω, with suitable boundary conditions, and A a nonlinear differential operator acting on u.

To compute approximate solutions to (3.25) we use similar techniques as for nonlinear algebraic equations, based on a fixed point iteration

$$u^{(k)} = G(u^{(k-1)}) = u^{(k-1)} + B(f - A(u^{(k-1)})), \tag{3.26}$$

with G and B differential operators. The fixed point iteration converges if $G : X \to Y$ is a *contraction*, in other words, a *bounded operator* between the normed vector spaces X and Y, for which there exists an $M < 1$, such that

$$\|G(x) - G(y)\|_Y \leq M\|x - y\|_X, \quad \forall x, y \in X.$$

We denote by $A[w](\cdot)$ a linearization of the nonlinear differential operator $A(\cdot)$ at w, defined so that $A[u](u) = A(u)$. Then the iteration

$$A[u^{(k-1)}](u^{(k)}) = f,$$

corresponds to choosing $B = A[u^{(k-1)}]^{-1}$ in (3.26).

For example, a linearization of the nonlinear differential operator

$$A(u) = -\Delta u + u^2,$$

can take the form

$$A[w](\cdot) = (-\Delta + w)(\cdot),$$

so that $A[w](u) = -\Delta u + wu$, and $A[u](u) = -\Delta u + u^2$.

Newton's method and the Fréchet derivative

For two normed vector spaces X and Y with $\Omega \subset X$ an open subset, we say that a function $f : \Omega \to Y$ is *Fréchet differentiable* at $x \in \Omega$, if there exists a bounded linear operator $J : X \to Y$, such that

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) + J(h)\|_Y}{\|h\|_X} = 0,$$

where $Df(x;\cdot) = J(\cdot)$ is the Fréchet derivative at x. If X and Y are finite dimensional vector spaces, J coincides with the Jacobian matrix.

If the function f is Fréchet differentiable at all $x \in \Omega$ we say that $f \in C^1(\Omega)$, and

$$\|f(x_1) - f(x_2) - Df(x_0; x_1 - x_2)\|_Y \leq \epsilon \|x_1 - x_2\|_X,$$

where $x_1, x_2 \in \Omega$ such that $\|Df(x_i;\cdot) - Df(x_0;\cdot)\|_Y \leq \epsilon$, for $i = 1, 2$.

If we choose $B(\cdot) = Df^{-1}(x_0;\cdot)$ in (3.26), we obtain a Newton method where in each iteration we solve the system

$$Df^{-1}(u^{(k-1)}; u^k - u^{(k-1)}) = f - A(u^{(k-1)}).$$

Existence and uniqueness

There is no general theory for existence and uniqueness of solutions to nonlinear partial differential equations as for linear elliptic equations. In rare cases analytical solution formulas can be found, and for some specific equations one can use fixed point theorems to prove existence and uniqueness. If the solution can be characterized as a critical point, techniques to solve minimization problems can be used.

3.6 Exercises

Problem 6. *Derive the variational formulation (3.10), and formulate the finite element method.*

Problem 7. *Prove that (3.8) with V defined by (3.9) has a unique solution. Hint: Friedrich's inequality states that for all $u \in H^1(\Omega)$,*

$$\|u - \bar{u}\|_1 \leq C \|\nabla u\|,$$

with

$$\bar{u} = \frac{1}{|\Omega|} \int_\Omega u(x)\, dx.$$

Problem 8. *Derive the variational formulation and formulate a Galerkin finite element method, for the following boundary value problem.*

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

with $a > 0$.

Chapter 4

Evolution equations

We now turn to evolution equations, partial differential equations that evolve in time. Specifically we focus on two linear evolution equations, the heat equation and the wave equation, two models of dissipation and wave phenomena, respectively, used extensively in science and engineering.

To compute approximate solutions we use semi-discretization, where space and time are discretized independently. This allows us to reuse the methods we have already developed, by using time stepping combined with a finite element method.

4.1 The heat equation

The heat equation

Diffusion processes can be modelled by the heat equation,

$$\begin{aligned} \dot{u}(x,t) - \epsilon \Delta u(x,t) &= f(x,t), && (x,t) \in \Omega \times I, \\ u(x,t) &= 0, && (x,t) \in \partial\Omega \times I, \\ u(x,0) &= u_0(x), && x \in \Omega, \end{aligned} \tag{4.1}$$

for a diffusion coefficient $\epsilon > 0$ and a scalar function $u : \Omega \times I \to \mathbb{R}$, in the domain $\Omega \subset \mathbb{R}^n$ with boundary $\partial\Omega$, and with the time interval $I = (0, T]$. To find an approximate solution to the heat equation, we can use *semi-discretization* where space and time are discretized separately, using a finite element method and time stepping, respectively.

We give the equation for homogeneous Dirichlet boundary conditions but the extension to general boundary conditions is analogous to the Poisson equation.

Semi-discretization

For each $t \in I$, multiply the equation by a test function $v \in V = H_0^1(\Omega)$, and integrate in space over Ω to get the variational formulation,

$$\int_\Omega \dot{u}(x,t)v(x)\,dx + \epsilon \int_\Omega \nabla u(x,t) \cdot \nabla v(x)\,dx = \int_\Omega f(x,t)v(x)\,dx. \qquad (4.2)$$

We formulate a finite element method based on a piecewise polynomial space $V_h \subset V$, spanned by the finite element basis functions $\{\varphi_i\}_{i=1}^M$: for each $t \in I$ find $U(t) \in V_h$, such that

$$\int_\Omega \dot{U}(x,t)v(x)\,dx + \epsilon \int_\Omega \nabla U(x,t) \cdot \nabla v(x)\,dx = \int_\Omega f(x,t)v(x)\,dx, \qquad (4.3)$$

for all $v \in V_h$. The discretized equations form a system of initial value problems,

$$M\dot{U}(t) + SU(t) = b(t), \qquad (4.4)$$

with

$$m_{ij} = \int_\Omega \phi_j(x)\phi_i(x)\,dx,$$

$$s_{ij} = \epsilon \int_\Omega \nabla\phi_j(x) \cdot \nabla\phi_i(x)\,dx,$$

$$b_i(t) = \int_\Omega f(x,t)\phi_i(x)\,dx,$$

which is solved by time stepping for each $t = t_n$, to get the approximate solution

$$U(x,t_n) = U_n(x) = \sum_{j=1}^M U_{j,n}\phi_j(x).$$

Alternatively, we first discretize the equation in time and after that in space. For example, an implicit Euler discretization of (4.1) gives,

$$U_n(x) - k_n\Delta U_n(x) = U_{n-1}(x) + k_n f(x,t_n), \quad x \in \Omega,$$

for each time step (t_{n-1}, t_n). A finite element discretization is then used to compute the approximation $U_n(x)$.

Stability estimates

By selecting the test function $v = u$ in (4.2), we obtain

$$\int_\Omega \dot{u}(x,t)u(x)\,dx + \epsilon \int_\Omega \nabla u(x,t)\cdot\nabla u(x)\,dx = \int_\Omega f(x,t)u(x)\,dx,$$

which is the same as

$$\frac{1}{2}\frac{d}{dt}\|u\|^2 + \epsilon\|\nabla u\|^2 = (f,u) \le \|f\|\|u\| \le \frac{1}{2\epsilon}\|f\|^2 + \frac{\epsilon}{2}\|u\|^2,$$

by Cauchy-Schwarz inequality and Young's inequality, so that

$$\frac{d}{dt}\|u\|^2 + \epsilon\|\nabla u\|^2 \le \frac{1}{\epsilon}\|f\|^2,$$

or

$$\|u(T)\|^2 + \epsilon\int_0^T \|\nabla u\|^2\,dt \le \|u_0\|^2 + \int_0^T \frac{1}{\epsilon}\|f\|^2\,dt, \tag{4.5}$$

from which we find that the norm of the gradient ∇u is bounded by the data, and that with $f = 0$ the norm of the solution u decreases with time, which illustrates the dissipative nature of solutions to the heat equation.

With a suitable time stepping method for (4.3), corresponding stability estimates can be proven for the approximation $U \approx u$.

The scalar heat equation

The scalar heat equation has the same characteristics as the heat equation in $\mathbb{R}^n$ for $n > 1$, but with simplified differential and boundary operators,

$$\begin{aligned} \dot{u}(x,t) - \epsilon u''(x,t) &= f(x,t), && (x,t) \in (a,b)\times I,\\ u(x,t) &= 0, && x = a \text{ or } x = b, t \in I,\\ u(x,0) &= u_0(x), && x \in (a,b), \end{aligned} \tag{4.6}$$

for $u : [a,b]\times I \to \mathbb{R}$, with $I = (0,T]$. For each $t \in I$, the weak formulation reads: find $u(t) \in V = H_0^1([a,b])$, such that

$$\int_\Omega \dot{u}(x,t)v(x)\,dx + \epsilon\int_\Omega u'(x,t)v'(x)\,dx = \int_\Omega f(x,t)v(x)\,dx, \tag{4.7}$$

for all $v \in V$. This leads us to the semi-discretized Galerkin finite element problem: for each $t \in I$, find $U(t) \in V_h \subset V$, such that

$$\int_\Omega \dot{U}(x,t)v(x)\,dx + \epsilon\int_\Omega U'(x,t)v'(x)\,dx = \int_\Omega f(x,t)v(x)\,dx,$$

for all $v \in V_h$, which represents a system of initial value problems that we can solve with a time stepping method of our choice.

4.2 The wave equation

The wave equation

The wave equation in $\mathbb{R}^n$ is a fundamental model in physics, which for a scalar function $u : \Omega \times I \to \mathbb{R}$ and $I = (0, T]$, with $\Omega \subset \mathbb{R}^n$, takes the form

$$\begin{aligned} \frac{1}{c^2}\ddot{u} - \Delta u &= f, && (x,t) \in \Omega \times I, \\ u &= g_D, && (x,t) \in \Gamma_D \times I, \\ \nabla u \cdot n &= g_N, && (x,t) \in \Gamma_N \times I, \\ u(x,0) &= u_0(x), && x \in \Omega, \\ \dot{u}(x,0) &= \dot{u}_0(x), && x \in \Omega, \end{aligned} \tag{4.8}$$

where the boundary $\partial\Omega = \Gamma_D \cap \Gamma_N$, c is the wave speed, and $f : \Omega \times I \to \mathbb{R}$ is a forcing function.

Plane waves

In one space dimension the wave equation reduces to

$$\frac{1}{c^2}\ddot{u} - u'' = f, \tag{4.9}$$

with corresponding simplifications of the boundary conditions. For the unbounded domain $\mathbb{R}$, with $f = 0$, the most general solution to the wave equation is

$$u(x,t) = v(ct - x + a) + w(ct + x + b),$$

with and two $v, w \in C^2(\mathbb{R})$ and $a, b \in \mathbb{R}$. Specifically, we have *plane wave* solutions to (4.9), of the form

$$u(x,t) = A_0 \exp(i(kx - \omega t - \varphi)),$$

with the wave speed $c = \omega/k$, $\omega = 2\pi/T$ the angular frequency with period T, $k = 2\pi/\lambda$ the wave number with wavelength λ, φ a phase shift, and A_0 the amplitude.

In multiple dimensions $\mathbb{R}^n$, the homogeneous wave equation, with $f = 0$ in (4.8), also has plane wave solutions in unbounded domains, in a direction defined by a unit vector $d \in \mathbb{R}^n$,

$$u(x,t) = A_0 \exp(i(k \cdot d - \omega t - \varphi)),$$

where $k \in \mathbb{R}^n$ now is a wave vector with $|k| = 2\pi/\lambda$.

Variational formulation

For each $t \in I$ we have the following variational formulation of the wave equation, assuming homogeneous Dirichlet boundary conditions: find $u(t) \in V = H_0^1(\Omega)$, such that

$$\frac{1}{c^2}(\ddot{u}, v) + (\nabla u, \nabla v) = (f, v), \quad \forall v \in V.$$

For $f = 0$ we have energy conservation, since by setting $v = \dot{u}(t)$,

$$0 = \frac{1}{c^2}(\ddot{u}, \dot{u}) + (\nabla u, \nabla \dot{u}) = \frac{1}{c^2}\frac{1}{2}\frac{d}{dt}\|\dot{u}\|^2 + \frac{1}{2}\frac{d}{dt}\|\nabla u\|^2,$$

we find that the sum of kinetic and potential energy is conserved,

$$\frac{d}{dt}\left(\frac{1}{c^2}\|\dot{u}\|^2 + \|\nabla u\|^2\right) = 0.$$

Semi-discretization

To compute approximations to the wave equation semi-discretization is used, with a Galerkin finite element method in space together with a time stepping method.

With a constant time step length k, we approximate the second order time derivative by a finite difference method,

$$\ddot{u}(t_n) \approx \frac{\dot{u}(t_{n+1/2}) - \dot{u}(t_{n-1/2})}{k} \approx \frac{\frac{U_{n+1}-U_n}{k} - \frac{U_n - U_{n-1}}{k}}{k} = \frac{U_{n+1} - 2U_n + U_{n-1}}{k^2},$$

and similarly we can use the approximations

$$u(t_n) \approx \bar{U}_n = \frac{U_{n+1} + U_{n-1}}{2}, \quad f(t_n) \approx \bar{f}_n = \frac{f(t_{n+1}) + f(t_{n-1})}{2},$$

so that with U_n, U_{n-1} given, we can formulate the following discrete method: find $U_{n+1} \in V_h \subset V$, such that

$$\frac{1}{k^2c^2}(U_{n+1} - 2U_n + U_{n-1}, v) + (\nabla \bar{U}_n, \nabla v) = (\bar{f}_n, v), \quad \forall v \in V_h.$$

For $f = 0$ and with $v = U_{n+1} - U_{n-1}$, we get the discrete energy equality

$$\frac{2}{c^2}\|\dot{U}_{n+1}\|^2 + \|\nabla U_{n+1}\|^2 = \frac{2}{c^2}\|\dot{U}_n\|^2 + \|\nabla U_{n-1}\|^2,$$

with

$$\dot{U}_{n+1} = \frac{U_{n+1} - U_n}{k}.$$

Specifically, the discrete energy equality gives a bound on space and time derivatives in the semi-discretization method.

The wave equation in mixed form

Introducing a vector function $v : \Omega \times I \to \mathbb{R}^n$, we can write the wave equation in mixed form as a first order system,

$$\begin{aligned} c_1 \dot{u}_1 + \nabla \cdot u_2 &= f_1, && (x,t) \in \Omega \times I, \\ c_2 \dot{u}_2 + \nabla u_1 &= f_2, && (x,t) \in \Omega \times I, \\ u_1 &= g_D, && (x,t) \in \Gamma_D \times I, \\ u_2 \cdot n &= g_N, && (x,t) \in \Gamma_N \times I, \\ u_1(x,0) &= u_0(x), && x \in \Omega, \\ u_2(x,0) &= v_0(x), && x \in \Omega, \end{aligned} \tag{4.10}$$

with $c_1 > 0$ and $c_2 > 0$ coefficients such that $(c_1 c_2)^{-1} = c^2$, and with forcing functions $f_u : \Omega \times I \to \mathbb{R}$ and $f_v : \Omega \times I \to \mathbb{R}^n$, such that $c_2 \dot{f}_1 - \nabla \cdot f_2 = f$.

Variational formulation

If we assume homogeneous Dirichlet boundary conditions, for each $t \in I$ we have the weak form: find $(u_1(t), u_2(t)) \in V_1 \times V_2$, such that

$$\begin{aligned} c_1(\dot{u}_1, v_1) + (\nabla \cdot u_2, v_1) &= (f_1, v_1), \\ c_2(\dot{u}_2, v_2) + (\nabla u_1, v_2) &= (f_2, v_2), \end{aligned}$$

for all $(v_1, v_2) \in V_1 \times V_2$, with

$$\begin{aligned} V_1 &= H_0^1(\Omega), \\ V_2 &= H^1(\Omega). \end{aligned}$$

For $f_1 = f_2 = 0$, set $(v_1, v_2) = (u_1, u_2)$, which gives energy conservation

$$\frac{d}{dt}(c_1 \|u_1\|^2 + c_2 \|u_2\|^2) = 0.$$

Semi-discretization

We formulate the following semi-discrete formulation of the wave equation in mixed form: for each $t \in I$, find $(U_1, U_2) \in V_{1,h} \times V_{2,h}$, such that

$$\begin{aligned} c_1(\dot{U}_1, v_1) + (\nabla \cdot U_2, v_1) &= (f_1, v_1), \\ c_2(\dot{U}_2, v_2) + (\nabla U_1, v_2) &= (f_2, v_2), \end{aligned}$$

for all $(v_1, v_2) \in V_{1,h} \times V_{2,h}$, with $V_{1,h} \subset V_1$ and $V_{2,h} \subset V_2$.

For this semi-discrete formulation we can then choose a suitable time discretization method, for example, a trapezoidal method where

$$\dot{U}_i(t_n) \approx \dot{U}_{i,n} = \frac{U_{i,n} - U_{i,n-1}}{k_n},$$

and

$$U_i(t_n) \approx \bar{U}_{i,n} = \frac{U_{i,n+1} + U_{i,n-1}}{2}, \quad f_i(t_n) \approx \bar{f}_{i,n} = \frac{f_i(t_{n+1}) + f_i(t_{n-1})}{2},$$

for $i = 1, 2$. We then seek $(U_{1,n}, U_{2,n}) \in V_{1,h} \times V_{2,h}$, such that

$$\begin{aligned} c_1(\dot{U}_{1,n}, v_1) + (\nabla \cdot \bar{U}_{2,n}, v_1) &= (\bar{f}_{1,n}, v_1), \\ c_2(\dot{U}_{2,n}, v_2) + (\nabla \bar{U}_{1,n}, v_2) &= (\bar{f}_{2,n}, v_2), \end{aligned}$$

for all $(v_1, v_2) \in V_{1,h} \times V_{2,h}$.

Stabilization

In contrast to the discretization of the second order wave equation, for the mixed wave equation we can only prove that the solution is bounded, which follows from choosing $(v_1, v_2) = (\bar{U}_{1,n}, \bar{U}_{2,n})$, with $f_1 = f_2 = 0$, so that

$$c_1\|U_{1,n}\|^2 + c_2\|U_{2,n}\|^2 = c_1\|U_{1,n-1}\|^2 + c_2\|U_{2,n-1}\|^2.$$

But we have no bounds for the gradient of the solution. The consequence is that the approximation may oscillate in space, which can be addressed by using a stabilization technique.

We formulate the following stabilized method based on the residual of the equation: Find $(U_{1,n}, U_{2,n}) \in V_{1,h} \times V_{2,h}$, such that

$$\begin{aligned} c_1(\dot{U}_{1,n}, v_1) + (\nabla \cdot \bar{U}_{2,n}, v_1) + (\tau_1 R_{1,n}, \nabla \cdot v_2) &= (\bar{f}_{1,n}, v_1), \\ c_2(\dot{U}_{2,n}, v_2) + (\nabla \bar{U}_{1,n}, v_2) + (\tau_2 R_2, \nabla v_1) &= (\bar{f}_{2,n}, v_2), \end{aligned}$$

for all $(v_1, v_2) \in V_{1,h} \times V_{2,h}$, where

$$\begin{aligned} R_{1,n} &= \dot{U}_{1,n} + \nabla \cdot \bar{U}_{2,n} - \bar{f}_{1,n}, \\ R_{2,n} &= \dot{U}_{2,n} + \nabla \bar{U}_{1,n} - \bar{f}_{2,n}, \end{aligned}$$

are the residuals, and

$$\tau_1 = Ch\sqrt{\frac{c_2}{c_1}}, \quad \tau_2 = Ch\sqrt{\frac{c_1}{c_2}},$$

are stabilization parameters. For $f_1 = f_2 = 0$, with $(v_1, v_2) = (\bar{U}_{1,n}, \bar{U}_{2,n})$,

$$\begin{aligned}
&c_1\|U_{1,n}\|^2 + c_2\|U_{2,n}\|^2 + 2k(\|\sqrt{\tau_1}\nabla\cdot\bar{U}_{2,n}\|^2 + \|\sqrt{\tau_2}\nabla\bar{U}_{1,n}\|^2)\\
&+2((\tau_1 U_{1,n}, \nabla\cdot\bar{U}_{2,n}) + (\tau_2 U_{2,n}, \nabla\bar{U}_{1,n}))\\
&-(\tau_1 U_{1,n-1}, \nabla\cdot U_{2,n}) - (\tau_2 U_{2,n-1}, \nabla U_{1,n})\\
&= c_1\|U_{1,n-1}\|^2 + c_2\|U_{2,n-1}\|^2\\
&+(\tau_1 U_{1,n-1}, \nabla\cdot U_{2,n-1}) + (\tau_2 U_{2,n-1}, \nabla U_{1,n-1}).
\end{aligned}$$

If we now assume that $c_1 = c_2 = c^{-1}$, so that $\tau_1 = \tau_2 = \tau$, then by the homogeneous Dirichlet boundary conditions, with $k, l \in \{n, n-1\}$,

$$\begin{aligned}
&(\tau U_{1,k}, \nabla\cdot U_{2,l}) + (\tau U_{2,l}, \nabla U_{1,k})\\
&\qquad = \int_\Omega \nabla\cdot(\tau U_{1,k}U_{2,l})\,dx = \int_{\partial\Omega} \tau U_{1,k}(U_{2,l}\cdot n)\,ds = 0,
\end{aligned}$$

so that we have the stability estimate

$$\begin{aligned}
&\frac{1}{2kc}(\|U_{1,n}\|^2 + \|U_{2,n}\|^2) + \|\sqrt{\tau}\nabla\cdot\bar{U}_{2,n}\|^2 + \|\sqrt{\tau}\nabla\bar{U}_{1,n}\|^2\\
&\qquad = \frac{1}{2kc}(\|U_{1,n-1}\|^2 + \|U_{2,n-1}\|^2),
\end{aligned}$$

which illustrates the effect of the stabilization terms.

4.3 General evolution equations

Linear parabolic equations

Consider the partial differential equation

$$\dot{u} + L(u) = f, \quad (x,t) \in \Omega \times I,$$

together with initial and boundary conditions, where $u : \Omega \times I \to \mathbb{R}$ and $f : \Omega \times I \to \mathbb{R}$, and L is a linear differential operator of the form

$$L(u) = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x,t)\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{n} b_i(x,t)\frac{\partial u}{\partial x_i} + c(x,t)u,$$

with given coefficient functions a_{ij}, b_i, c, for $i, j = 1, ..., n$. If the bilinear form corresponding to $L(u)$ is elliptic, we say that the evolution equation is *parabolic*, in which case there exists a unique solution to the evolution equation if the data and domain are sufficiently regular.

Linearization of nonlinear evolution equations

Consider the nonlinear evolution equation

$$\dot{u} + A(u) = f, \quad (x,t) \in \Omega \times I,$$

together with initial and boundary conditions, where A is a nonlinear differential operator, which takes on scalar or vector values.

Since the methods we have developed are based on solving matrix equations where the matrix is generated from a linear operator, if A is nonlinear we may need to construct a linearized operator $A[w](\cdot)$ such that $A[u](u) = A(u)$.

For example, if $A(u) = (u \cdot \nabla)u$, we can construct the linearized operator $A[w](\cdot) = (w \cdot \nabla)(\cdot)$, such that $A[u](u) = (u \cdot \nabla)u$.

Time stepping nonlinear evolution equations

To advance a nonlinear evolution equation in time, for each time step $I_n = [t_{n-1}, t_n]$, we treat the nonlinear operator A differently depending on what time stepping method we use.

If we use an explicit time stepping method the nonlinear operator is evaluated at the previous time step, so that $A(u(t_n)) \approx A(U_{n-1})$.

In contrast, if we use an implicit time stepping method we need to solve a system of nonlinear equations at each time step. For example, in the case of the implicit Euler method, at each time step we need to solve the nonlinear equation

$$U_n = U_{n-1} + k_n(f_n - A[U_n]U_n),$$

for the new solution U_n, where $A[U_n]$ now is a matrix that corresponds to a spatial discretization of the linearized operator, and $k_n = t_n - t_{n-1}$ the time step length.

To solve the nonlinear equation we can use a fixed point iteration, or Newton's method, in which we need to reassemble the matrix $A[U_n^k]$ for each new approximation U_n^k in the nonlinear iteration.

Space-time finite element methods

In contrast to semi-discretization, we could choose to discretize the space-time continuum with one single finite element method. We refer to such finite element methods as *space-time finite element methods*, for which we seek a piecewise polynomial approximation in the space-time domain $\Omega \times I$.

But also with a space-time finite element method we may exploit the sequential nature of time by using a structured discretization in the time dimension to form space-time slabs $S_n = \Omega \times I_n$, with $I_n = (t_{n-1}, t_n)$. This way, each space-time slab S_n can be solved in sequence, in a form of time stepping method.

Operator splitting

If the nonlinear operator A naturally splits into separate operators that are different in character, we may employ individual methods tailored to each operator instead of using A directly. For example, if $A = A_1 + A_2$, with A_1 linear and A_2 nonlinear.

We distinguish between two splitting methods, simple splitting or symmetric (Strang) splitting. Simple splitting introduces a splitting error of order $\mathcal{O}(k_n^2)$ for each time step, so the global splitting error is first order accurate, whereas symmetric splitting is second order accurate. Thus we do not gain by using a higher order time stepping method for simple splitting, whereas to achieve second order accuracy with symmetric splitting we need a time stepping method that is second order accurate.

Simple splitting is based on computing an intermediate approximation $U_{n-1/2}$, so that in the case of using implicit Euler time stepping for A_1 and explicit Euler time stepping for A_2, we get

1. $\dfrac{U_{n-1/2} - U_{n-1}}{k_n} + A_1(U_{n-1/2}) = 0,$
2. $\dfrac{U_n - U_{n-1/2}}{k_n} + A_2(U_{n-1}) = f_{n-1},$

while for symmetric splitting with the trapezoidal rule,

1. $\dfrac{\tilde{U}_{n-1/2} - U_{n-1}}{k_n/2} + \dfrac{1}{2}(A_1(\tilde{U}_{n-1/2}) + A_1(U_{n-1})) = \dfrac{1}{2}(f(t_{n-1/2}) + f(t_{n-1})),$
2. $\dfrac{U_{n-1/2} - \tilde{U}_{n-1/2}}{k_n} + \dfrac{1}{2}(A_2(U_{n-1/2}) + A_2(\tilde{U}_{n-1/2})) = 0,$
3. $\dfrac{U_n - U_{n-1/2}}{k_n/2} + \dfrac{1}{2}(A_1(U_n) + A_1(U_{n-1/2})) = \dfrac{1}{2}(f(t_n) + f(t_{n-1/2})).$

We here considered the fully discretized case where A is a matrix, but for the for the case of semi-discretizaton where A is a differential operator, the idea is the same.

4.4 Exercises

Problem 9. *Derive (4.4) from the variational formulation of the heat equation.*

Problem 10. *Multiply (4.1) by $u(x,t)$ and integrate over Ω, to show that for $f(x,t) = 0$,*

$$\frac{d}{dt}\|u(t)\|^2 \leq 0.$$

Problem 11. *Derive (4.5) from (4.1).*

Problem 12. *Multiply (4.8) by $u(x,t)$ and integrate over Ω, to show that for $f(x,t) = 0$,*

$$\frac{d}{dt}(\|u(t)\|^2 + \|\nabla u(t)\|^2) = 0.$$

Chapter 5

Equations of fluid mechanics

The Navier-Stokes equations represent the fundamental model for fluid mechanics. Conservation of mass, momentum and energy, together with constitutive laws, is the foundation for the equations, which for different parameter regimes express viscous linear transport, wave propagation, or nonlinear transport phenomena, most notably turbulence and shocks. In this chapter we focus on incompressible flow, a realistic model for water, and air at low Mach numbers, which excludes shock waves and contact discontinuities.

The existence of weak solutions to the incompressible Navier-Stokes equations was established in 1934 by Jean Leray, and the question whether it exists a strong solution is formulated as a Clay Institute $1 million Prize problem. Ultimately this comes down to turbulence, if turbulent solutions are differentiable functions that satisfy the Navier-Stokes equations pointwise, or analogous to shock waves in compressible flow, weak solutions that dissipate energy.

5.1 Conservation laws

A general conservation law

Consider an arbitrary open subdomain $\omega \subset \mathbb{R}^n$. For a time $t > 0$, the total flow of a quantity with density $\phi(x,t)$ through the boundary $\partial\omega$ is given by

$$\int_{\partial\omega} \phi u \cdot n \, ds,$$

where n is the outward unit normal of $\partial\omega$, and $u = u(x,t)$ is the velocity of the flow. The change of the total quantity ϕ in ω is equal to the volume

source or sink $s = s(x,t)$, minus the total flow of the quantity through the boundary $\partial\omega$,

$$\frac{d}{dt}\int_\omega \phi(x,t)\,dx = -\int_{\partial\omega} \phi u \cdot n\,ds + \int_\omega s(x,t)\,dx,$$

a conservation equation which is sometimes referred to as *Reynolds transport theorem*. By Gauss' theorem,

$$\int_\omega \left(\frac{\partial}{\partial t}\phi(x,t) + \nabla\cdot(\phi u) - s\right)\,dx = 0,$$

and assuming the integrand is continuous in ω, we are lead to the general conservation equation

$$\dot\phi + \nabla\cdot(\phi u) - s = 0, \tag{5.1}$$

for $t > 0$ and $x \in \omega$, with $\omega \subset \mathbb{R}^n$ any open domain for which the equation is sufficiently regular.

Conservation of mass

Now consider the flow of a continuum with $\rho = \rho(x,t)$ the mass density of the continuum. The general continuity equation (5.1) with $\phi = \rho$ and zero source $s = 0$, gives the equation for conservation of mass

$$\dot\rho + \nabla\cdot(\rho u) = 0.$$

We say that a flow is *incompressible* if

$$\nabla\cdot u = 0,$$

or equivalently, if the *material derivative* is zero,

$$\frac{D\rho}{Dt} = \dot\rho + u\cdot\nabla\rho = 0,$$

since

$$0 = \dot\rho + \nabla\cdot(\rho u) = \frac{D\rho}{Dt} + \rho\nabla\cdot u.$$

Conservation of momentum

Newton's 2nd Law states that the change of *momentum* ρu over an arbitrary open subdomain $\omega \subset \mathbb{R}^n$, is equal to the sum of all forces, including *volume forces*,

$$\int_\omega \rho f\,dx,$$

for a force density $f = f(x,t) = (f_1(x,t), ..., f_n(x,t))$, and *surface forces*,

$$\int_{\partial\omega} n \cdot \sigma \, ds,$$

with the *Cauchy stress tensor* $\sigma = \sigma(x,t) = (\sigma_{ij}(x,t))$, and where we define $n \cdot \sigma = n^T\sigma = (\sigma_{ji}n_j)$. Gauss' theorem gives the total force as

$$\int_\omega \rho f \, dx + \int_{\partial\omega} n \cdot \sigma \, ds = \int_\omega (\rho f + \nabla \cdot \sigma) \, dx.$$

The general continuity equation with $\phi = \rho u$, and the source given by the sum of all forces, leads to the equation for conservation of momentum

$$\frac{\partial}{\partial t}(\rho u) + \nabla \cdot (\rho u \otimes u) = \rho f + \nabla \cdot \sigma, \tag{5.2}$$

with $u \otimes u = uu^T$, the tensor product of the velocity vector field u. With the help of conservation of mass, we can rewrite the left hand side as

$$\frac{\partial}{\partial t}(\rho u) + \nabla \cdot (\rho u \otimes u) = u(\dot\rho + \nabla \cdot (\rho u)) + \rho(\dot u + (u \cdot \nabla)u) = \rho(\dot u + (u \cdot \nabla)u),$$

so that

$$\rho(\dot u + (u \cdot \nabla)u) = \rho f + \nabla \cdot \sigma. \tag{5.3}$$

We say that (5.2) is an equation on *conservation form*, whereas (5.3) is on *non-conservation form*.

The Cauchy stress tensor

The Cauchy stress tensor can be represented by a symmetric 3×3 matrix σ, and thus has three real eigenvalues λ_i, the *principal stresses*, and three orthogonal eigenvectors v_i, the *principal directions*. In the basis of the principal directions, the Cauchy stress tensor is a diagonal matrix with the principal stresses on the diagonal. The largest normal stress is equal to the largest principal stress $\max_i \lambda_i$, and acts on the plane defined by the corresponding principal direction as normal vector.

The principal stresses can be combined to form the *stress invariants*,

$$\begin{aligned} I_1 &= \lambda_1 + \lambda_2 + \lambda_3, \\ I_2 &= \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1, \\ I_3 &= \lambda_1\lambda_2\lambda_3. \end{aligned}$$

Specific states of stress can be characterized through the principal stresses. For example, $\lambda_2 = \lambda_3 = 0$ corresponds to a state of a uniaxial stress λ_1 in the direction of v_1, and $\lambda_1 = \lambda_2 = \lambda_3$ to a state of pure *pressure*.

With the principal stresses, we can also formulate a scalar measure of stress, the *von Mises stress*

$$\sigma_n = \sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2}{2}}.$$

We define the *mechanical pressure* as the mean normal stress,

$$p_{mech} = -\frac{1}{3}\operatorname{tr}(\sigma) = -\frac{1}{3}I_1,$$

and the *deviatoric stress tensor* $\tau = \sigma + p_{mech}I$, with $\operatorname{tr}(\tau) = 0$, such that

$$\sigma = -p_{mech}I + \tau,$$

and we can write conservation of momentum as

$$\rho(\dot{u} + (u \cdot \nabla)u) = \rho f - \nabla p_{mech} + \nabla \cdot \tau.$$

Conservation of energy

The first law of thermodynamics states that the total energy should be conserved, where the total energy $e = k + \theta$ is the sum of kinetic energy

$$k = \rho\frac{|u|^2}{2},$$

and internal heat energy θ which can be related to temperature and the thermodynamic pressure. To close a system of conservation laws, an additional equation is needed to determine the internal energy θ, an equation of state such as the perfect gas law.

5.2 The Navier-Stokes equations

Compressible flow

The *Mach number* is defined by

$$M = |u|/c,$$

and relates the local speed $|u|$ to the speed of sound c. For high Mach numbers the flow is compressible, in which case shock waves, rarefactions and contact discontinuities can develop, whereas for low Mach numbers the flow can be modelled as incompressible.

Incompressible flow

Now consider incompressible flow, and assume the density and viscosity to be constant. In this case the conservation laws for mass and momentum can be solved as one coupled system, where the pressure is an independent variable, different from the thermodynamic pressure used in a model of compressible flow.

To determine the deviatoric stress we need a constitutive model of the fluid. For a Newtonian fluid, the deviatoric stress depends linearly on the *strain rate tensor*

$$\epsilon = \frac{1}{2}(\nabla u + (\nabla u)^T) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right),$$

with $\tau = 2\mu\epsilon$, where μ is the *dynamic viscosity*, which we here assume to be constant.

The *incompressible Navier-Stokes equations* then takes the form,

$$\dot{u} + (u \cdot \nabla)u + \nabla p - \nu\Delta u = f, \tag{5.4}$$

$$\nabla \cdot u = 0, \tag{5.5}$$

with the *kinematic viscosity* $\nu = \mu/\rho$, and the kinematic pressure $p = p_{mech}/\rho$.

Non-dimensionalization

Solutions to the Navier-Stokes equations may have quite different characters, depending on the balance of the inertial and dissipative terms of the equations. To exhibit this balance, we express the Navier-Stokes equations in terms of the non-dimensional variables u_*, p_*, f_*, x_*, t_*,

$$u = Uu_*, \quad p = Pp_*, \quad x = Lx_*, \quad f = Ff_*, \quad t = Tt_*,$$

where U, P, L, T are characteristic scales of the velocity, pressure, force, length and time, respectively. The resulting non-dimensionalized differential operators are scaled as,

$$\frac{\partial}{\partial t} = \frac{1}{T}\frac{\partial}{\partial t_*}, \quad \nabla = \frac{1}{L}\nabla_*, \quad \Delta = \frac{1}{L^2}\Delta_*,$$

which gives

$$\frac{U}{T}\frac{\partial}{\partial t_*}u_* + \frac{U^2}{L}(u_* \cdot \nabla_*)u_* + \frac{P}{L}\nabla_* p_* - \frac{\nu U}{L^2}\Delta_* u_* = Ff_*,$$

$$\frac{U}{L}\nabla \cdot u_* = 0,$$

or,

$$\begin{aligned} &\dot{u} + (u \cdot \nabla)u + \nabla p - Re^{-1}\Delta u = f, \\ &\nabla \cdot u = 0. \end{aligned}$$

Here we have dropped the non-dimensional notation for simplicity, with

$$T = L/U, \quad P = U^2, \quad F = \frac{U^2}{L}, \quad Re = \frac{UL}{\nu},$$

where the *Reynolds number* Re determines the balance between inertial and viscous characteristics in the flow. For low Re linear viscous effects dominate, whereas for high Re we have a flow dominated by nonlinear inertial effect, and turbulence for sufficiently high Reynolds numbers.

Formally, in the limit $Re \to \infty$, the viscous term vanishes and we are left with the inviscid *Euler equations*,

$$\begin{aligned} &\dot{u} + (u \cdot \nabla)u + \nabla p = f, \\ &\nabla \cdot u = 0, \end{aligned}$$

traditionally seen as a model for flow at high Reynolds numbers. Although, this simple analysis is too naive and relies on strong assumptions on the regularity of solutions to the equations, which is an open problem posed as one of the Clay Institute $1 million Prize problems.

In the limit $Re \to 0$, we obtain the *Stokes equations* as a model of viscous flow,

$$\begin{aligned} -\Delta u + \nabla p &= f, \\ \nabla \cdot u &= 0, \end{aligned}$$

with now a different scaling of the pressure and the force,

$$P = \frac{\nu U}{L}, \quad F = \frac{\nu U}{L^2}.$$

5.3 Stokes flow

The Stokes equations

The Stokes equations for a domain $\Omega \subset \mathbb{R}^n$ with boundary $\nabla\Omega = \Gamma_D \cup \Gamma_N$, and associated normal n, takes the form

$$\begin{aligned} -\Delta u + \nabla p &= f, && x \in \Omega, \\ \nabla \cdot u &= 0, && x \in \Omega, \\ u &= g_D, && x \in \Gamma_D, \\ -\nabla u \cdot n + pn &= g_N, && x \in \Gamma_N. \end{aligned}$$

Homogeneous Dirichlet boundary conditions

First assume that $\partial\Omega = \Gamma_D$ and $g_D = 0$, that is, homogeneous Dirichlet boundary conditions for the velocity. We then seek a weak solution to the Stokes equations in the following spaces,

$$V = H_0^1(\Omega) \times ... \times H_0^1(\Omega) = [H_0^1(\Omega)]^n,$$
$$Q = \{q \in L^2(\Omega) : \int_\Omega q\, dx = 0\},$$

where the extra condition in the vector space Q is needed to assure uniqueness of the pressure, which otherwise is undetermined up to a constant.

We derive the variational formulation by taking the inner product of the momentum equation with a test function $v \in V$, and the inner product of the continuity equation with a test function $q \in Q$. By Green's formula and the homogeneous Dirichlet boundary condition, we obtain the variational formulation as: find $(u, p) \in V \times Q$, such that

$$a(u, v) + b(v, p) = (f, v), \quad \forall v \in V, \tag{5.6}$$
$$-b(u, q) = 0, \quad \forall q \in Q, \tag{5.7}$$

for $f \in V'$, with

$$a(v, w) = (\nabla v, \nabla w) = \int_\Omega \nabla v : \nabla w\, dx, \tag{5.8}$$
$$b(v, q) = -(\nabla \cdot v, q) = -\int_\Omega (\nabla \cdot v) q\, dx, \tag{5.9}$$

and

$$\nabla v : \nabla w = \sum_{i,j=1}^{3} \frac{\partial v_i}{\partial x_j} \frac{\partial w_i}{\partial x_j}.$$

Existence and uniqueness of solutions

To prove the existence of a unique solution to the Stokes problem (5.6)-(5.7), we will use the Lax-Milgram theorem. Specifically, we denote by

$$T : [H^{-1}(\Omega)]^n \to [H_0^1(\Omega)]^n = V,$$

the solution operator for the Poisson equation with homogeneous Dirichlet boundary conditions, which exists by Lax-Milgram theorem and which we extend to a vector operator for each component individually, defined by

$$a(Tf, v) = (f, v), \quad \forall v \in V.$$

First set $v = T\nabla q$ in (5.6), to get

$$a(T\nabla p, T\nabla q) = (f, T\nabla q), \quad \forall q \in Q, \tag{5.10}$$

since $a(u, T\nabla q) = (u, \nabla q) = b(u, q) = 0$, by (5.7), and

$$b(T\nabla q, p) = -(\nabla \cdot T\nabla q, p) = (\nabla q, T\nabla p) = a(T\nabla q, T\nabla p).$$

To prove that there exists a unique solution $p \in Q$ to (5.10), we need to show that the bilinear form is elliptic, and because

$$a(T\nabla q, T\nabla q)^{1/2} = \frac{a(T\nabla q, T\nabla q)}{a(T\nabla q, T\nabla q)^{1/2}} = \sup_{v\in V} \frac{a(T\nabla q, v)}{a(v,v)^{1/2}} = \sup_{v\in V} \frac{b(v,q)}{\|v\|_V},$$

we are lead to the following condition for $\beta > 0$,

$$\sup_{v\in V} \frac{b(v,q)}{\|v\|_V} \geq \beta \|q\|_Q, \quad \forall q \in Q,$$

or equivalently,

$$\inf_{q\in Q} \sup_{v\in V} \frac{b(v,q)}{\|v\|_V \|q\|_Q} \geq \beta > 0,$$

which we refer to as the *inf-sup condition*.

If this inf-sup condition is satisfied there exists a $p \in Q$ which is the unique solution to (5.10), so that we can solve the momentum equation (5.6) to get the unique solution $u \in V$ for which (5.7) is satisfied, which exists by the Lax-Milgram theorem if the bilinear form is elliptic.

These conditions on the bilinear forms guarantees a unique solution for the general abstract variational problem (5.6)-(5.7).

Theorem 7. *The variational problem (5.6)-(5.7) has a unique weak solution $(u, p) \in V \times Q$, which satisfies the following stability inequality,*

$$\|u\|_V + \|q\|_Q \leq C\|f\|_{-1},$$

if the following conditions hold,

(i) $a(\cdot,\cdot)$ is bounded and coercive, i.e. that exists a constant $\alpha > 0$,

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

for all $v \in Z = \{v \in V : b(v,q) = 0, \forall q \in Q\}$,

(ii) $b(\cdot,\cdot)$ is bounded and satisfies the inf-sup condition, i.e. there exists a constant $\beta > 0$,

$$\inf_{q\in Q} \sup_{v\in V} \frac{b(v,q)}{\|v\|_V \|q\|_Q} \geq \beta.$$

Proof of stability estimate. Set $(v, q) = (u, p)$ then add (5.6) to (5.7). By (i) there exists a constant $\alpha > 0$,

$$\|u\|_V \leq \frac{\alpha^{-1}}{\|u\|_V} a(u,u) = \frac{\alpha^{-1}}{\|u\|_V}(f,u) \leq \alpha^{-1}\|f\|_{-1}.$$

Hence, by (ii), for all $q \in Q$ there exists a constant $\beta > 0$, such that

$$\begin{aligned} \|q\|_Q &\leq \beta^{-1} \sup_{v\in V} \frac{|b(v,q)|}{\|v\|_V} = \beta^{-1} \sup_{v\in V} \frac{|(f,v) - a(u,v)|}{\|v\|_V} \\ &\leq \beta^{-1}(\|f\|_{-1} + \|u\|_V) \leq \beta^{-1}(1+\alpha^{-1})\|f\|_{-1}, \end{aligned}$$

and thus

$$\|u\|_V + \|q\|_Q \leq (\alpha^{-1} + \beta^{-1} + \alpha^{-1}\beta^{-1})\|f\|_{-1} = C\|f\|_{-1}.$$

□

Mixed finite element approximation

We now formulate a finite element method for solving Stokes equations. Since we use different approximation spaces for the velocity and the pressure, we refer to the method as a mixed finite element method.

We seek an approximation $(U, P) \in V_h \times Q_h$, such that,

$$a(U,v) + b(v,P) = (f,v), \tag{5.11}$$

$$-b(U,q) = 0, \tag{5.12}$$

for all $(v, q) \in V_h \times Q_h$, where V_h and Q_h are finite element approximation spaces. There exists a unique solution to (5.11)-(5.12), under similar conditions as for the continuous variational problem.

Theorem 8. *The mixed finite element problem (5.11)-(5.12) has a unique solution $(U, P) \in V_h \times Q_h$, if*

(i) $a(\cdot,\cdot)$ is coercive, i.e. that exists a constant $\alpha_h > 0$, such that

$$a(v,v) \geq \alpha_h \|v\|_V,$$

for all $v \in Z_h = \{v \in V_h : b(v,q) = 0, \forall q \in Q_h\}$,

(ii) $b(\cdot,\cdot)$ satisfies the inf-sup condition, i.e. there exists a constant $\beta_h > 0$,

$$\inf_{q\in Q_h} \sup_{v\in V_h} \frac{b(v,q)}{\|v\|_V \|q\|_Q} \geq \beta_h,$$

and this unique solution satisfies the following error estimate,

$$\|u - U\|_V + \|p - P\|_Q \leq C \left(\inf_{v \in V_h} \|u - v\| + \inf_{q \in Q_h} \|p - q\| \right),$$

for a constant $C > 0$.

The pair of approximation spaces must be chosen to satisfy the inf-sup condition, with the velocity space sufficiently rich compared to the pressure space. For example, continuous piecewise quadratic approximation of the velocity and continuous piecewise linear approximation of the pressure, referred to as the Taylor-Hood elements. On the other hand, continuous piecewise linear approximation of both velocity and pressure is not inf-sup stable.

Schur complement methods

We seek finite element approximations in the following spaces,

$$V_h = \{v = (v_1, v_2, v_3) : v_k(x) = \sum_{j=1}^{N} v_k^j \phi_j(x), k = 1, 2, 3\}$$

and

$$Q_h = \{q : q(x) = \sum_{j=1}^{M} q^j \psi_j(x)\},$$

which leads to a discrete system in matrix form,

$$\begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix},$$

with u and p vectors holding the coordinates of U and P in the respective bases of V_h and Q_h.

The matrix A is symmetric positive definite and thus invertible, so we can express

$$u = A^{-1}(f - Bp),$$

and since $B^T u = 0$,

$$B^T A^{-1} B p = B^T A^{-1} f,$$

which is the *Schur complement* equation. If $\text{null}(B) = \{0\}$, then the matrix $S = B^T A^{-1} B$ is symmetric positive definite and can also be inverted.

Schur complement methods take the form

$$p_k = p_{k-1} - C^{-1}(B^T A^{-1} B p_{k-1} - B^T A^{-1} f),$$

where C^{-1} is a preconditioner for $S = B^T A^{-1} B$. The Usawa algorithm is based on C^{-1} as a scaled identity matrix, which gives

1. Solve $Au_k = f - Bp_{k-1}$,
2. Set $p_k = p_{k-1} + \alpha B^T u_k$.

Stabilized methods

Approximation spaces of equal order is possible, by stabilization of the standard Galerkin finite element method: find $(U, P) \in V_h \times Q_h$, such that,

$$\begin{aligned} a(U, v) + b(v, P) &= (f, v), \\ -b(U, q) + s(P, q) &= 0, \end{aligned}$$

for all $(v, q) \in V_h \times Q_h$, where $s(P, q)$ is a pressure stabilization term. The resulting discrete system takes the form,

$$\begin{bmatrix} A & B \\ B^T & S \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix},$$

where the stabilization term is chosen so that the matrix S is invertible. For example, the *Brezzi-Pitkäranta* stabilization takes the form,

$$s(P, q) = C \int_\Omega h^2 \nabla P \cdot \nabla q \, dx,$$

with $C > 0$ a constant.

5.4 The transient Navier-Stokes equations

Semi-discretization

We now formulate a finite element method for solving the unsteady Navier-Stokes equations (5.4)-(5.5) by semi-discretization.

For each $t > 0$, we seek approximations $(U(t), P(t)) \in V_h \times Q_h$, with $U(t) = (U_1(t), U_2(t), U_3(t))$, of the form,

$$U_k(x, t) = \sum_{j=1}^{N} U_k^j(t) \phi_j(x), \; k = 1, 2, 3, \quad P(x, t) = \sum_{j=1}^{M} P^j(t) \psi_j(x),$$

such that

$$(\dot{U}, v) + c(U; U, v) + a(U, v) + b(v, P) - b(U, q) = (f, v),$$

for all $(v, q) \in V_h \times Q_h$, where the bilinear forms are defined by (5.8)-(5.9), with the trilinear form,

$$c(u; v, w) = ((u \cdot \nabla)v, w) = \int_\Omega (u \cdot \nabla)v \cdot w \, dx. \tag{5.13}$$

The semi-discretization (5.4) is a system of ODEs, which we can solve by a suitable time-stepping method. For $\nabla \cdot u = 0$, we have that

$$c(u; v, w) = \bar{c}(u; v, w),$$

with

$$\bar{c}(u; v, w) = \frac{1}{2}((u \cdot \nabla)v, w) - \frac{1}{2}(v, (u \cdot \nabla)w). \tag{5.14}$$

We may alternatively use this form in (5.4), where in particular we note that $c(u; v, v) = 0$.

The θ-method

Semi-discretizaton by the θ-method takes the form: for each time interval $I_n = (t_{n-1}, t_n)$, with the time step length $k_n = t_n - t_{n-1}$, find $(U_n, P_n) = (U(t_n), P(t_n)) \in V_h \times Q_h$, such that

$$\frac{1}{k_n}((U_n, v) - (U_{n-1}, v)) + c(U_\theta; U_\theta, v) + \nu a(U_\theta, v) + b(v, P_\theta) - b(U_\theta, q) = (f, v),$$

for all $(v, q) \in V_h \times Q_h$, with

$$U_\theta = (1 - \theta)U_n + \theta U_{n-1}, \quad P_\theta = (1 - \theta)P_n + \theta P_{n-1}.$$

Here e.g. $\theta = 0$ corresponds to the Implicit Euler method, and $\theta = 0.5$ corresponds to the Trapezoidal method.

5.5 Stabilized finite element methods

Stabilization techniques

Previously we have found that the inf-sup condition of the Stokes equations could be circumvented by a pressure stabilization technique to allow for equal order approximation spaces for the velocity and the pressure. The idea of stabilization through a small regularizing perturbation is fundamental for finite element methods, not only in the context of saddle-point problems.

Whereas low Reynolds number flow is dominated by viscosity, for high Reynolds numbers the dominating phenomenon is nonlinear transport. Standard Galerkin finite element methods are not optimal for discretization of transport dominated equations, instead we will use stabilization techniques to formulate suitable finite element methods.

Stabilized finite element methods for the Navier-Stokes equations is an active area of research, with different classes of techniques being developed, many of them related. The methods presented in this chapter are chosen as examples of how we can address the two main challenges for constant density incompressible flow, the saddle-point problem and instabilities associated with the transport dominated problem. Alternative stabilization techniques are based on controlling the jump in gradients, methods derived from local multiscale arguments, or localized artifical viscosity, for example.

Linear transport model problem

We first consider the linear transport equation for a scalar quantity $u = u(x,t)$, convected by a divergence-free vector field $\beta = \beta(x,t)$,

$$\begin{aligned} \dot{u} + (\beta \cdot \nabla)u - \epsilon \Delta u = f, \qquad & (x,t) \in \Omega \times I, \\ \nabla \cdot \beta = 0, \qquad & (x,t) \in \Omega \times I, \end{aligned}$$

with suitable initial and boundary conditions, and $\epsilon > 0$ a small diffusion coefficient. To understand the basic mechanism we analyze the following simple model problem in one space dimension,

$$\begin{aligned} -\epsilon u'' + u' = 0, \quad & x \in (0,1), \\ u(0) = 1,\, u(1) = 0, & \end{aligned}$$

for which we formulate a standard Galerkin finite element method: find $U \in V_h$ such that,

$$\int_0^1 \epsilon u' v' \, dx + \int_0^1 u' v \, dx = 0,$$

for all test functions $v \in V_h^0$, with

$$\begin{aligned} V_h &= \{v \in H^1(0,1) : v(0) = 1, v(1) = 0\}, \\ V_h^0 &= \{v \in H^1(0,1) : v(0) = 0, v(1) = 0\}. \end{aligned}$$

Divide the interval $(0,1)$ into M uniform subintervals $I_i = (x_{i-1}, x_i)$ of length $h = x_i - x_{i-1}$, with nodes $\{x_i\}_{i=0}^{M+1}$ and associated piecewise linear basis functions $\phi_i = \phi_i(x)$.

Then we can write the finite element approximation as

$$U(x) = \sum_{j=1}^{M} u_j \phi_j(x) + u_0 \phi_0(x) + u_{M+1} \phi_{M+1}(x),$$

with $u_j = u(x_j)$ (since we have a nodal basis), and from the boundary conditions we have that

$$U(x) = \sum_{j=1}^{M} u_j \phi_j(x) + \phi_0(x).$$

The discrete system takes the form $Ax = b$, with $A = (a_{ij})$, $b = (b_i)$ and $x = (x_j)$,

$$a_{ij} = \int_0^1 \epsilon \phi_j'(x) \phi_i'(x)\ dx + \int_0^1 \phi_j'(x) \phi_i(x)\ dx,$$
$$b_i = \int_0^1 \epsilon \phi_0'(x) \phi_i'(x)\ dx + \int_0^1 \phi_0'(x) \phi_i(x)\ dx.$$

Equation i takes the form

$$\sum_{j=1}^{M} a_{ij} x_j = x_{i-1} \left(-\frac{\epsilon}{h} - \frac{1}{2} \right) + x_i \frac{2\epsilon}{h} + x_{i+1} \left(-\frac{\epsilon}{h} + \frac{1}{2} \right) = 0.$$

We observe two different regimes,

$$\epsilon >> h \Rightarrow -x_{i-1} + 2x_i - x_{i+1} = 0,$$
$$\epsilon << h \Rightarrow -x_{i-1} + x_{i+1} = 0,$$

with a combination of the two when $\epsilon \approx h$. In the convection dominated case, the boundary conditions lead to two cases depending on if M is an odd or even number; either no solution exists, or the solution oscillates between 0 and 1.

To obtain a finite element approximation that is close to the exact solution in the convection dominated case, we stabilize the method by an artificial diffusion $\epsilon = h/2$. We also refer to this as an *upwind method*, since the resulting equation takes the form

$$-x_{i-1} + x_i = 0,$$

where information is propagated from the upwind direction.

Streamline diffusion stabilization

For the Navier-Stokes equations we can use artificial viscosity to stabilize the finite element method for high Reynolds numbers. But there are also more accurate, less diffusive, stabilization methods.

For example, we may use streamline diffusion stabilization, where we add artificial viscosity in the streamline direction β only, which is enough to remove the spurious oscillations. Combined with pressure stabilization to allow for equal order approximation spaces, the method takes the following form.

For each $t > 0$, find $(U(t), P(t)) \in V_h \times Q_h$, such that

$$(\dot{U}, v) + \bar{c}(U; U, v) + a(U, v) + b(v, P) - b(U, q) + s_1(U; U, v) + s_2(P, q) = (f, v),$$

for all $(v, q) \in V_h \times Q_h$, with the stabilization terms

$$\begin{aligned} s_1(U; U, v) &= (\delta_1(U \cdot \nabla)U, (U \cdot \nabla)v), \\ s_2(P, q) &= (\delta_2 \nabla P, \nabla q), \end{aligned}$$

with stabilization parameters $\delta_1 \sim h/U_{n-1}$ and $\delta_2 \sim h$.

By choosing $(v, q) = (U, P)$, we obtain a stability estimate of the method,

$$\frac{d}{dt}\frac{1}{2}\|U\|^2 + \|\sqrt{\nu}\nabla U\|^2 + \|\sqrt{\delta_1}(U \cdot \nabla)U\|^2 + \|\sqrt{\delta_2}\nabla P\|^2 = 0,$$

where we can observe the regularizing effect of the stabilization terms.

Least squares stabilization of the residual

The Galerkin Least Squares (GLS) method is based on a combination of Galerkin's method with a least squares minimization of the residual of the Navier-Stokes equations. GLS is a *consistent* method in the sense that all terms in the method are based on the residual of the equations, no artificial stabilization terms are added.

For each $t > 0$, find $(U(t), P(t)) \in V_h \times Q_h$, such that

$$(\dot{U}, v) + \bar{c}(U; U, v) + a(U, v) + b(v, P) - b(U, q) + s_1(U; U, v) + s_2(U, v) = (f, v),$$

for all $(v, q) \in V_h \times Q_h$, with the stabilization terms

$$\begin{aligned} s_1(w; U, v) &= (\delta_1(\dot{U} + (w \cdot \nabla)U + \nabla P), \dot{v} + (w \cdot \nabla)v + \nabla q) \\ s_2(U, v) &= (\delta_2 \nabla \cdot U, \nabla \cdot v), \end{aligned}$$

with stabilization parameters $\delta_1 \sim h/U_{n-1}$ and $\delta_2 \sim hU_{n-1}$.

By choosing $(v, q) = (U, P)$, we obtain a stability estimate of the method,

$$\frac{d}{dt}\frac{1}{2}\|U\|^2+\|\sqrt{\nu}\nabla U\|^2+\|\sqrt{\delta_1}(\dot{U}+(U\cdot\nabla)U+\nabla P-\nu\Delta U)\|^2+\|\sqrt{\delta_2}\nabla\cdot U\|^2 = 0,$$

where we can observe the regularizing effect of the stabilization terms.

5.6 Computational fluid dynamics

CFD discretization methods

Simulation of fluid dynamics by computational methods is referred to as *computational fluid dynamics* (CFD), and this chapter is only but a small sample of this field.

The analysis of finite element methods closely follows the mathematical analysis of the variational form of the Navier-Stokes equations, which is an advantage in the development of stabilization methods and error analysis of the methods. But other types of discretization methods are also used. For example, Galerkin methods with global orthonormal basis functions, referred to as *spectral methods*, can be used in periodic domains, and for simple domains which can be well approximated by a structured grid, finite difference methods can be used.

Finite volume methods are today the most widely used class of CFD discretization methods in industry, in part by tradition but also since the methods enjoy some advantageous properties. A finite volume method is based on local conservation over mesh cells that each represent a control volume, and the flow between cells is determined by flux functions. Hence the methods by design are locally conservative. Mathematically, there is a strong relation between finite volume methods and discontinuous Galerkin finite element methods (DGFEM). In fact, a low order DGFEM is equivalent to a finite volume method.

Particle methods in the form of *smooth particle hydrodynamics* are popular in the computer graphics community due to their speed, but the accuracy is typically too low to be used for quantitative predictions in CFD. *Lattice Boltzmann methods* is another particle based approach which has gained some popularity in the CFD community for similar reasons, but accuracy is also here a concern.

Most CFD discretization methods fall in the categories of mesh based methods, such as finite element or finite volume methods, meshfree methods, such as particle methods, or integral methods based on fundamental solutions of the equations.

Turbulence and a $1 million prize problem

Turbulence is an outstanding challenge to computational fluid dynamics, since the computational cost to fully resolve the turbulent scales of the flow is often unsurmountable, even with the most powerful supercomputers available. The consequence is that a whole subfield of computational fluid dynamics is devoted to the development of turbulence models, that seek to model the effect of the smallest turbulent scales without full resolution of the flow.

The question of existence of unique solutions to the Navier-Stokes equations in three dimensions is open, in fact it is formulated as one of the Clay Institute $1 million prize problems. The main mathematical result available is due to Jean Leray from 1934 who proved the existence of a weak solution, or turbulent solutions in his own terminology.

Uniqueness of such weak solutions is an open problem, but can be approached by studying functionals of the weak solutions in the form of mean values, such as time averaged forces from the headwind experienced by a car. In such a mathematical framework, simulation of turbulent flow is based on computational approximation of weak solutions to the Navier-Stokes equations, for example, by a finite element method.

5.7 Exercises

Problem 13. *Implement a finite element method to solve the Stokes equations.*

Problem 14. *Implement a stablized finite element method to solve the time-dependent Navier-Stokes equations.*