Numerical solution of partial differential equations

Endre Süli

Mathematical Institute, University of Oxford,
Radcliffe Observatory Quarter,
Woodstock Road, Oxford OX2 6GG, UK

1 Introduction

Numerical solution of PDEs is rich and active field of modern applied mathematics. The steady growth of the subject is stimulated by ever-increasing demands from the natural sciences, engineering and economics to provide accurate and reliable approximations to mathematical models involving partial differential equations (PDEs) whose exact solutions are either too complicated to determine in closed form or, in many cases, are not known to exist. While the history of numerical solution of ordinary differential equations is firmly rooted in 18th and 19th century mathematics, the mathematical foundations of the field of numerical solution of PDEs are much more recent: they were first formulated in the landmark paper *Über die partiellen Differenzengleichungen der mathematischen Physik* (On the partial difference equations of mathematical physics) by Richard Courant, Karl Friedrichs, and Hans Lewy, published in 1928. There is a vast array of powerful numerical techniques for specific PDEs: level set and fast-marching methods for front-tracking and interface problems; numerical methods for PDEs on, possibly evolving, manifolds; immersed boundary methods; mesh-free methods; particle methods; vortex methods; various numerical homogenization methods and specialized numerical techniques for multiscale problems; wavelet-based multiresolution methods; sparse finite difference/finite element methods, greedy algorithms and tensorial methods for high-dimensional PDEs; domain-decomposition methods for geometrically complex problems, and numerical methods for PDEs with stochastic coefficients that feature in a number of applications, including uncertainty quantification problems. Our brief review cannot do justice to this huge and rapidly evolving subject. We shall therefore confine ourselves to the most standard and well-established techniques for the numerical solution of PDEs: finite difference methods, finite element methods, finite volume methods and spectral methods. Before embarking on our survey, it is appropriate to take a brief excursion into the theory of PDEs in order to fix the relevant notational conventions and to describe some typical model problems.

2 Model partial differential equations

A linear partial differential operator L of *order* m with real-valued coefficients $a_\alpha = a_\alpha(x)$, $|\alpha| \le m$, on a domain $\Omega \subset \mathbb{R}^d$, defined by

$$L := \sum_{|\alpha| \le m} a_\alpha(x)\partial^\alpha, \quad x \in \Omega,$$

is called *elliptic* if, for every $x := (x_1, \dots, x_d) \in \Omega$ and every nonzero $\xi := (\xi_1, \dots, \xi_d) \in \mathbb{R}^d$,

$$Q_m(x,\xi) := \sum_{|\alpha|=m} a_\alpha(x)\xi^\alpha \neq 0.$$

Here $\alpha := (\alpha_1, \dots, \alpha_d)$ is a d-component vector with nonnegative integer entries, called a *multi-index*, $|\alpha| := \alpha_1 + \cdots + \alpha_d$ is the *length* of the multi-index α, $\partial^\alpha := \partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d}$, with $\partial_{x_j} := \partial/\partial x_j$, and $\xi^\alpha := \xi_1^{\alpha_1} \cdots \xi_d^{\alpha_d}$. In the case of complex-valued coefficients a_α the definition above is modified by demanding that $|Q_m(x,\xi)| \neq 0$ for all $x \in \Omega$ and all nonzero $\xi \in \mathbb{R}^d$. A typical example of a first-order elliptic operator with complex coefficients is the *Cauchy–Riemann operator* $\partial_{\bar z} := \frac{1}{2}\big(\partial_x + \imath\partial_y\big)$, where $\imath := \sqrt{-1}$. With this general definition of ellipticity even-order operators can exhibit some rather disturbing properties. For example, the Bitsadze equation $\partial_{xx}u + 2\imath\partial_{xy}u - \partial_{yy}u = 0$ admits infinitely many solutions in the unit disc Ω in $\mathbb{R}^2$ centered at the origin, all of which vanish on the boundary $\partial\Omega$ of Ω. Indeed, with $z = x + \imath y$, $u(x,y) = (1 - |z|^2)f(z)$ is a solution that vanishes on $\partial\Omega$ for any complex analytic function f. Thus a stronger requirement, referred to as *uniform ellipticity*, is frequently imposed; for real-valued coefficients a_α, $|\alpha| \le m$, and $m = 2k$ where k is a positive integer, uniform ellipticity demands the existence of a constant $C > 0$ such that $(-1)^k Q_{2k}(x,\xi) \ge C|\xi|^{2k}$ for all $x \in \Omega$ and all nonzero $\xi \in \mathbb{R}^d$.

The archetypal linear second-order uniformly elliptic PDE is $-\Delta u + c(x)u = f(x)$, $x \in \Omega$. Here c and f are real-valued functions defined on Ω and $\Delta := \sum_{i=1}^{d} \partial_{x_i}^2$ is the *Laplace operator.* When $c < 0$ the equation is called the *Helmholtz equation.* In the special case when $c(x) \equiv 0$ the equation is referred to as *Poisson's equation*, and when $c(x) \equiv 0$ and $f(x) \equiv 0$ as *Laplace's equation.* Elliptic PDEs arise in a range of mathematical models in continuum mechanics, physics, chemistry, biology, economics and finance. For example, in a two-dimensional flow of an incompressible fluid with flow-velocity $u = (u_1, u_2, 0)$ the stream-function ψ, related to u by $u = \nabla \times (0, 0, \psi)$, satisfies Laplace's equation. The potential Φ of a gravitational field, due to an attracting massive object of density ρ, satisfies Poisson's equation $\Delta\Phi = 4\pi G\rho$, where G is the universal gravitational constant.

More generally, one can consider fully nonlinear second-order PDEs:

$$F(x, u, \nabla u, D^2 u) = 0,$$

where F is a real-valued function defined on the set $\Upsilon := \Omega \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d\times d}_{\text{symm}}$, with a typical element $\upsilon := (x, z, p, R)$, where $x \in \Omega$, $z \in \mathbb{R}$, $p \in \mathbb{R}^d$ and $R \in \mathbb{R}^{d\times d}_{\text{symm}}$, Ω is an open set in $\mathbb{R}^d$, D^2u denotes the Hessian matrix of u, and $\mathbb{R}^{d\times d}_{\text{symm}}$ is the $d(d+1)/2$-dimensional linear space of real symmetric $d \times d$ matrices, $d \geq 2$. An equation of this form is said to be elliptic on Υ if the $d \times d$ matrix whose entries are $\partial F/\partial R_{ij}$, $i, j = 1, \ldots, d$, is positive definite at each $\upsilon \in \Upsilon$. An important example, encountered in connection with optimal transportation problems, is the Monge–Ampère equation: $\det D^2u = f(x)$ with $x \in \Omega$; for the equation to be elliptic it is necessary to demand that the twice continuously differentiable function u is uniformly convex at each point of Ω, and for such a solution to exist we must also have f positive.

Parabolic and hyperbolic PDEs typically arise in mathematical models where one of the independent physical variables is time, t. For example,

$$\partial_t u + Lu = f \quad \text{and} \quad \partial_{tt} u + Lu = f,$$

where L is a uniformly elliptic partial differential operator of order $2m$ and u and f are functions of $(t, x_1, \ldots, x_d)$, are *uniformly parabolic* and *uniformly hyperbolic* PDEs, respectively. The simplest examples are the (uniformly parabolic) unsteady heat equation and the (uniformly hyperbolic) second-order wave equation, where

$$Lu := -\sum_{i,j=1}^{d} \partial_{x_j} \left(a_{ij}(t, x)\partial_{x_i} u\right),$$

and $a_{ij}(t, x) = a_{ij}(t, x_1, \ldots, x_d)$, $i, j = 1, \ldots, d$, are the entries of a $d \times d$ matrix, which is positive definite, uniformly with respect to $(t, x_1, \ldots, x_d)$.

Not all PDEs are of a certain fixed type. For example, the following PDEs are *mixed elliptic-hyperbolic*; they are elliptic for $x > 0$ and hyperbolic for $x < 0$:

$$\begin{aligned}
\partial_{xx} + \text{sign}(x)\partial_{yy} u = 0 & \qquad \text{(Lavrentiev equation)},\\
\partial_{xx} u + x\partial_{yy} u = 0 & \qquad \text{(Tricomi equation)},\\
x\partial_{xx} + \partial_{yy} u = 0 & \qquad \text{(Kel'dish equation)}.
\end{aligned}$$

Stochastic analysis is a fertile source of PDEs of *nonnegative characteristic form*, such as

$$\partial_t u - \sum_{i,j=1}^{d} \partial_{x_j} \left(a_{ij}\partial_{x_i} u\right) + \sum_{i=1}^{d} b_i \, \partial_{x_i} u + cu = f,$$

where b_i, c and f are real-valued functions of $(t, x_1, \ldots, x_d)$, and $a_{ij} = a_{ij}(t, x_1, \ldots, x_d)$, $i, j = 1, \ldots, d$, are the entries of a *positive semidefinite* matrix; since the a_{ij} are dependent on the temporal variable t, the equation is, potentially, of *changing type.* An important special case is when the a_{ij} are all identically equal to zero, resulting in the first-order hyperbolic equation, also referred to as *advection* (or *transport*) *equation*:

$$\partial_t u + \sum_{i=1}^{d} b_i(t, x) \, \partial_{x_i} u + c(t, x)u = f(t, x).$$

The nonlinear counterpart of this equation,

$$\partial_t u + \sum_{i=1}^{d} \partial_{x_i} [f(t, x, u)] = 0,$$

plays an important role in compressible fluid dynamics, traffic flow models and flow in porous media. Special cases include the Burgers equation $\partial_t u + \partial_x(\frac{1}{2}u^2) = 0$ and the Buckley–Leverett equation $\partial_t u + \partial_x(u^2/(u^2 + \frac{1}{4}(1-u)^2)) = 0$.

PDEs are rarely considered in isolation: additional information is typically supplied in the form of boundary conditions, imposed on the boundary $\partial\Omega$ of the domain $\Omega \subset \mathbb{R}^d$ in which the PDE is studied, or, in the case of parabolic and hyperbolic equations, also as initial conditions at $t = 0$. The PDE in tandem with the boundary/initial conditions is referred to as a *boundary-value problem/initial-value problem*, or when both boundary and initial data are supplied, as an *initial-boundary-value problem.*

3 Finite difference methods

We begin by considering finite difference methods for elliptic boundary-value problems. The basic idea behind the construction of finite difference methods is to *discretize* the closure, $\overline{\Omega}$, of the (bounded) domain of definition $\Omega \subset \mathbb{R}^d$ of the solution (the, so-called, *analytical solution*) to the PDE by approximating it with a finite set of points in $\mathbb{R}^d$, called the *mesh points* or *grid points*, and replacing the partial derivatives of the analytical solution appearing in the equation by *divided differences* (difference quotients) of a *grid-function*, i.e. a function that is defined at all points of the finite difference grid. The process results in a finite set of equations with a finite number of unknowns: the values of the grid-function representing the finite difference approximation to the analytical solution over the finite difference grid. We illustrate the construction by considering a simple second-order uniformly elliptic PDE subject to a *homogeneous Dirichlet boundary condition*:

$$
\begin{aligned}
-\Delta u + c(x,y)u &= f(x,y) \qquad \text{in } \Omega, && (1)\\
u &= 0 \qquad \text{on } \partial\Omega, && (2)
\end{aligned}
$$

on the unit square $\Omega := (0,1)^2$; here c and f are real-valued functions that are defined and continuous on Ω, and $c \geq 0$ on Ω. Let us suppose for simplicity that the grid-points are equally spaced. Thus we take $h := 1/N$, where $N \geq 2$ is an integer. The corresponding finite difference grid is then $\overline{\Omega}_h := \{(x_i, y_j) : i,j = 0,\dots,N\}$, where $x_i := ih$ and $y_j := jh$, $i,j = 0,\dots,N$. We also define $\Omega_h := \overline{\Omega}_h \cap \Omega$ and $\partial\Omega_h := \overline{\Omega}_h \setminus \Omega_h$.

It is helpful to introduce the following notation for *first-order divided differences*:

$$D_x^+ u(x_i,y_j) := \frac{u(x_{i+1},y_j) - u(x_i,y_j)}{h}$$

and

$$D_x^- u(x_i,y_j) := \frac{u(x_i,y_j) - u(x_{i-1},y_j)}{h},$$

with $D_y^+ u(x_i,y_j)$ and $D_y^-(x_i,y_j)$ defined analogously. Then, $D_x^2 u(x_i,y_j) := D_x^- D_x^+ u(x_i,y_j)$ and $D_y^2 u(x_i,y_j) := D_y^- D_y^+ u(x_i,y_j)$ are referred to as the *second-order divided difference* of u in the x- and y-direction, respectively, at $(x_i,y_j) \in \Omega_h$.

Assuming that $u \in C^4(\overline{\Omega})$ (i.e. that u and all of its partial derivatives up to and including those of fourth order are defined and continuous on $\overline{\Omega}$), we have that, at any $(x_i,y_j) \in \Omega_h$,

$$D_x^2 u(x_i,y_j) = \frac{\partial^2 u}{\partial x^2}(x_i,y_j) + \mathcal{O}(h^2) \qquad (3)$$

and

$$D_y^2 u(x_i,y_j) = \frac{\partial^2 u}{\partial y^2}(x_i,y_j) + \mathcal{O}(h^2), \qquad (4)$$

as $h \to 0$. Omission of the $\mathcal{O}(h^2)$ terms in (3) and (4) above yields that

$$D_x^2 u(x_i,y_j) \approx \tfrac{\partial^2 u}{\partial x^2}(x_i,y_j),\ D_y^2 u(x_i,y_j) \approx \tfrac{\partial^2 u}{\partial y^2}(x_i,y_j),$$

where the symbol $\approx$ signifies approximate equality in the sense that as $h \to 0$ the expression to the left of $\approx$ converges to the expression to the right of $\approx$. Hence,

$$
\begin{aligned}
-\big(D_x^2 u(x_i,y_j) + D_y^2 u(x_i,y_j)\big) &+ c(x_i,y_j)u(x_i,y_j)\\
\approx f(x_i,y_j) &\quad \text{for all } (x_i,y_j) \in \Omega_h, && (5)\\
u(x_i,y_j) = 0 &\quad \text{for all } (x_i,y_j) \text{ in } \partial\Omega_h. && (6)
\end{aligned}
$$

It is instructive to note the similarity between (1) and (5), and (2) and (6), respectively. Motivated by the form of (5) and (6), we seek a grid-function U, whose value at the grid-point $(x_i,y_j) \in \overline{\Omega}_h$, denoted by U_{ij}, approximates $u(x_i,y_j)$, the unknown exact solution to the boundary-value problem (1), (2) evaluated at (x_i,y_j), $i,j = 0,\dots,N$. We define U as the solution to the following system of linear algebraic equations:

$$
\begin{aligned}
-\,(D_x^2 U_{ij} + D_y^2 U_{ij}) &+ c(x_i,y_j)U_{ij}\\
= f(x_i,y_j) &\quad \text{for all } (x_i,y_j) \in \Omega_h, && (7)\\
U_{ij} = 0 &\quad \text{for all } (x_i,y_j) \in \partial\Omega_h. && (8)
\end{aligned}
$$

As each equation in (7) involves five values of the grid-function U (namely, U_{ij}, $U_{i-1,j}$, $U_{i+1,j}$, $U_{i,j-1}$, $U_{i,j+1}$), the finite difference method (7) is called the *five-point difference scheme.* The matrix of the linear system (7), (8) is sparse, symmetric and positive definite, and for given functions c and f it can be efficiently solved by iterative techniques from [REF **??**]IV.xy, including [REF **??**]II.xy type methods (e.g. the conjugate gradient method) and multigrid methods. Multigrid methods were developed in the 1970s and 1980 and are widely used as the iterative solver of choice for large systems of linear algebraic equations that arise from finite difference and finite element approximations in many industrial applications. The fundamental idea behind multigrid methods is to accelerate the convergence of standard relaxation methods (such as the Jacobi iteration or successive over-relaxation (SOR)) by using a hierarchy of coarser-to-finer grids.

A multigrid method with an intentionally reduced convergence tolerance can also be used as an efficient *preconditioner* for a Krylov subspace iteration. The preconditioner P for a nonsingular matrix A is an approximation of A^{-1}, whose purpose is to ensure that PA is a good approximation of the identity matrix, and therefore iterative algorithms for the solution of the preconditioned version, $PAx = Pb$, of the system of linear algebraic equations $Ax = b$ exhibit rapid convergence.

One of the central questions in the numerical analysis of PDEs is the mathematical study of the approximation properties of numerical methods. We shall illustrate this by considering the finite difference method (7), (8). The grid-function T defined on Ω_h by

$$\begin{aligned} T_{ij} := &- \left(D_x^2 u(x_i, y_j) + D_y^2 u(x_i, y_j)\right) \\ &+ c(x_i, y_j)u(x_i, y_j) - f(x_i, y_j) \end{aligned} \tag{9}$$

is called the *truncation error* of the finite difference method (7), (8). Assuming that $u \in C^4(\overline{\Omega})$, it follows from (3)–(5) that, at each grid point $(x_i, y_j) \in \Omega_h$, $T_{ij} = \mathcal{O}(h^2)$ as $h \to 0$. The exponent of h in the statement $T_{ij} = \mathcal{O}(h^2)$ (which, in this case, is equal to 2) is called the *order of accuracy* (or *order of consistency*) of the method.

It can be shown that there exists a positive constant c_0, independent of h, U and f, such that

$$\begin{aligned} \Bigg(h^2 \sum_{i=1}^{N} \sum_{j=1}^{N-1} |D_x^- U_{ij}|^2 + h^2 \sum_{i=1}^{N-1} \sum_{j=1}^{N} |D_y^- U_{ij}|^2 \\ + h^2 \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} |U_{ij}|^2 \Bigg)^{\frac{1}{2}} \\ \le c_0 \left(h^2 \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} |f(x_i, y_j)|^2 \right)^{\frac{1}{2}}. \end{aligned} \tag{10}$$

Such an inequality, expressing the fact that the numerical solution $U \in S_{h,0}$, is bounded by the data (in this case $f \in S_h$), uniformly with respect to the grid size h, where $S_{h,0}$ denotes the linear space of all grid-functions defined on $\overline{\Omega}_h$ that vanish on $\partial\Omega_h$ and S_h is the linear space of all grid functions defined on Ω_h, is called a *stability inequality.* The smallest real number $c_0 > 0$ for which (10) holds is called the *stability constant* of the method. It follows in particular from (10) that if $f_{ij} = 0$ for all $i, j = 1, \ldots, N-1$, then $U_{ij} = 0$ for all $i, j = 0, \ldots, N$. Therefore the matrix of the system of linear equations (7), (8) is nonsingular, which then implies the existence of a unique solution U to (7), (8) for any $h = 1/N$, $N \ge 2$. Consider the difference operator $L_h : U \in S_{h,0} \mapsto f = L_h U \in S_h$ defined by (7), (8). The left-hand side of (10) is sometimes denoted by $\|U\|_{1,h}$ and the right-hand side by $\|f\|_{0,h}$; hence, the stability inequality (10) can be rewritten as

$$\|U\|_{1,h} \le c_0 \|f\|_{0,h}$$

with $f = L_h U$, and stability can then be seen to be demanding the existence of the inverse to the linear finite difference operator $L_h : S_{h,0} \to S_h$, and its boundedness, uniformly with respect to the discretization parameter h. The mapping $U \in S_{h,0} \mapsto \|U\|_{1,h} \in \mathbb{R}$ is a norm on $S_{h,0}$, called the *discrete (Sobolev)* $H^1(\Omega)$ *norm*, and the mapping $f \in S_h \mapsto \|f\|_{0,h} \in \mathbb{R}$ is a norm on S_h, called the *discrete* $L^2(\Omega)$ *norm.* It should be noted that the stability properties of finite difference methods depend on the choice of norm for the data and for the associated solution.

In order to quantify the closeness of the approximate solution U to the analytical solution u at the grid-points, we define the *global error* e

of the method (7), (8) by $e_{ij} := u(x_i, y_j) - U_{ij}$. Clearly, the grid-function $e = u - U$ satisfies (7), (8) if $f(x_i, y_j)$ on the right-hand side of (7) is replaced by T_{ij}. Hence, by the stability inequality, $\|u - U\|_{1,h} = \|e\|_{1,h} \leq c_0\|T\|_{0,h}$. Under the assumption that $u \in C^4(\overline{\Omega})$ we thus deduce that $\|u - U\|_{1,h} \leq c_1 h^2$, where c_1 is a positive constant, independent of h. The exponent of h on the right-hand side (which is 2 is this case) is referred to as the *order of convergence* of the finite difference method and is equal to the order of accuracy. Indeed, the fundamental idea that stability and consistency together imply convergence is a recurring theme in the analysis of numerical methods for differential equations.

The five-point difference scheme can be generalized in various ways. For example, instead of using the same grid-size h in both co-ordinate directions, one could have used a grid-size $\Delta x = 1/M$ in the x-direction and a possibly different grid-size $\Delta y = 1/N$ in the y-direction, where $M, N \geq 2$ are integers. One can also consider boundary-value problems on more complicated polygonal domains Ω in $\mathbb{R}^2$ such that each edge of Ω is parallel with one of the co-ordinate axes: for example, the L-shaped domain $(-1, 1)^2 \setminus [0, 1]^2$. The construction above can be extended to domains with curved boundaries in any number of dimensions; at grid-points that are on (or next to) the boundary, divided differences with unequally spaced grid-points are then used.

In the case of nonlinear elliptic boundary-value problems, such as the Monge–Ampère equation on a bounded open set $\Omega \subset \mathbb{R}^d$, subject to the nonhomogeneous Dirichlet boundary condition $u = g$ on $\partial\Omega$, a finite difference approximation is easily constructed by replacing at each grid-point $(x_i, y_j) \in \Omega$ the value $u(x_i, y_j)$ of the analytical solution u (and its partial derivatives) in the PDE with the numerical solution U_{ij} (and its divided differences), and imposing the numerical boundary condition $U_{ij} = g(x_i, y_j)$ for all $(x_i, y_j) \in \partial\Omega_h$. Unfortunately, such a simple-minded method does not explicitly demand the convexity of U in any sense, and this can lead to instabilities. In fact, there is no reason why the sequence of finite difference solutions should converge to the (convex) analytical solution of the Monge–Ampère equation as $h \to 0$. Even in two space dimensions the resulting method may have multiple solutions, and special iterative solvers need to be used to select the convex solution. Enforcing convexity of the finite difference solution in higher dimensions is much more difficult. A recent successful development in this field has been the construction of so-called *wide-angle finite difference methods*, which are monotone, and the convergence theory of Barles and Souganidis therefore ensures convergence of the sequence of numerical solutions, as $h \to 0$, to the unique viscosity solution of the Monge–Ampère equation.

We close this section on finite difference methods with a brief discussion about their application to time-dependent problems. A key result is the *Lax equivalence theorem*, which states that, for a finite difference method that is consistent with a well-posed initial-value problem for a linear PDE, stability of the method implies convergence of the sequence of grid-functions defined by the method on the grid to the analytical solution as the grid-size converges to zero, and vice versa. Consider the unsteady heat equation $u_t - \Delta u + u = 0$ for $t \in (0, T]$, with $T > 0$ given, and x in the unit square $\Omega = (0, 1)^2$, subject to the homogeneous Dirichlet boundary condition $u = 0$ on $(0, T] \times \partial\Omega$ and the initial condition $u(0, x) = u_0(x)$, $x \in \Omega$, where u_0 and f are given real-valued continuous functions. The computational domain $[0, T] \times \overline{\Omega}$ is discretized by the grid $\{t^m = m\Delta t : m = 0, \ldots, M\} \times \overline{\Omega}_h$, where $\Delta t = T/M$, $M \geq 1$, and $h = 1/N$, $N \geq 2$. We consider the θ-method

$$\frac{U_{ij}^{m+1} - U_{ij}^m}{\Delta t} - (D_x^2 U_{ij}^{m+\theta} + D_y^2 U_{ij}^{m+\theta}) + U_{ij}^{m+\theta} = 0$$

for all $i, j = 1, \ldots, N-1$ and $m = 0, \ldots, M-1$, supplemented with the initial condition $U_{ij}^0 = u_0(x_i, y_j)$, $i, j = 0, \ldots, N$, and the boundary condition $U_{ij}^{m+1} = 0$, $m = 0, \ldots, M-1$, for all (i, j) such that $(x_i, y_j) \in \partial\Omega_h$. Here $\theta \in [0, 1]$ and $U_{ij}^{m+\theta} := (1-\theta)U_{ij}^m + \theta U_{ij}^{m+1}$, with U_{ij}^m and U_{ij}^{m+1} representing the approximations to $u(t^m, x_i, y_j)$ and $u(t^{m+1}, x_i, y_j)$, respectively. The values $\theta = 0, \frac{1}{2}, 1$ are particularly relevant; the corresponding finite difference methods are called the *forward* (or *explicit*) *Euler method*, the *Crank–Nicolson method*, and the *backward* (or *implicit*) *Euler method*, respectively; their truncation errors are defined by:

$$\begin{aligned}T_{ij}^{m+1} &:= \frac{u(t^{m+1}, x_i, y_j) - u(t^m, x_i, y_j)}{\Delta t} \\ &\quad - (1-\theta)(D_x^2 u(t^m, x_i, y_j) + D_y^2 u(t^m, x_i, y_j)) \\ &\quad - \theta(D_x^2 u(t^{m+1}, x_i, y_j) + D_y^2 u(t^{m+1}, x_i, y_j)) \\ &\quad + (1-\theta)u(t^m, x_i, y_j) + \theta u(t^{m+1}, x_i, y_j),\end{aligned}$$

for $i, j = 1, \ldots, N-1$, $m = 0, \ldots, M-1$. Assuming that u is sufficiently smooth, Taylor series expansion yields that $T_{ij} = \mathcal{O}(\Delta t + h^2)$ for $\theta \neq 1/2$ and $T_{ij} = \mathcal{O}((\Delta t)^2 + h^2)$ for $\theta = 1/2$. Thus in particular the forward and backward Euler methods are first-order accurate with respect to the temporal variable t and second-order accurate with respect to the spatial variables x and y, whereas the Crank–Nicolson method is second-order accurate with respect to both the temporal variable and the spatial variables. The stability properties of the θ-method are also influenced by the choice of $\theta \in [0, 1]$: we have that

$$\max_{1 \leq m \leq M} \|U^m\|_{0,h}^2 + \Delta t \sum_{m=0}^{M-1} \|U^{m+\theta}\|_{1,h}^2 \leq \|U^0\|_{0,h}^2$$

for $\theta \in [0, \frac{1}{2})$, provided that $2d(1-2\theta)\Delta t \leq h^2$, with $d = 2$ (space dimensions) in our case; and for $\theta \in [\frac{1}{2}, 1]$, irrespective of the choice of Δt and h. Thus in particular the forward (explicit) Euler method is *conditionally stable*, the condition being that $2d\Delta t \leq h^2$, with $d = 2$ here, while the Crank–Nicolson and backward (implicit) Euler methods are *unconditionally stable*.

A finite difference method approximates the analytical solution by a grid-function that is defined over a finite difference grid contained in the computational domain. We shall next consider finite element methods, which involve piecewise polynomial approximations of the analytical solution, defined over the computational domain.

4 Finite element methods

Finite element methods (FEMs) are a powerful and general class of techniques for the numerical solution of PDEs. Their historical roots can be traced back to a paper by Richard Courant published in 1943, which proposed the use of continuous piecewise affine approximations for the numerical solution of variational problems. This represented a significant advance from the practical point of view over earlier techniques by Ritz and Galerkin from the early 1900s, which were based on the use of linear combinations of smooth functions (e.g. eigenfunctions of the differential operator under consideration). The importance of Courant's contribution was, unfortunately, not recognized at the time and the idea was forgotten, until the early 1950s, when it was rediscovered by engineers. FEMs have been since developed into an effective and flexible computational tool with a firm mathematical foundation.

4.1 FEMs for elliptic PDEs

Suppose that $\Omega \subset \mathbb{R}^d$ is a bounded open set in $\mathbb{R}^d$ with a Lipschitz-continuous boundary $\partial\Omega$. We shall denote by $L^2(\Omega)$ the space of square-integrable functions (in the sense of Lebesgue), equipped with the norm $\|v\|_0 := \left(\int_\Omega |v|^2 \,\mathrm{d}x\right)^{1/2}$. Let $H^m(\Omega)$ denote the *Sobolev space* consisting of all functions $v \in L^2(\Omega)$ whose (weak) partial derivatives $\partial^\alpha v$ belong to $L^2(\Omega)$ for all α such that $|\alpha| \leq m$. $H^m(\Omega)$ is equipped with the norm $\|v\|_m := \left(\sum_{|\alpha| \leq m} \|\partial^\alpha v\|_0^2\right)^{1/2}$. We denote by $H_0^1(\Omega)$ the set of all functions $v \in H^1(\Omega)$ that vanish on $\partial\Omega$.

Let a and c be real-valued functions, defined and continuous on $\overline{\Omega}$, and suppose that there exists a positive constant c_0 such that $a(x) \geq c_0$ for all $x \in \overline{\Omega}$. Assume further that b_i, $i = 1, \ldots, d$, are continuously differentiable real-valued functions defined on $\overline{\Omega}$, such that $c - \frac{1}{2}\nabla \cdot b \geq c_0$ on $\overline{\Omega}$, where $b := (b_1, \ldots, b_d)$, and let $f \in L^2(\Omega)$. Consider the boundary-value problem:

$$-\nabla \cdot (a(x)\nabla u) + b(x) \cdot \nabla u + c(x)u = f(x),$$

for $x \in \Omega$, with $u|_{\partial\Omega} = 0$. The construction of the finite element approximation of this boundary-value problem commences by considering the following *weak formulation* of the problem: find $u \in H_0^1(\Omega)$ such that

$$B(u, v) = \ell(v) \quad \forall v \in H_0^1(\Omega), \tag{11}$$

where the bilinear form $B(\cdot, \cdot)$ is defined by

$$B(w, v) := \int_\Omega [a(x)\nabla w \cdot \nabla v + b(x) \cdot \nabla w\, v + c(x)wv] \mathrm{d}x$$

and $\ell(v) := \int_\Omega f v \,\mathrm{d}x$, with $w, v \in H_0^1(\Omega)$. If u is sufficiently smooth, for example, $u \in H^2(\Omega) \cap$

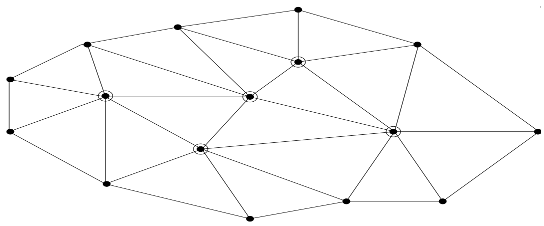

Figure 1: Finite element triangulation of the computational domain $\overline{\Omega}$, a polygonal region of $\mathbb{R}^2$. Vertices on $\partial\Omega$ are denoted by solids dots, and vertices internal to Ω by circled solid dots.

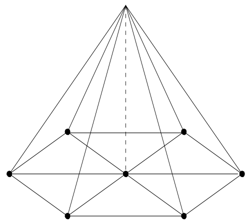

Figure 2: Piecewise linear nodal basis function. The basis function is identically zero outside a patch of triangles surrounding the central node, at which the height of the function is equal to 1.

$H_0^1(\Omega)$, then integration by parts in (11) implies that u is a *strong solution* of the boundary-value problem; i.e. $-\nabla \cdot (a(x)\nabla u) + b(x) \cdot \nabla u + c(x)u = f(x)$ almost everywhere in Ω, and $u|_{\partial\Omega} = 0$. More generally, in the absence of such an additional assumption about smoothness, the function $u \in H_0^1(\Omega)$ satisfying (11) is called a *weak solution* of this elliptic boundary-value problem. Under our assumptions on a, b, c and f, the existence of a unique weak solution follows from the Lax–Milgram theorem.

We shall consider the finite element approximation of (11) in the special case when Ω is a bounded open polygonal domain in $\mathbb{R}^2$. The first step in the construction of the FEM is to define a *triangulation* of $\overline{\Omega}$. A triangulation of $\overline{\Omega}$ is a tessellation of $\overline{\Omega}$ into a finite number of closed triangles T_i, $i = 1, \dots, M$, whose interiors are pairwise disjoint, and for each $i, j \in \{1, \dots, M\}$, $i \neq j$, for which $T_i \cap T_j$ is nonempty, $T_i \cap T_j$ is either a common vertex or a common edge of T_i and T_j (see Fig. 1). The vertices in the triangulation are also referred to as *nodes*.

Let h_T denote the longest edge of a triangle T in the triangulation, and let h be the largest among the h_T. Let, further, S_h denote the linear space of all real-valued continuous functions v_h defined on $\overline{\Omega}$ such that the restriction of v_h to any triangle in the triangulation is an affine function, and define $S_{h,0} := S_h \cap H_0^1(\Omega)$. The finite element approximation of the problem (11) is: find u_h in the *finite element space* $S_{h,0}$ such that

$$B(u_h, v_h) = \ell(v_h) \quad \forall v_h \in S_{h,0}. \qquad (12)$$

Let us denote by x_i, $i = 1, \dots, L$, the set of all vertices (nodes) in the triangulation (see Fig. 1), and let $N = N(h)$ denote the dimension of the finite element space $S_{h,0}$. We shall assume that the vertices x_i, $i = 1, \dots, L$, are numbered so that x_i, $i = 1, \dots, N$, are within Ω and the remaining $L - N$ vertices are on $\partial\Omega$. Let further $\{\varphi_j \ : \ j = 1, \dots, N\} \subset S_{h,0}$, denote the so-called *nodal basis* for $S_{h,0}$, where the basis functions are defined by $\varphi_j(x_i) = \delta_{ij}$, $i = 1, \dots, L$, $j = 1, \dots, N$. A typical piecewise linear nodal basis function is shown in Fig. 2. Thus, there exists a vector $U = (U_1, \dots, U_N)^{\mathrm{T}} \in \mathbb{R}^N$ such that

$$u_h(x) = \sum_{j=1}^{N} U_j \varphi_j(x). \qquad (13)$$

Substitution of this expansion into (12) and taking $v_h = \varphi_k$, $k = 1, \dots, N$, yields the following system of N linear algebraic equations in the N unknowns, $U_1, \dots, U_N$:

$$\sum_{j=1}^{N} B(\varphi_j, \varphi_k) U_j = \ell(\varphi_k), \quad k = 1, \dots, N. \qquad (14)$$

By recalling the definition of $B(\cdot, \cdot)$, we see that the matrix $A := ([B(\varphi_j, \varphi_k)]_{j,k=1}^N)^{\mathrm{T}}$ of this system of linear equations is sparse, positive definite (and if b is identically zero then also symmetric). The unique solution $U = (U_1, \dots, U_N)^{\mathrm{T}} \in \mathbb{R}^N$ of the linear system, upon substitution into (13), yields the computed approximation u_h to the analytical solution u on the given triangulation of the computational domain $\overline{\Omega}$, using numerical algorithms for [REF **??**]IV.NLA Sec 6.

As $S_{h,0}$ is a (finite-dimensional) linear subspace of $H_0^1(\Omega)$, $v = v_h$ is a legitimate choice in (11). By subtracting (12) from (11), with $v = v_h$, we deduce that

$$B(u - u_h, v_h) = 0 \quad \forall v_h \in S_{h,0}, \tag{15}$$

which is referred to as the *Galerkin orthogonality property* of the FEM. Hence, for any $v_h \in S_{h,0}$,

$$\begin{aligned} c_0 \|u - u_h\|_1^2 &\le B(u - u_h, u - u_h) \\ &= B(u - u_h, u - v_h) \\ &\le c_1 \|u - u_h\|_1 \|u - v_h\|_1, \end{aligned}$$

where $c_1 := (M_a^2 + M_b^2 + M_c^2)^{1/2}$, with $M_v := \max_{x \in \overline{\Omega}} |v(x)|$, $v \in \{a, b, c\}$. We thus have that

$$\|u - u_h\|_1 \le \frac{c_1}{c_0} \min_{v_h \in S_{h,0}} \|u - v_h\|_1. \tag{16}$$

This result is known as *Céa's lemma*, and is an important tool in the analysis of FEMs. Suppose, for example, that $u \in H^2(\Omega) \cap H_0^1(\Omega)$ and denote by I_h the *finite element interpolant* of u defined by

$$I_h u(x) := \sum_{j=1}^{N} u(x_j) \varphi_j(x).$$

It follows from (16) that $\|u - u_h\|_1 \le \frac{c_1}{c_0} \|u - I_h u\|_1$. Assuming further that the triangulation is *shape-regular* in the sense that there exists a positive constant c_*, independent of h, such that for each triangle in the triangulation the ratio of the longest edge to the radius of the circumscribed circle is bounded below by c_*, arguments from approximation theory imply the existence of a positive constant $\hat{c}$, independent of h, such that $\|u - I_h u\|_1 \le \hat{c} h \|u\|_2$. Hence, the following *a priori error bound* holds in the H^1 norm:

$$\|u - u_h\|_1 \le (c_1/c_0) \hat{c} h \|u\|_2.$$

We deduce from this inequality that, as the triangulation is refined by letting $h \to 0$, the sequence of finite element approximations u_h computed on successively refined triangulations converges to the analytical solution u in the H^1 norm. It is also possible to derive *a priori* error bounds in other norms, such as the L^2 norm.

The inequality (16) of Céa's lemma can be seen to express the fact that the approximation $u_h \in S_{h,0}$ to the solution $u \in H_0^1(\Omega)$ of (11) delivered by the FEM (12) is the *near-best approximation* to u from the linear subspace $S_{h,0}$ of $H_0^1(\Omega)$. Clearly, $c_1/c_0 \ge 1$. When the constant $c_1/c_0 \gg 1$, the numerical solution u_h supplied by the FEM is typically a poor approximation to u in the $\|\cdot\|_1$ norm, unless h is very small; for example, if $a(x) = c(x) \equiv \varepsilon$ and $b(x) = (1,1)^{\mathrm{T}}$, then $c_1/c_0 = \sqrt{2}(1 + \varepsilon^2)^{1/2}/\varepsilon \gg 1$ if $0 < \varepsilon \ll 1$. Such nonselfadjoint elliptic boundary-value problems arise in mathematical models of diffusion-advection-reaction, where advection dominates diffusion and reaction in the sense that $|b(x)| \gg a(x) > 0$ and $|b(x)| \gg c(x) > 0$ for all $x \in \overline{\Omega}$. The stability and approximation properties of the classical FEM (12) for such advection-dominated problems can be improved by modifying, in a consistent manner, the definitions of $B(\cdot,\cdot)$ and $\ell(\cdot)$ through the addition of 'stabilization terms', or by enriching the finite element space with special basis functions that are designed so as to capture sharp boundary and interior layers exhibited by typical solutions of advection-dominated problems. The resulting FEMs are generally referred to as *stabilized finite element methods*. A typical example is the *streamline-diffusion finite element method*, in which the bilinear form of the standard FEM is supplemented with an additional numerical diffusion term, which acts in the stream-wise direction only, i.e. in the direction of the vector b, in which classical FEMs tend to exhibit undesirable numerical oscillations.

If, on the other hand, b is identically zero on $\overline{\Omega}$, then $B(\cdot,\cdot)$ is a *symmetric* bilinear form, in the sense that $B(w,v) = B(v,w)$ for all $w, v \in H_0^1(\Omega)$. The norm $\|\cdot\|_B$ defined by $\|v\|_B := [B(v,v)]^{1/2}$ is called the *energy norm* on $H_0^1(\Omega)$ associated with the elliptic boundary-value problem (11). In fact, (11) can then be restated as the following, equivalent, variational problem: find $u \in H_0^1(\Omega)$ such that

$$J(u) \le J(v) \quad \forall v \in H_0^1(\Omega),$$

where

$$J(v) := \frac{1}{2} B(v,v) - \ell(v).$$

Analogously, the FEM (12) can then be restated equivalently as follows: find $u_h \in S_{h,0}$ such that $J(u_h) \le J(v_h)$ for all $v_h \in S_{h,0}$. Furthermore, Céa's lemma, in terms of the energy norm, $\|\cdot\|_B$, becomes $\|u - u_h\|_B = \min_{v_h \in S_{h,0}} \|u - v_h\|_B$. Thus, in the case when the function b is identically zero the numerical solution $u_h \in S_{h,0}$ delivered by the FEM is the *best approximation* to the analytical solution $u \in H^1_0(\Omega)$ in the energy norm $\|\cdot\|_B$.

We illustrate the extension of these ideas to nonlinear elliptic PDEs through a simple model problem. For a real number $p \in (1,\infty)$, let $L^p(\Omega) := \{v : \int_\Omega |v|^p \,\mathrm{d}x < \infty\}$ and $W^{1,p}(\Omega) := \{v \in L^p(\Omega) : |\nabla v| \in L^p(\Omega)\}$. Let further $W^{1,p}_0(\Omega)$ denote the set of all $v \in W^{1,p}(\Omega)$ such that $v|_{\partial\Omega} = 0$. For $f \in L^q(\Omega)$, where $1/p + 1/q = 1$, $p \in (1,\infty)$, consider the problem of finding the minimizer $u \in W^{1,p}_0(\Omega)$ of the functional

$$J(v) := \frac{1}{p}\int_\Omega |\nabla v|^p \,\mathrm{d}x - \int_\Omega f v \,\mathrm{d}x, \quad v \in W^{1,p}_0(\Omega).$$

With $S_{h,0}$ as above, the finite element approximation of the problem then consists of finding $u_h \in S_{h,0}$ that minimizes $J(v_h)$ over all $v_h \in S_{h,0}$. The existence and uniqueness of the minimizers $u \in W^{1,p}_0(\Omega)$ and $u_h \in S_{h,0}$ in the respective problems is a direct consequence of the convexity of the functional J. Moreover as $h \to 0$, u_h converges to u in the norm of the Sobolev space $W^{1,p}(\Omega)$.

Problems in electromagnetism and continuum mechanics are typically modeled by systems of PDEs involving several dependent variables, which may need to be approximated from different finite element spaces because of the disparate physical nature of the variables and the different boundary conditions that they may be required to satisfy. The resulting finite element methods are called *mixed finite element methods*. In order for a mixed FEM to possess a unique solution and for the method to be stable, the finite element spaces from which the approximations to the various components of the vector of unknowns are sought cannot be chosen arbitrarily, but need to satisfy a certain compatibility condition, usually referred to as the *inf-sup condition*.

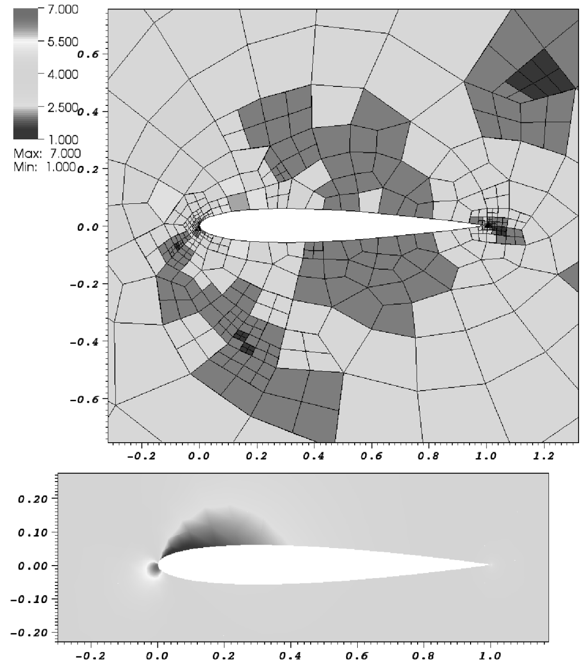

Figure 3: An *hp*-adaptive finite element grid, using polynomials with degrees $1,\dots,7$ (indicated by the colour-coding), in a discontinuous Galerkin finite element approximation of the compressible Euler equations of gas dynamics (top) and the colour contours of the approximate density on the grid (bottom). (By courtesy of Paul Houston).

FEMs of the kind described in this section, where the finite element space containing the approximate solution is a subset of the function space in which the weak solution to the problem is sought, are called *conforming finite element methods*. Otherwise, the FEM is called *nonconforming*. *Discontinuous Galerkin finite element methods* (DGFEM) are an extreme instance of a nonconforming FEM, in the sense that pointwise inter-element continuity requirements in the piecewise polynomial approximation are completely abandoned, and the analytical solution is approximated by discontinuous piecewise polynomial functions. DGFEMs have several advantages over finite difference methods: the concept of higher-order discretization is inherent to DGFEMs; it is, in addition, particularly convenient from the point of view of adaptivity that DGFEMs can easily accommodate very general tessellations of the computational domain, with local polynomial degrees in the approximation that may vary from element to element. Indeed, the notion of *adaptivity* is a powerful and important idea in the field of numerical approximation of PDEs, which we shall now further elaborate on in the context of finite element methods.

4.2 A posteriori error analysis and adaptivity

Provided that the analytical solution is sufficiently smooth, *a priori* error bounds guarantee that, as the grid size h tends to 0, the corresponding sequence of numerical approximations converges to the exact solution of the boundary-value problem. In practice one may unfortunately only afford to compute on a small number of grids/triangulations, the minimum grid size attainable being limited by the computational resources available. A further practical consideration is that the regularity of the analytical solution may exhibit large variations over the computational domain, with singularities localized at particular points (e.g. corners and edges of the domain) or low-dimensional manifolds in the interior of the domain (e.g. shocks and contact discontinuities in nonlinear conservation laws, or steep internal layers in advection-dominated diffusion equations). The error between the unknown analytical solution and numerical solutions computed on locally refined grids, which are best suited for such problems, cannot be accurately quantified by typical *a priori* error bounds and asymptotic convergence results that presuppose uniform refinement of the computational grid as the grid-size tends to 0. The alternative is to perform a computation on a chosen computational grid/triangulation and use the computed approximation to the exact solution to quantify the approximation error *a posteriori*, and also to identify parts of the computational domain where the grid-size was inadequately chosen, necessitating local, so called, *adaptive*, refinement or coarsening of the computational grid/triangulation (*h-adaptivity*). In FEMs it is also possible to locally vary the degree of the piecewise polynomial function in the finite element space (*p-adaptivity*). Finally, one may also make adjustments to the computational grid/triangulation, by moving/relocating the grid points (*r-adaptivity*). The adaptive loop for an h-adaptive FEM has the form:

`SOLVE` $\to$ `ESTIMATE` $\to$ `MARK` $\to$ `REFINE`.

Thus, a finite element approximation is first computed on a certain fixed, typically coarse, triangulation of the computational domain. Then, in the second step, an *a posteriori* error bound is used to estimate the error in the computed solution: a typical *a posteriori* error bound for an elliptic boundary-value problem $Lu = f$, where L is a second-order uniformly elliptic operator and f is a given right-hand side, is of the form $\|u - u_h\|_1 \leq C_* ||R(u_h)||_*$, where C_* is a (computable) constant, $|| \cdot ||_*$ is a certain norm, depending on the problem, and $R(u_h) = f - Lu_h$ is the (computable) *residual*, which measures the extent to which the computed numerical solution u_h fails to satisfy the PDE $Lu = f$. In the third step, on the basis of the *a posteriori* error bound, selected triangles in the triangulation are marked as those whose size is inadequate (i.e. too large or too small, relative to a fixed local tolerance, which is usually chosen as a suitable fraction of the prescribed overall tolerance `TOL`), and finally the marked triangles are refined or coarsened, as the case may be. This four-step adaptive loop is repeated either until a certain termination criterion is reached (e.g. $C_*||R(u_h)||_* <$ `TOL`) or until the computational resources are exhausted. A similar adaptive loop can be used in p-adaptive FEMs, except that the step `REFINE` is then interpreted as adjustment (i.e. increase or decrease) of the local polynomial degree, which then, instead of being a fixed integer over the entire triangulation, may vary from triangle to triangle. It is also possible to combine different adaptive strategies: for example, simultaneous h and p adaptivity is referred to as *hp-adaptivity*; thanks to the simple communication at the boundaries of adjacent elements in the subdivision of the computational domain, *hp*-adaptivity is particularly easy to incorporate into DGFEMs; see Fig. 3.

5 Finite volume methods

Finite volume methods have been developed for the numerical solution of PDEs in divergence form, such as conservation laws that arise from continuum mechanics. Consider, for example, the following system of nonlinear PDEs:

$$\frac{\partial u}{\partial t} + \nabla \cdot f(u) = 0, \tag{17}$$

where $u := (u_1, \dots, u_n)^{\mathrm{T}}$ is an n-component vector-function of the variables t and $x_1, \dots, x_d$; the vector-function $f(u) := (f_1(u), \dots, f_d(u))^{\mathrm{T}}$ is the corresponding *flux function*. The PDE (17) is supplemented with the initial condition $u(0, x) = u_0(x)$, $x \in \mathbb{R}^d$. Suppose that $\mathbb{R}^d$ has been tessellated into disjoint closed symplices κ

(intervals if $d=1$, triangles if $d=2$, and tetrahedra if $d=3$), whose union is the whole of $\mathbb{R}^d$ and such that each pair of distinct simplices from the tessellation is either disjoint, or has only closed symplices of dimension $\leq d-1$ in common. In the theory of finite volume methods the symplices κ are usually referred to as *cells* (rather than elements). For each particular cell κ in the tesselation of $\mathbb{R}^d$ the PDE (17) is integrated over κ, which gives

$$\int_\kappa \frac{\partial u}{\partial t}\,\mathrm{d}x + \int_\kappa \nabla \cdot f(u)\,\mathrm{d}x = 0. \tag{18}$$

By defining the *volume-average*

$$\bar{u}_\kappa(t) := \frac{1}{|\kappa|}\int_\kappa u(t,x)\,\mathrm{d}x, \quad t \geq 0,$$

where $|\kappa|$ is the measure of κ, and applying the divergence theorem, we deduce that

$$\frac{\mathrm{d}\bar{u}_\kappa}{\mathrm{d}t} + \frac{1}{|\kappa|}\oint_{\partial\kappa} f(u)\cdot\nu\,\mathrm{d}S = 0,$$

where $\partial\kappa$ is the boundary of κ and ν is the unit outward normal vector to $\partial\kappa$. In the present construction the constant volume-average is assigned to the barycenter of a cell, and the resulting finite volume method is therefore referred to as a *cell-centre finite volume method.* In the theory of finite volume methods the local region κ over which the PDE is integrated is called a *control volume.* Thus in the case of cell-centre finite volume methods the control volumes coincide with the cells in the tessellation. An alternative choice, resulting in *vertex-centred finite volume methods*, is that for each vertex in the computational grid one considers the patch of cells surrounding the vertex, and assigns to the vertex a control volume contained in the patch of elements (e.g., in the case of $d=2$, the polygonal domain defined by connecting the barycenters of cells that surround a vertex).

Thus far no approximation has taken place. In order to construct a practical numerical method, the integral over $\partial\kappa$ is rewritten as a sum of integrals over all $(d-1)$-dimensional open faces contained in $\partial\kappa$, and the integral over each face is approximated by replacing the normal flux $f(u)\cdot\nu$ over the face, appearing as integrand, by interpolation or extrapolation of control volume averages. This procedure can be seen as a replacement of the exact normal flux over a face of a control volume with a *numerical flux function.* Thus, for example, denoting by $e_{\kappa\lambda}$ the $(d-1)$-dimensional face of the control volume κ that is shared with a neighboring control volume λ, we have that

$$\oint_{\partial\kappa} f(u)\cdot\nu\,\mathrm{d}S \approx \sum_{\lambda\,:\,e_{\kappa\lambda}\subset\partial\kappa} g_{\kappa\lambda}(\bar{u}_\kappa, \bar{u}_\lambda),$$

where the numerical flux function $g_{\kappa\lambda}$ is required to possess the following two crucial properties:

- *Conservation* ensures that fluxes from adjacent control volumes sharing a mutual interface exactly cancel when summed. This is achieved by demanding that the numerical flux satisfies the identity

  $$g_{\kappa\lambda}(u,v) = -g_{\lambda\kappa}(v,u),$$

  for each pair of neighboring control volumes κ and λ.

- *Consistency* ensures that, for each face of each control volume, the numerical flux with identical state arguments reduces to the true total flux of that same state passing through the face, i.e.,

  $$g_{\kappa\lambda}(u,u) = \int_{e_{\kappa\lambda}} f(u)\cdot\nu\,\mathrm{d}S,$$

  for each pair of neighboring control volumes κ and λ with common face $e_{\kappa\lambda} := \kappa\cap\lambda$.

The resulting spatial discretization of the nonlinear conservation law is then further discretized with respect to the temporal variable t by time stepping, in steps of Δt, starting from the given initial datum u_0, the simplest choice being to use the explicit Euler method.

The historical roots of this construction date back to the work of Sergei Godunov in 1959 on the gas dynamics equations; Godunov used piecewise constant solution representations in each control volume with value equal to the average over the control volume and calculated a single numerical flux from the local solution of the Riemann problem posed at the interfaces. Additional resolution beyond the first-order accuracy of the

Godunov scheme can be attained by *reconstruction/recovery* from the computed cell-averages (as in the MUSCL scheme of Van Leer based on piecewise linear reconstruction, or by piecewise quadratic reconstruction as in the piecewise parabolic method (PPM) of Colella and Woodward), by exactly evolving discontinuous piecewise linear states instead of piecewise constant states, or by completely avoiding the use of Riemann solvers (as in the Nessyahu–Tadmor and Kurganov–Tadmor central difference methods).

Thanks to their in-built conservation properties, finite volume methods have been widely and successfully used for the numerical solution of both scalar nonlinear conservation laws and systems of nonlinear conservation laws, including the compressible Euler equations of gas dynamics. There is a satisfactory convergence theory of finite volume methods for scalar multidimensional conservation laws; efforts to develop a similar body of theory for multidimensional systems of nonlinear conservation laws are however hampered by the incompleteness of the theory of well-posedness for such PDE systems.

6 Spectral methods

While finite difference methods provide approximate solutions to PDEs at the points of the chosen computational grid, and finite element and finite volume methods supply continuous or discontinuous piecewise polynomial approximations on tessellations of the computational domain, spectral methods deliver approximate solutions in the form of polynomials of a certain fixed degree, which are, by definition, smooth functions over the entire computational domain. If the solution to the underlying PDE is a smooth function, a spectral method will provide a highly accurate numerical approximation to it.

Spectral approximations are typically sought as linear combinations of [REF **??**]II.xy over the computational domain. Consider a nonempty open interval (a, b) of the real line and a nonnegative *weight-function* w, which is positive on (a, b), except perhaps at countably many points in (a, b), and such that

$$\int_a^b w(x)|x|^k \, \mathrm{d}x < \infty \quad \forall k \in \{0, 1, 2, \dots\}.$$

Let, further, $L^2_w(a, b)$ denote the set of all real-valued functions v defined on (a, b) such that

$$\|v\|_w := \left(\int_a^b w(x)|v(x)|^2 \mathrm{d}x \right)^{1/2} < \infty.$$

Then, $\| \cdot \|_w$ is a norm on $L^2_w(a, b)$, induced by the inner product $(u, v)_w := \int_a^b w(x)u(x)v(x) \, \mathrm{d}x$. We say that $\{P_k\}_{k=0}^{\infty}$ is a *system of orthogonal polynomials* on (a, b) if P_k is a polynomial of exact degree k and $(P_m, P_n)_w = 0$ when $m \neq n$. For example, if $(a, b) = (-1, 1)$ and $w(x) = (1 - x)^{\alpha}(1 + x)^{\beta}$, with $\alpha, \beta \in (-1, 1)$ fixed, then the resulting system of orthogonal polynomials are the Jacobi polynomials, special cases of which are the Gegenbauer (or ultraspherical) polynomials ($\alpha = \beta \in (-1, 1)$), Chebyshev polynomials of the first kind ($\alpha = \beta = -1/2$), Chebyshev polynomials of the second kind ($\alpha = \beta = 1/2$) and Legendre polynomials ($\alpha = \beta = 0$). On a multidimensional domain $\Omega \subset \mathbb{R}^d$, $d \geq 2$, that is the cartesian product of nonempty open intervals (a_k, b_k), $k = 1, \dots, d$, of the real line and a multivariate weight-function w of the form $w(x) = w_1(x_1) \cdots w_d(x_d)$, where $x = (x_1, \dots, x_d)$ and w_k is a univariate weight-function of the variable $x_k \in (a_k, b_k)$, $k = 1, \dots, d$, orthogonal polynomials with respect to the inner product $(\cdot, \cdot)_w$ defined by $(u, v)_w = \int_\Omega w(x)u(x)v(x) \, \mathrm{d}x$ are simply products of univariate orthogonal polynomials with respect to the weights w_k, defined on the intervals (a_k, b_k), $k = 1, \dots, d$, respectively.

Spectral Galerkin methods for PDEs are based on transforming the PDE problem under consideration into a suitable weak form by multiplication with a *test function*, integration of the resulting expression over the computational domain Ω, and integration by parts, if necessary, in order to incorporate boundary conditions. Similarly as in the case of finite element methods, an approximate solution u_N to the analytical solution u is sought from a finite-dimensional linear space $S_N \subset L^2_w(\Omega)$, which is now, however, spanned by the first $(N + 1)^d$ elements of a certain system of orthogonal polynomials with respect to the weight-function w, and satisfying the associated Dirichlet boundary condition (if any); u_N is required to satisfy the same weak formulation as the analytical solution, except that the test functions are confined to the finite-dimensional linear space S_N. In order to exploit the orthogonality properties of the chosen system of orthogonal polynomials, the weight-function w has to be incorporated into the weak formulation of the problem, which is not always easy, unless of course the weight-function w already appears as a coefficient in the differential equation, or if the orthogonal polynomials in question are the Legendre polynomials (since then $w(x) \equiv 1$). We describe the construction for a uniformly elliptic PDE subject to a *homogeneous Neumann* boundary condition:

$$\begin{aligned} -\Delta u + u &= f(x) \qquad && x \in \Omega := (-1,1)^d, \\ \frac{\partial u}{\partial \nu} &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where $f \in L^2(\Omega)$ and ν denotes the unit outward normal vector to $\partial\Omega$ (or, more precisely, to the $(d-1)$-dimensional open faces contained in $\partial\Omega$). Let us consider the finite-dimensional linear space

$$\begin{aligned} S_N := \operatorname{span}\{L_\alpha := L_{\alpha_1} \cdots L_{\alpha_d} :\\ 0 \le \alpha_k \le N,\ k = 1, \dots, d\}, \end{aligned}$$

where L_{α_k} is the univariate Legendre polynomial of degree α_k of the variable $x_k \in (-1,1)$, $k = 1, \dots, d$. The *Legendre–Galerkin spectral approximation* of the boundary value problem is defined as follows: find $u_N \in S_N$ such that

$$B(u_N, v_N) = \ell(v_N) \quad \forall v_N \in S_N, \tag{19}$$

where the linear functional $\ell(\cdot)$ and the bilinear form $B(\cdot,\cdot)$ are defined by $\ell(v) := \int_\Omega f v \,\mathrm{d}x$ and $B(w,v) := \int_\Omega (\nabla w \cdot \nabla v + wv)\,\mathrm{d}x$, respectively, with $w, v \in H^1(\Omega)$. As $B(\cdot,\cdot)$ is a symmetric bilinear form and S_N is a finite-dimensional linear space, the task of determining u_N is equivalent to solving a system of linear algebraic equations with a symmetric square matrix $A \in \mathbb{R}^{K\times K}$ with $K := \dim(S_N) = (N+1)^d$. Since $B(V,V) = \|V\|_1^2 > 0$ for all $V \in S_N \setminus \{0\}$, where, as before, $\|\cdot\|_1$ denotes the $H^1(\Omega)$ norm, the matrix A is positive definite, and therefore invertible. Thus we deduce the existence and uniqueness of a solution to (19). Céa's lemma (see (16)) for (19) takes the form

$$\|u - u_N\|_1 = \min_{v_N \in S_N} \|u - v_N\|_1. \tag{20}$$

Assuming that $u \in H^s(\Omega)$, $s > 1$, results from approximation theory imply that the right-hand side of (20) is bounded by a constant multiple of $N^{1-s}\|u\|_s$, and we thus deduce the error bound

$$\|u - u_N\|_1 \le C N^{1-s} \|u\|_s, \quad s > 1.$$

Furthermore, if $u \in C^\infty(\overline{\Omega})$ (i.e. all partial derivatives of u of any order are continuous on $\overline{\Omega}$), then $\|u - u_N\|_1$ will converge to zero at a rate that is faster than any algebraic rate of convergence; such a superalgebraic convergence rate is usually referred to as *spectral convergence* and is the hallmark of spectral methods.

Since $u_N \in S_N$, there exist $U_\alpha \in \mathbb{R}$, with multi-indices $\alpha = (\alpha_1, \dots, \alpha_d) \in \{0, \dots, N\}^d$, such that

$$u_N(x) = \sum_{\alpha \in \{0,\dots,N\}^d} U_\alpha L_\alpha(x).$$

Substituting this expansion into (19) and taking $v_N = L_\beta$, with $\beta = (\beta_1, \dots, \beta_d) \in \{0, \dots, N\}^d$, we obtain the system of linear algebraic equations

$$\sum_{\alpha \in \{0,\dots,N\}^d} B(L_\alpha, L_\beta) U_\alpha = \ell l(L_\beta),\ \beta \in \{0, \dots, N\}^d \tag{21}$$

for the unknowns U_α, $\alpha \in \{0, \dots, N\}^d$, which is reminiscent of the system of linear equations (14) encountered in connection with finite element methods. There is, however, a fundamental difference: whereas the matrix of the linear system (14) was symmetric positive definite and *sparse*, the one appearing in (21) is symmetric positive definite and *full*. It has to be noted that because

$$B(L_\alpha, L_\beta) = \int_\Omega \nabla L_\alpha \cdot \nabla L_\beta \,\mathrm{d}x + \int_\Omega L_\alpha L_\beta \,\mathrm{d}x,$$

in order for the matrix of the system to become diagonal, instead of Legendre polynomials one would need to use a system of polynomials that are orthogonal in the *energy inner product* $(u,v)_B := B(u,v)$, induced by B.

If the homogeneous Neumann boundary condition considered above is replaced with a 1-periodic boundary condition in each of the d co-ordinate directions and the function f appearing on the right-hand side of the PDE $-\Delta u + u = f(x)$ on $\Omega = (0,1)^d$ is a 1-periodic function in each co-ordinate direction, then one can use trigonometric polynomials instead of Legendre polynomials in the expansion of the numerical solution. This will then result in what is known as a *Fourier–Galerkin spectral method.* Because trigonometric polynomials are orthogonal in both the $L^2(\Omega)$ and the $H^1(\Omega)$ inner product, the matrix of the resulting system of linear equations will be diagonal, which greatly simplifies the solution process. Having said this, the presence of (periodic) nonconstant coefficients in the PDE will still destroy orthogonality in the associated energy inner product $(\cdot,\cdot)_B$, and the matrix of the resulting system of linear equations will then, again, become full. Nevertheless, significant savings can be made in spectral computations through the use of fast transform methods, such as the fast Fourier transform (FFT) or the fast Chebyshev transform, and this has contributed to the popularity of Fourier and Chebyshev spectral methods.

Spectral collocation methods seek a numerical solution u_N from a certain finite-dimensional space S_N, spanned by orthogonal polynomials, just as spectral Galerkin methods, except that after expressing u_N as a finite linear combination of orthogonal polynomials and substituting this linear combination into the differential equation, rather than requiring that the difference between the left-hand side and the right-hand side of the resulting expression is orthogonal to S_N, one demands instead that this difference vanishes at certain carefully chosen points, called the *collocation points.* Boundary and initial conditions are enforced analogously. A trivial requirement in selecting the collocation points is that one ends up with as many equations as the number of unknowns, which is, in turn, equal to the dimension of the linear space S_N.

We illustrate the procedure by considering the parabolic equation

$$\partial_t u - \partial^2_{xx} u = 0, \quad (t,x) \in (0,\infty) \times (-1,1),$$

subject to the initial condition $u(0,x) = u_0(x)$ with $x \in [-1,1]$ and the homogeneous Dirichlet boundary conditions $u(t,-1) = 0$, $u(t,1) = 0$, $t \in (0,\infty)$. A numerical approximation u_N is sought in the form of the finite linear combination

$$u_N(t,x) = \sum_{k=0}^{N} a_k(t) T_k(x)$$

with $(t,x) \in [0,\infty) \times [-1,1]$, where $T_k(x) := \cos(k \arccos(x))$, $x \in [-1,1]$, is the *Chebyshev polynomial* (of the first kind) of degree $k \geq 0$. Note that there are $N+1$ unknowns: the coefficients $a_k(t)$, $k = 0,1,\ldots,N$. We thus require the same number of equations. The function u_N is substituted into the PDE and it is demanded that, for $t \in (0,\infty)$ and $k = 1,\ldots,N-1$,

$$\partial_t u_N(t,x_k) - \partial^2_{xx} u_N(t,x_k) = 0;$$

and $u_N(t,-1) = 0$ and $u_N(t,1) = 0$ for $t \in (0,\infty)$, supplemented by the initial condition $u_N(0,x_k) = u_0(x_k)$ for $k = 0,\ldots,N$, where the $(N+1)$ collocation points are defined by $x_k := \cos(k\pi/N)$, $k = 0,\ldots,N$; these are the $(N+1)$ points of extrema of T_N on the interval $[-1,1]$. By writing $u^k(t) := u_N(t,x_k)$, after some calculation based on properties of Chebyshev polynomials one arrives at the following set of ordinary differential equations:

$$\frac{\mathrm{d}u^k(t)}{\mathrm{d}t} = \sum_{l=1}^{N-1} (D_N^2)_{kl} u^l(t), \quad k = 1,\ldots,N-1,$$

where D_N^2 is the *spectral differentiation matrix of second order*, whose entries $(D_N^2)_{kl}$ can be explicitly calculated. One can then use any standard numerical method for a system of ordinary differential equations to evolve the values $u^k(t) = u_N(t,x_k)$ of the approximate solution u_N at the collocations points x_k, $k = 1,\ldots,N-1$, contained in $(-1,1)$, from the values of the initial datum u_0 at the same points.

7 Concluding remarks

We have concentrated on four general and widely applicable families of numerical methods — finite difference, finite element, finite volume and spectral methods. For additional details the reader is referred to the books listed under the heading *Further Reading* below, and to the rich literature on numerical methods for PDEs for the construction and analysis of other important techniques for specialized PDE problems.

Further Reading

1. Bangerth, W. and Rannacher, R., 2003. Adaptive finite element methods for differential equations. Lectures in Mathematics ETH Zürich. Birkhuser Verlag, Basel.

2. Brenner, S. C. and Scott, L. R., 2008. The mathematical theory of finite element methods. Third edition. Texts in Applied Mathematics, **15**. Springer, New York.
3. Brezzi, F. and Fortin, M., 1991. Mixed and hybrid finite element methods. Springer Series in Computational Mathematics, **15**. Springer, New York.
4. Canuto, C., Hussaini, M., Quarteroni, A., and Zhang, T., 2006. Spectral methods: fundamentals in single domains. Springer, Berlin.
5. Ciarlet, P. G., 2002. The finite element method for elliptic problems. Classics in Applied Mathematics, **40**. SIAM, Philadelphia, PA.
6. Eymard R., Galluoët, T., and Herbin R., 2000. Finite volume methods. In: Handbook of Numerical Analysis **7**, 713–1020, North Holland, Amsterdam.
7. Gustafsson, B., Kreiss, H.-O., and Oliger, J., 1995. Time dependent problems and difference methods. Wiley-Interscience, New York.
8. Di Pietro, D. A. and Ern, A., 2012. Mathematical aspects of discontinuous Galerkin methods. Mathematics & Applications, **69**. Springer, Heidelberg.
9. Hackbusch, W., 1985. Multigrid methods and applications. Springer Series in Computational Mathematics, **4**. Springer-Verlag, Berlin.
10. Johnson, C., 2009. Numerical solution of partial differential equations by the finite element method. Reprint of the 1987 edition. Dover Publications, Inc., Mineola, New York.
11. LeVeque, R. J., 2002. Finite volume methods for hyperbolic problems. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge.
12. Richtmyer, R. D. and Morton, K. W., 1994. Difference methods for initial-value problems. Reprint of the second edition. Robert E. Krieger Publishing Co., Inc., Malabar, FL.
13. Schwab, Ch., 1998. p- and hp-finite element methods. Oxford University Press, New York.
14. Trefethen, L. N., 2000. Spectral Methods in Matlab. SIAM, Philadelphia, PA.
15. Verfürth, R., 2013. A posteriori error estimation techniques for finite element methods. The Clarendon Press, Oxford University Press, Oxford.