Chapter 11
Discretization of Differential Equations

Many processes in science and technology can be described by differential equations involving the rate of changes in time or space of a continuous variable, the unknown function. While the simplest differential equations can be solved exactly, a numerical treatment is necessary in most cases and the equations have to be discretized to turn them into a finite system of equations which can be solved by computers [6, 155, 200]. In this chapter we discuss different methods to discretize differential equations. The simplest approach is the method of finite differences, which replaces the differential quotients by difference quotients (Chap. 3). It is often used for the discretization of time. Finite difference methods for the space variables work best on a regular grid. Finite volume methods are very popular in computational fluid dynamics. They take averages over small control volumes and can be easily used with irregular grids. Finite differences and finite volumes belong to the general class of finite element methods which are prominent in the engineering sciences and use an expansion in piecewise polynomials with small support. Spectral methods, on the other hand, expand the solution as a linear combination of global basis functions like polynomials or trigonometric functions. A general concept for the discretization of differential equations is the method of weighted residuals which minimizes the weighted residual of a numerical solution. Most popular is Galerkin's method which uses the expansion functions also as weight functions. Simpler are the point collocation and sub-domain collocation methods which fulfill the differential equation only at certain points or averaged over certain control volumes. More demanding is the least-squares method which has become popular in computational fluid dynamics and computational electrodynamics. The least-square integral provides a measure for the quality of the solution which can be used for adaptive grid size control.

If the Green's function is available for a problem, the method of boundary elements is an interesting alternative. It reduces the dimensionality and is, for instance, very popular in chemical physics to solve the Poisson-Boltzmann equation.

P.O.J. Scherer, *Computational Physics*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-00401-3_11,

11.1 Classification of Differential Equations

An ordinary differential equation (ODE) is a differential equation for a function of one single variable, like Newton's law for the motion of a body under the influence of a force field

$$m\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{x}(t) = \mathbf{F}(\mathbf{x}, t), \tag{11.1}$$

a typical initial value problem where the solution in the domain $t_0 \le t \le T$ is determined by position and velocity at the initial time

$$\mathbf{x}(t = t_0) = \mathbf{x}_0 \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t = t_0) = \mathbf{v}_0. \tag{11.2}$$

Such equations of motion are discussed in Chap. 12. They also appear if the spatial derivatives of a partial differential equation have been discretized. Usually this kind of equation is solved by numerical integration over finite time steps $\Delta t = t_{n+1} - t_n$. Boundary value problems, on the other hand, require certain boundary conditions[1] to be fulfilled, for instance the linearized Poisson-Boltzmann equation in one dimension (Chap. 17)

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Phi - \kappa^2\Phi = -\frac{1}{\varepsilon}\rho(x) \tag{11.3}$$

where the value of the potential is prescribed on the boundary of the domain $x_0 \le x \le x_1$

$$\Phi(x_0) = \Phi_0 \quad \Phi(x_1) = \Phi_1. \tag{11.4}$$

Partial differential equations (PDE) finally involve partial derivatives with respect to at least two different variables, in many cases time and spatial coordinates.

11.1.1 Linear Second Order PDE

A very important class are second order linear partial differential equations of the general form

$$\left[\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{N} b_i \frac{\partial}{\partial x_i} + c\right] f(x_1 \dots x_N) + d = 0 \tag{11.5}$$

where the coefficients a_{ij}, b_i, c, d are functions of the variables $x_1 \dots x_N$ but do not depend on the function f itself. The equation is classified according to the eigenvalues of the coefficient matrix a_{ij} as [141]

[1] Dirichlet b.c. concern the function values, Neumann b.c. the derivative, Robin b.c. a linear combination of both, Cauchy b.c. the function value and the normal derivative and mixed b.c. have different character on different parts of the boundary.

- *elliptical* if all eigenvalues are positive or all eigenvalues are negative, like for the Poisson equation (Chap. 17)

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\Phi(x,y,z)=-\frac{1}{\varepsilon}\rho(x,y,z) \tag{11.6}$$

- *hyperbolic* if one eigenvalue is negative and all the other eigenvalues are positive or vice versa, for example the wave equation in one spatial dimension (Chap. 18)

$$\frac{\partial^2}{\partial t^2}f-c^2\frac{\partial^2}{\partial x^2}f=0 \tag{11.7}$$

- *parabolic* if at least one eigenvalue is zero, like for the diffusion equation (Chap. 19)

$$\frac{\partial}{\partial t}f(x,y,z,t)-D\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)f(x,y,z,t)=S(x,y,z,t) \tag{11.8}$$

- *ultra-hyperbolic* if there is no zero eigenvalue and more than one positive as well as more than one negative eigenvalue. Obviously the dimension then must be 4 at least.

11.1.2 Conservation Laws

One of the simplest first order partial differential equations is the advection equation

$$\frac{\partial}{\partial t}f(x,t)+u\frac{\partial}{\partial x}f(x,t)=0 \tag{11.9}$$

which describes transport of a conserved quantity with density f (for instance mass, number of particles, charge etc.) in a medium streaming with velocity u. This is a special case of the class of conservation laws (also called continuity equations)

$$\frac{\partial}{\partial t}f(\mathbf{x},t)+\operatorname{div}\mathbf{J}(\mathbf{x},t)=g(\mathbf{x},t) \tag{11.10}$$

which are very common in physics. Here $\mathbf{J}$ describes the corresponding flux and g is an additional source (or sink) term. For instance the advection-diffusion equation (also known as convection equation) has this form which describes quite general transport processes:

$$\frac{\partial}{\partial t}C=\operatorname{div}(D\operatorname{grad}C-\mathbf{u}C)+S(\mathbf{x},t)=-\operatorname{div}\mathbf{J}+S(\mathbf{x},t) \tag{11.11}$$

where one contribution to the flux

$$\mathbf{J}=-D\operatorname{grad}C+\mathbf{u}C \tag{11.12}$$

is proportional to the gradient of the concentration C (Fick's first law) and the second part depends on the velocity field $\mathbf{u}$ of a streaming medium. The source term

S represents the effect of chemical reactions. Equation (11.11) is also similar to the drift-diffusion equation in semiconductor physics and closely related to the Navier Stokes equations which are based on the Cauchy momentum equation [1]

$$\rho \frac{d\mathbf{u}}{dt} = \rho\left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\,\mathrm{grad}\,\mathbf{u}\right) = \mathrm{div}\,\sigma + \mathbf{f} \tag{11.13}$$

where σ denotes the stress tensor. Equation (11.10) is the strong or differential form of the conservation law. The requirements on the smoothness of the solution are reduced by using the integral form which is obtained with the help of Gauss' theorem

$$\int_V \left(\frac{\partial}{\partial t} f(\mathbf{x},t) - g(\mathbf{x},t)\right) dV + \oint_{\partial V} \mathbf{J}(\mathbf{x},t)\, d\mathbf{A} = 0. \tag{11.14}$$

An alternative integral form results from Galerkin's [98] method of weighted residuals which introduces a weight function $w(\mathbf{x})$ and considers the equation

$$\int_V \left(\frac{\partial}{\partial t} f(\mathbf{x},t) + \mathrm{div}\,\mathbf{J}(\mathbf{x},t) - g(\mathbf{x},t)\right) w(\mathbf{x})\, dV = 0 \tag{11.15}$$

or after applying Gauss' theorem

$$\int_V \left\{\left(\frac{\partial}{\partial t} f(\mathbf{x},t) - g(\mathbf{x},t)\right) w(\mathbf{x}) - \mathbf{J}(\mathbf{x},t)\,\mathrm{grad}\, w(\mathbf{x})\right\} dV$$
$$+ \oint_{\partial V} w(\mathbf{x})\mathbf{J}(\mathbf{x},t)\, d\mathbf{A} = 0. \tag{11.16}$$

The so called weak form of the conservation law states that this equation holds for arbitrary weight functions w.

11.2 Finite Differences

The simplest method to discretize a differential equation is to introduce a grid of equidistant points and to discretize the differential operators by finite differences (FDM) as described in Chap. 3. For instance, in one dimension the first and second derivatives can be discretized by

$$x \to x_m = m\Delta x \quad m = 1 \dots M \tag{11.17}$$

$$f(x) \to f_m = f(x_m) \quad m = 1 \dots M \tag{11.18}$$

$$\frac{\partial f}{\partial x} \to \left(\frac{\partial}{\partial x} f\right)_m = \frac{f_{m+1} - f_m}{\Delta x} \quad \text{or} \quad \left(\frac{\partial}{\partial x} f\right)_m = \frac{f_{m+1} - f_{m-1}}{2\Delta x} \tag{11.19}$$

$$\frac{\partial^2 f}{\partial x^2} \to \left(\frac{\partial^2}{\partial x^2} f\right)_m = \frac{f_{m+1} + f_{m-1} - 2f_m}{\Delta x^2}. \tag{11.20}$$

These expressions are not well defined at the boundaries of the grid $m = 1, M$ unless the boundary conditions are taken into account. For instance, in case of a Dirichlet problem f_0 and f_{M+1} are given boundary values and

$$\left(\frac{\partial}{\partial x}f\right)_1 = \frac{f_2 - f_0}{2\Delta x} \quad \left(\frac{\partial^2}{\partial x^2}f\right)_1 = \frac{f_2 - 2f_1 + f_0}{\Delta x^2} \tag{11.21}$$

$$\left(\frac{\partial}{\partial x}f\right)_M = \frac{f_{M+1} - f_M}{\Delta x} \text{ or } \frac{f_{M+1} - f_{M-1}}{2\Delta x}$$

$$\left(\frac{\partial^2}{\partial x^2}f\right)_M = \frac{f_{M-1} - 2f_M + f_{M+1}}{\Delta x^2}. \tag{11.22}$$

Other kinds of boundary conditions can be treated in a similar way.

11.2.1 Finite Differences in Time

Time derivatives can be treated similarly using an independent time grid

$$t \to t_n = n\Delta t \quad n = 1 \dots N \tag{11.23}$$

$$f(t,x) \to f_m^n = f(t_n, x_m) \tag{11.24}$$

and finite differences like the first order forward difference quotient

$$\frac{\partial f}{\partial t} \to \frac{f_m^{n+1} - f_m^n}{\Delta t} \tag{11.25}$$

or the symmetric difference quotient

$$\frac{\partial f}{\partial t} \to \frac{f_m^{n+1} - f_m^{n-1}}{2\Delta t} \tag{11.26}$$

to obtain a system of equations for the function values at the grid points f_m^n. For instance for the diffusion equation in one spatial dimension

$$\frac{\partial f(x,t)}{\partial t} = D\frac{\partial^2}{\partial x^2}f(x,t) + S(x,t) \tag{11.27}$$

the simplest discretization is the FTCS (forward in time, centered in space) scheme

$$\left(f_m^{n+1} - f_m^n\right) = D\frac{\Delta t}{\Delta x^2}\left(f_{m+1}^n + f_{m-1}^n - 2f_m^n\right) + S_m^n \Delta t \tag{11.28}$$

which can be written in matrix notation as

$$\mathbf{f}_{n+1} - \mathbf{f}_n = D\frac{\Delta t}{\Delta x^2}M\mathbf{f}_n + \mathbf{S}_n\Delta t \tag{11.29}$$

with

$$\mathbf{f}_n = \begin{pmatrix} f_1^n \\ f_2^n \\ f_3^n \\ \vdots \\ f_M^n \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 \end{pmatrix}. \tag{11.30}$$

11.2.2 Stability Analysis

Fully discretized linear differential equations provide an iterative algorithm of the type[2]

$$\mathbf{f}_{n+1} = A\mathbf{f}_n + \mathbf{S}_n \Delta t \tag{11.31}$$

which propagates numerical errors according to

$$\mathbf{f}_{n+1} + \varepsilon_{n+1} = A(\mathbf{f}_n + \varepsilon_n) + \mathbf{S}_n \Delta t \tag{11.32}$$

$$\varepsilon_{j+1} = A\varepsilon_j. \tag{11.33}$$

Errors are amplified exponentially if the absolute value of at least one eigenvalue of A is larger than one. The algorithm is stable if all eigenvalues of A are smaller than one in absolute value (Sect. 1.4). If the eigenvalue problem is difficult to solve, the von Neumann analysis is helpful which decomposes the errors into a Fourier series and considers the Fourier components individually by setting

$$\mathbf{f}_n = g^n(k) \begin{pmatrix} e^{ik} \\ \vdots \\ e^{ikM} \end{pmatrix} \tag{11.34}$$

and calculating the amplification factor

$$\left| \frac{f_m^{n+1}}{f_m^n} \right| = |g(k)|. \tag{11.35}$$

The algorithm is stable if $|g(k)| \leq 1$ for all k.

Example For the discretized diffusion equation (11.28) we find

$$g^{n+1}(k) = g^n(k) + 2D\frac{\Delta t}{\Delta x^2} g^n(k)(\cos k - 1) \tag{11.36}$$

$$g(k) = 1 + 2D\frac{\Delta t}{\Delta x^2}(\cos k - 1) = 1 - 4D\frac{\Delta t}{\Delta x^2}\sin^2\left(\frac{k}{2}\right) \tag{11.37}$$

$$1 - 4D\frac{\Delta t}{\Delta x^2} \leq g(k) \leq 1 \tag{11.38}$$

hence stability requires

$$D\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}. \tag{11.39}$$

[2] Differential equations which are higher order in time can be always brought to first order by introducing the time derivatives as additional variables.

11.2.3 Method of Lines

Alternatively time can be considered as a continuous variable. The discrete values of the function then are functions of time (so called lines)

$$f_m(t) \tag{11.40}$$

and a set of ordinary differential equations has to be solved. For instance for diffusion in one dimension (11.27) the equations

$$\frac{\mathrm{d}f_m}{\mathrm{d}t} = \frac{D}{h^2}(f_{m+1} + f_{m-1} - 2f_m) + S_m(t) \tag{11.41}$$

which can be written in matrix notation as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} f_1 \\ f_1 \\ f_2 \\ \vdots \\ f_M \end{pmatrix} = \frac{D}{\Delta x^2}\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 \end{pmatrix}\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_M \end{pmatrix} + \begin{pmatrix} S_1 + \frac{D}{h^2}f_0 \\ S_2 \\ S_3 \\ \vdots \\ S_M + \frac{D}{h^2}f_{M+1} \end{pmatrix} \tag{11.42}$$

or briefly

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{f}(t) = A\mathbf{f}(t) + \mathbf{S}(t). \tag{11.43}$$

Several methods to integrate such a semi-discretized equation will be discussed in Chap. 12. If eigenvectors and eigenvalues of A are easy available, an eigenvector expansion can be used.

11.2.4 Eigenvector Expansion

A homogeneous system

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{f}(t) = A\mathbf{f}(t) \tag{11.44}$$

where the matrix A is obtained from discretizing the spatial derivatives, can be solved by an eigenvector expansion. From the eigenvalue problem

$$A\mathbf{f} = \lambda\mathbf{f} \tag{11.45}$$

we obtain the eigenvalues λ and eigenvectors $\mathbf{f}_\lambda$ which provide the particular solutions:

$$\mathbf{f}(t) = \mathrm{e}^{\lambda t}\mathbf{f}_\lambda \tag{11.46}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{\lambda t}\mathbf{f}_\lambda\right) = \lambda\left(\mathrm{e}^{\lambda t}\mathbf{f}_\lambda\right) = A\left(\mathrm{e}^{\lambda t}\mathbf{f}_\lambda\right). \tag{11.47}$$

These can be used to expand the general solution

$$\mathbf{f}(t) = \sum_\lambda C_\lambda \mathrm{e}^{\lambda t} \mathbf{f}_\lambda. \tag{11.48}$$

The coefficients C_λ follow from the initial values by solving the linear equations

$$\mathbf{f}(t=0) = \sum_\lambda C_\lambda \mathbf{f}_\lambda. \tag{11.49}$$

If the differential equation is second order in time

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{f}(t) = A\mathbf{f}(t) \tag{11.50}$$

the particular solutions are

$$\mathbf{f}(t) = \mathrm{e}^{\pm t\sqrt{\lambda}}\mathbf{f}_\lambda \tag{11.51}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\mathrm{e}^{\pm t\sqrt{\lambda}}\mathbf{f}_\lambda\right) = \lambda\left(\mathrm{e}^{\pm t\sqrt{\lambda}}\mathbf{f}_\lambda\right) = A\left(\mathrm{e}^{\pm t\sqrt{\lambda}}\mathbf{f}_\lambda\right) \tag{11.52}$$

and the eigenvector expansion is

$$\mathbf{f}(t) = \sum_\lambda \left(C_{\lambda+}\mathrm{e}^{t\sqrt{\lambda}} + C_{\lambda-}\mathrm{e}^{-t\sqrt{\lambda}}\right)\mathbf{f}_\lambda. \tag{11.53}$$

The coefficients $C_{\lambda\pm}$ follow from the initial amplitudes and velocities

$$\begin{aligned} \mathbf{f}(t=0) &= \sum_\lambda (C_{\lambda+} + C_{\lambda-})\mathbf{f}_\lambda \\ \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{f}(t=0) &= \sum_\lambda \sqrt{\lambda}(C_{\lambda+} - C_{\lambda-})\mathbf{f}_\lambda. \end{aligned} \tag{11.54}$$

For a first order inhomogeneous system

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{f}(t) = A\mathbf{f}(t) + \mathbf{S}(t) \tag{11.55}$$

the expansion coefficients have to be time dependent

$$\mathbf{f}(t) = \sum_\lambda C_\lambda(t)\mathrm{e}^{\lambda t}\mathbf{f}_\lambda \tag{11.56}$$

and satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{f}(t) - A\mathbf{f}(t) = \sum_\lambda \frac{\mathrm{d}C_\lambda}{\mathrm{d}t}\mathrm{e}^{\lambda t}\mathbf{f}_\lambda = \mathbf{S}(t). \tag{11.57}$$

After taking the scalar product with $\mathbf{f}_\mu$[3]

$$\frac{\mathrm{d}C_\mu}{\mathrm{d}t} = \mathrm{e}^{-\mu t}\left(\mathbf{f}_\mu \mathbf{S}(t)\right) \tag{11.58}$$

[3] If A is not Hermitian we have to distinguish left- and right-eigenvectors.

can be solved by a simple time integration. For a second order system

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{f}(t) = A\mathbf{f}(t) + \mathbf{S}(t) \tag{11.59}$$

we introduce the first time derivative as a new variable

$$\mathbf{g} = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{f} \tag{11.60}$$

to obtain a first order system of double dimension

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\mathbf{f}\\ \mathbf{g}\end{pmatrix} = \begin{pmatrix}0 & 1\\ A & 0\end{pmatrix}\begin{pmatrix}\mathbf{f}\\ \mathbf{g}\end{pmatrix} + \begin{pmatrix}\mathbf{S}\\ 0\end{pmatrix} \tag{11.61}$$

where eigenvectors and eigenvalues can be found from those of A (11.45)

$$\begin{pmatrix}0 & 1\\ A & 0\end{pmatrix}\begin{pmatrix}\mathbf{f}_\lambda\\ \pm\sqrt{\lambda}\mathbf{f}_\lambda\end{pmatrix} = \begin{pmatrix}\pm\sqrt{\lambda}\mathbf{f}_\lambda\\ \lambda\mathbf{f}_\lambda\end{pmatrix} = \pm\sqrt{\lambda}\begin{pmatrix}\mathbf{f}_\lambda\\ \pm\sqrt{\lambda}\mathbf{f}_\lambda\end{pmatrix} \tag{11.62}$$

$$\left(\pm\sqrt{\lambda}\,\mathbf{f}_\lambda^T\ \mathbf{f}_\lambda^T\right)\begin{pmatrix}0 & 1\\ A & 0\end{pmatrix} = \left(\lambda\,\mathbf{f}_\lambda^T\ \pm\sqrt{\lambda}\mathbf{f}_\lambda^T\right) = \pm\sqrt{\lambda}\left(\pm\sqrt{\lambda}\,\mathbf{f}_\lambda^T\ \mathbf{f}_\lambda^T\right). \tag{11.63}$$

Insertion of

$$\sum_\lambda C_{\lambda+}\mathrm{e}^{\sqrt{\lambda}t}\begin{pmatrix}\mathbf{f}_\lambda\\ \sqrt{\lambda}\mathbf{f}_\lambda\end{pmatrix} + C_{\lambda-}\mathrm{e}^{-\sqrt{\lambda}t}\begin{pmatrix}\mathbf{f}_\lambda\\ -\sqrt{\lambda}\mathbf{f}_\lambda\end{pmatrix}$$

gives

$$\sum_\lambda \frac{\mathrm{d}C_{\lambda+}}{\mathrm{d}t}\mathrm{e}^{\sqrt{\lambda}t}\begin{pmatrix}\mathbf{f}_\lambda\\ \sqrt{\lambda}\mathbf{f}_\lambda\end{pmatrix} + \frac{\mathrm{d}C_{\lambda-}}{\mathrm{d}t}\mathrm{e}^{\sqrt{\lambda}t}\begin{pmatrix}\mathbf{f}_\lambda\\ -\sqrt{\lambda}\mathbf{f}_\lambda\end{pmatrix} = \begin{pmatrix}\mathbf{S}(t)\\ 0\end{pmatrix} \tag{11.64}$$

and taking the scalar product with one of the left-eigenvectors we end up with

$$\frac{\mathrm{d}C_{\lambda+}}{\mathrm{d}t} = \frac{1}{2}\big(\mathbf{f}_\lambda\mathbf{S}(t)\big)\mathrm{e}^{-\sqrt{\lambda}t} \tag{11.65}$$

$$\frac{\mathrm{d}C_{\lambda-}}{\mathrm{d}t} = -\frac{1}{2}\big(\mathbf{f}_\lambda\mathbf{S}(t)\big)\mathrm{e}^{\sqrt{\lambda}t}. \tag{11.66}$$

11.3 Finite Volumes

Whereas the finite differences method uses function values

$$f_{i,j,k} = f(x_i, y_j, z_k) \tag{11.67}$$

at the grid points

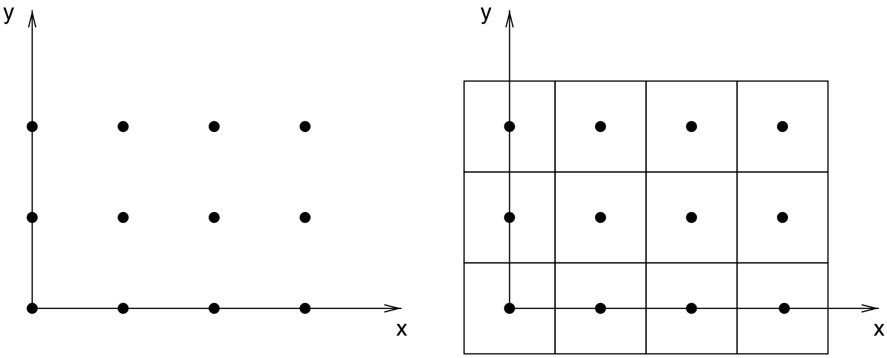


Fig. 11.1 (Finite volume method) The domain V is divided into small control volumes V_r, in the simplest case cubes around the grid points $\mathbf{r}_{ijk}$

$$\mathbf{r}_{ijk} = (x_i, y_j, z_k), \tag{11.68}$$

the finite volume method (FVM) [79] averages function values and derivatives over small control volumes V_r which are disjoint and span the domain V (Fig. 11.1)

$$V = \bigcup_r V_r \qquad V_r \cap V_{r'} = \emptyset \quad \forall r \neq r'. \tag{11.69}$$

The averages are

$$\overline{f}_r = \frac{1}{V_r}\int_{V_r} dV\, f(\mathbf{r}) \tag{11.70}$$

or in the simple case of cubic control volumes of equal size h^3

$$\overline{f}_{ijk} = \frac{1}{h^3}\int_{x_i-h/2}^{x_i+h/2} dx \int_{y_j-h/2}^{y_j+h/2} dy \int_{z_k-h/2}^{z_k+h/2} dz\, f(x,y,z). \tag{11.71}$$

Such average values have to be related to discrete function values by numerical integration (Chap. 4). The midpoint rule (4.17), for instance replaces the average by the central value

$$\overline{f}_{ijk} = f(x_i, y_j, z_k) + O\big(h^2\big) \tag{11.72}$$

whereas the trapezoidal rule (4.13) implies the average over the eight corners of the cube

$$\overline{f}_{ijk} = \frac{1}{8}\sum_{m,n,p=\pm 1} f(x_{i+m/2}, y_{j+n/2}, z_{k+p/2}) + O\big(h^2\big). \tag{11.73}$$

In (11.73) the function values refer to a dual grid [79] centered around the vertices of the original grid (11.68) (Fig. 11.2),

$$\mathbf{r}_{i+1/2,j+1/2,k+1/2} = \left(x_i + \frac{h}{2}, y_j + \frac{h}{2}, z_k + \frac{h}{2}\right). \tag{11.74}$$

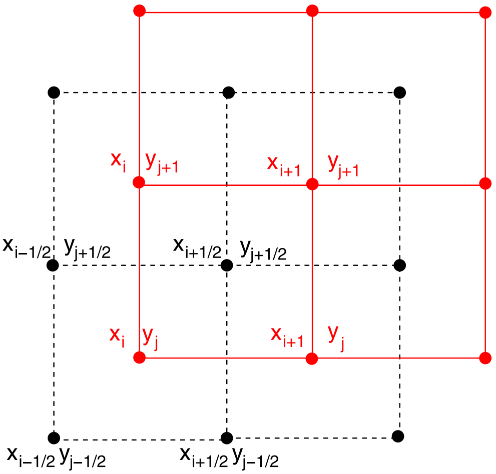


Fig. 11.2 (Dual grid) The dual grid (*black*) is centered around the vertices of the original grid (*red*)

The average gradient can be rewritten using the generalized Stokes' theorem as

$$\overline{\text{grad } f_{ijk}} = \frac{1}{V}\int_{V_{ijk}} dV \text{ grad } f(\mathbf{r}) = \oint_{\partial V_{ijk}} f(\mathbf{r})\, d\mathbf{A}. \tag{11.75}$$

For a cubic grid we have to integrate over the six faces of the control volume

$$\overline{\text{grad } f_{ijk}} = \frac{1}{h^3}\begin{pmatrix} \int_{z_k-h/2}^{z_k+h/2} dz \int_{y_j-h/2}^{y_j+h/2} dy (f(x_i+\frac{h}{2}, y, z) - f(x_i-\frac{h}{2}, y, z)) \\ \int_{z_k-h/2}^{z_k+h/2} dz \int_{x_i-h/2}^{x_i+h/2} dx (f(x_i, y+\frac{h}{2}, z) - f(x_i, y-\frac{h}{2}, z)) \\ \int_{x_i-h/2}^{x_i+h/2} dx \int_{y_j-h/2}^{y_j+h/2} dy (f(x_i, y, z+\frac{h}{2}) - f(x_i, y, z-\frac{h}{2})) \end{pmatrix}. \tag{11.76}$$

The integrals have to be evaluated numerically. Applying as the simplest approximation the midpoint rule (4.17)

$$\int_{x_i-h/2}^{x_i+h/2} dx \int_{y_j-h/2}^{y_j+h/2} dy\, f(x,y) = h^2\big(f(x_i, y_j) + O\big(h^2\big)\big) \tag{11.77}$$

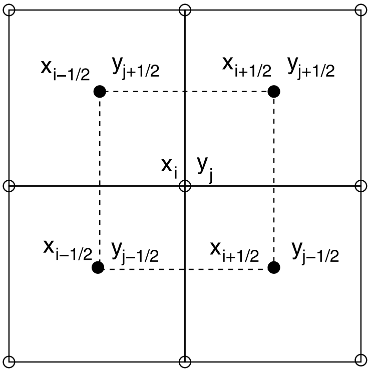
this becomes

$$\overline{\text{grad } f_{ijk}} = \frac{1}{h}\begin{pmatrix} f(x_i+\frac{h}{2}, y_j, z_k) - f(x_i-\frac{h}{2}, y_j, z_k) \\ f(x_i, y_j+\frac{h}{2}, z_k) - f(x_i, y_j-\frac{h}{2}, z_k) \\ f(x_i, y_j, z_k+\frac{h}{2}) - f(x_i, y_j, z_k-\frac{h}{2}) \end{pmatrix} \tag{11.78}$$

which involves symmetric difference quotients. However, the function values in (11.78) refer neither to the original nor to the dual grid. Therefore we interpolate (Fig. 11.3)

$$f\left(x_i \pm \frac{h}{2}, y_j, z_k\right) \approx \frac{1}{2}\big(f(x_i, y_j, z_k) + f(x_{i\pm 1}, y_j, z_k)\big) \tag{11.79}$$

Fig. 11.3 (Interpolation between grid points) Interpolation is necessary to relate the averaged gradient (11.78) to the original or dual grid



$$\frac{1}{h}\left(f\left(x_i+\frac{h}{2},y_j,z_k\right)-f\left(x_i-\frac{h}{2},y_j,z_k\right)\right) \approx \frac{1}{2h}\big(f(x_{i+1},y_j,z_k)-f(x_{i-1},y_j,z_k)\big) \tag{11.80}$$

or

$$f\left(x_i\pm\frac{h}{2},y_j,z_k\right)\approx\frac{1}{4}\sum_{m,n=\pm1}f\left(x_i\pm\frac{h}{2},y_j+m\frac{h}{2},z_k+n\frac{h}{2}\right). \tag{11.81}$$

The finite volume method is capable of treating discontinuities and is very flexible concerning the size and shape of the control volumes.

11.3.1 Discretization of fluxes

Integration of (11.10) over a control volume and application of Gauss' theorem gives the integral form of the conservation law

$$\frac{1}{V}\oint \mathbf{J}\,d\mathbf{A}+\frac{\partial}{\partial t}\frac{1}{V}\int f\,dV=\frac{1}{V}\int g\,dV \tag{11.82}$$

which involves the flux $\mathbf{J}$ of some property like particle concentration, mass, energy or momentum density or the flux of an electromagnetic field. The total flux through a control volume is given by the surface integral

$$\Phi=\oint_{\partial V}\mathbf{J}\,d\mathbf{A} \tag{11.83}$$

which in the special case of a cubic volume element of size h^3 becomes the sum over the six faces of the cube (Fig. 11.4)

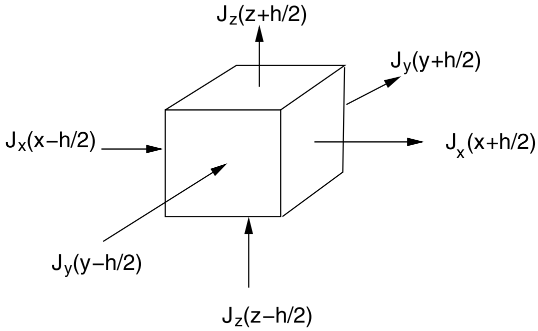


Fig. 11.4 Flux through a control volume

$$
\begin{aligned}
\Phi &= \sum_{r=1}^{6} \int_{A_r} \mathbf{J}\, d\mathbf{A} \\
&= \int_{x_i-h/2}^{x_i+h/2} dx \int_{y_j-h/2}^{y_j+h/2} dy \left(J_z\left(x, y, z_k + \frac{h}{2}\right) - J_z\left(x, y, z_k - \frac{h}{2}\right)\right) \\
&\quad + \int_{x_i-h/2}^{x_i+h/2} dx \int_{z_k-h/2}^{z_k+h/2} dz \left(J_y\left(x, y_j + \frac{h}{2}, z\right) - J_z\left(x, y_j - \frac{h}{2}, z\right)\right) \\
&\quad + \int_{z_k-h/2}^{z_k+h/2} dz \int_{y_j-h/2}^{y_j+h/2} dy \left(J_x\left(x_i + \frac{h}{2}, y, z\right) - J_z\left(x_i - \frac{h}{2}, y, z\right)\right).
\end{aligned}
\tag{11.84}
$$

The surface integral can be evaluated numerically (Chap. 4). Using the midpoint approximation (11.77) we obtain

$$
\begin{aligned}
\frac{1}{V}\Phi(x_i, y_j, z_k) = \frac{1}{h}\big(&J_z(x_i, y_j, z_{k+1/2}) - J_z(x_i, y_j, z_{k-1/2}) \\
&+ J_y(x_i, y_{j+1/2}, z_k) - J_y(x_i, y_{j-1/2}, z_k) \\
&+ J_x(x_{i+1/2}, y_j, z_k) - J_x(x_{i-1/2}, y_j, z_k)\big).
\end{aligned}
\tag{11.85}
$$

The trapezoidal rule (4.13) introduces an average over the four corners (Fig. 11.3)

$$
\begin{aligned}
&\int_{x_i-h/2}^{x_i+h/2} dx \int_{y_j-h/2}^{y_j+h/2} dy\, f(x, y) \\
&\quad = h^2\left(\frac{1}{4}\sum_{m,n=\pm 1} f(x_{i+m/2}, y_{j+n/2}) + O\left(h^2\right)\right)
\end{aligned}
\tag{11.86}
$$

which replaces the flux values in (11.85) by the averages

$$
J_x(x_{i\pm 1/2}, y_j, z_k) = \frac{1}{4}\sum_{m=\pm 1, n=\pm 1} J_z(x_{i\pm 1/2}, y_{j+m/2}, z_{k+n/2}) \tag{11.87}
$$

$$
J_y(x_i, y_{j\pm 1/2}, z_k) = \frac{1}{4}\sum_{m=\pm 1, n=\pm 1} J_z(x_{i+m/2}, y_{j\pm 1/2}, z_{k+n/2}) \tag{11.88}
$$

$$J_z(x_i, y_j, z_{k\pm 1/2}) = \frac{1}{4} \sum_{m=\pm 1, n=\pm 1} J_z(x_{i+m/2}, y_{j+n/2}, z_{k\pm 1/2}). \tag{11.89}$$

One advantage of the finite volume method is that the flux is strictly conserved.

11.4 Weighted Residual Based Methods

A general method to discretize partial differential equations is to approximate the solution within a finite dimensional space of trial functions.[4] The partial differential equation is turned into a finite system of equations or a finite system of ordinary differential equations if time is treated as a continuous variable. This is the basis of spectral methods which make use of polynomials or Fourier series but also of the very successful finite element methods. Even finite difference methods and finite volume methods can be formulated as weighted residual based methods.

Consider a differential equation[5] on the domain V which is written symbolically with the differential operator $\mathcal{T}$

$$\mathcal{T}\big[u(\mathbf{r})\big] = f(\mathbf{r}) \quad \mathbf{r} \in V \tag{11.90}$$

and corresponding boundary conditions which are expressed with a boundary operator $\mathcal{B}$[6]

$$\mathcal{B}\big[u(\mathbf{r})\big] = g(\mathbf{r}) \quad \mathbf{r} \in \partial V. \tag{11.91}$$

The basic principle to obtain an approximate solution $\tilde{u}(\mathbf{r})$ is to choose a linear combination of expansion functions $N_i(\mathbf{r})$ $i = 1 \ldots r$ as a trial function which fulfills the boundary conditions[7]

$$\tilde{u} = \sum_{i=1}^{r} u_i N_i(\mathbf{r}) \tag{11.92}$$

$$\mathcal{B}\big[\tilde{u}(\mathbf{r})\big] = g(\mathbf{r}). \tag{11.93}$$

In general (11.92) is not an exact solution and the residual

$$R(\mathbf{r}) = \mathcal{T}[\tilde{u}](\mathbf{r}) - f(\mathbf{r}) \tag{11.94}$$

will not be zero throughout the whole domain V. The function $\tilde{u}$ has to be determined such that the residual becomes "small" in a certain sense. To that end weight functions[8] w_j $j = 1 \ldots r$ are chosen to define the weighted residuals

[4] Also called expansion functions.

[5] Generalization to systems of equations is straightforward.

[6] One or more linear differential operators, usually a combination of the function and its first derivatives.

[7] This requirement can be replaced by additional equations for the u_i, for instance with the tau method [195].

[8] Also called test functions.

$$R_j(u_1 \dots u_r) = \int dV\, w_j(\mathbf{r})\big(\mathcal{T}[\tilde{u}](\mathbf{r}) - f(\mathbf{r})\big). \tag{11.95}$$

The optimal parameters u_i are then obtained from the solution of the equations

$$R_j(u_1 \dots u_r) = 0 \quad j = 1 \dots r. \tag{11.96}$$

In the special case of a linear differential operator these equations are linear

$$\sum_{i=1}^{r} u_i \int dV\, w_j(\mathbf{r})\mathcal{T}\big[N_i(\mathbf{r})\big] - \int dV\, w_j(\mathbf{r}) f(\mathbf{r}) = 0. \tag{11.97}$$

Several strategies are available to choose suitable weight functions.

11.4.1 Point Collocation Method

The collocation method uses the weight functions $w_j(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_j)$, with certain collocation points $\mathbf{r}_j \in V$. The approximation $\tilde{u}$ obeys the differential equation at the collocation points

$$0 = R_j = \mathcal{T}[\tilde{u}](\mathbf{r}_j) - f(\mathbf{r}_j) \tag{11.98}$$

and for a linear differential operator

$$0 = \sum_{i=1}^{r} u_i \mathcal{T}[N_i](\mathbf{r}_j) - f(\mathbf{r}_j). \tag{11.99}$$

The point collocation method is simple to use, especially for nonlinear problems. Instead of using trial functions satisfying the boundary conditions, extra collocation points on the boundary can be added (mixed collocation method).

11.4.2 Sub-domain Method

This approach uses weight functions which are the characteristic functions of a set of control volumes V_i which are disjoint and span the whole domain similar as for the finite volume method

$$V = \bigcup_j V_j \qquad V_j \cap V_{j'} = \emptyset \quad \forall j \neq j' \tag{11.100}$$

$$w_j(\mathbf{r}) = \begin{cases} 1 & \mathbf{r} \in V_j \\ 0 & \text{else.} \end{cases} \tag{11.101}$$

The residuals then are integrals over the control volumes and

$$0 = R_j = \int_{V_j} dV\, \big(\mathcal{T}[\tilde{u}](\mathbf{r}) - f(\mathbf{r})\big) \tag{11.102}$$

respectively

$$0 = \sum_i u_i \int_{V_j} dV\, \mathcal{T}[N_i](\mathbf{r}) - \int_{V_j} dV\, f(\mathbf{r}). \tag{11.103}$$

11.4.3 Least Squares Method

Least squares methods have become popular for first order systems of differential equations in computational fluid dynamics and computational electrodynamics [30, 140].

The L2-norm of the residual (11.94) is given by the integral

$$S = \int_V dV\, R(\mathbf{r})^2. \tag{11.104}$$

It is minimized by solving the equations

$$0 = \frac{\partial S}{\partial u_j} = 2 \int_V dV\, \frac{\partial R}{\partial u_j} R(\mathbf{r}) \tag{11.105}$$

which is equivalent to choosing the weight functions

$$w_j(\mathbf{r}) = \frac{\partial R}{\partial u_j} R(\mathbf{r}) = \frac{\partial}{\partial u_j} \mathcal{T}\left[\sum_i u_i N_i(\mathbf{r})\right] \tag{11.106}$$

or for a linear differential operator simply

$$w_j(\mathbf{r}) = \mathcal{T}\big[N_j(\mathbf{r})\big]. \tag{11.107}$$

Advantages of the least squares method are that boundary conditions can be incorporated into the residual and that S provides a measure for the quality of the solution which can be used for adaptive grid size control. On the other hand S involves a differential operator of higher order and therefore much smoother trial functions are necessary.

11.4.4 Galerkin Method

Galerkin's widely used method [87, 98] chooses the basis functions as weight functions

$$w_j(\mathbf{r}) = N_j(\mathbf{r}) \tag{11.108}$$

and solves the following system of equations

$$\int dV\, N_j(\mathbf{r}) \mathcal{T}\left[\sum_i u_i N_i(\mathbf{r})\right] - \int dV\, N_j(\mathbf{r}) f(\mathbf{r}) = 0 \tag{11.109}$$

or in the simpler linear case

$$\sum u_i \int_V dV\, N_j(\mathbf{r}) \mathcal{T}\big(N_i(\mathbf{r})\big) = \int_V dV\, N_j(\mathbf{r}) f(\mathbf{r}, t). \tag{11.110}$$

11.5 Spectral and Pseudo-spectral Methods

Spectral methods use basis functions which are nonzero over the whole domain, the trial functions being mostly polynomials or Fourier sums [35]. They can be used to solve ordinary as well as partial differential equations. The combination of a spectral method with the point collocation method is also known as pseudo-spectral method.

11.5.1 Fourier Pseudo-spectral Methods

Linear differential operators become diagonal in Fourier space. Combination of Fourier series expansion and point collocation leads to equations involving a discrete Fourier transformation, which can be performed very efficiently with the Fast Fourier Transform methods.

For simplicity we consider only the one-dimensional case. We choose equidistant collocation points

$$x_m = m\Delta x \quad m = 0, 1 \ldots M - 1 \tag{11.111}$$

and expansion functions

$$N_j(x) = \mathrm{e}^{\mathrm{i}k_j x} \quad k_j = \frac{2\pi}{M\Delta x} j \quad j = 0, 1 \ldots M - 1. \tag{11.112}$$

For a linear differential operator

$$\mathcal{L}\big[\mathrm{e}^{\mathrm{i}k_j x}\big] = l(k_j)\mathrm{e}^{\mathrm{i}k_j x} \tag{11.113}$$

and the condition on the residual becomes

$$0 = R_m = \sum_{j=0}^{M-1} u_j l(k_j)\mathrm{e}^{\mathrm{i}k_j x_m} - f(x_m) \tag{11.114}$$

or

$$f(x_m) = \sum_{j=0}^{M-1} u_j l(k_j)\mathrm{e}^{\mathrm{i}2\pi mj/M} \tag{11.115}$$

which is nothing but a discrete Fourier back transformation (Sect. 7.2, (7.19)) which can be inverted to give

$$u_j l(k_j) = \frac{1}{N} \sum_{m=0}^{M-1} f(x_m)\mathrm{e}^{-\mathrm{i}2\pi mj/M}. \tag{11.116}$$

Instead of exponential expansion functions, sine and cosine functions can be used to satisfy certain boundary conditions, for instance to solve the Poisson equation within a cube (Sect. 17.1.2).

11.5.2 Example: Polynomial Approximation

Let us consider the initial value problem (Fig. 11.5)

$$\frac{\mathrm{d}}{\mathrm{d}x}u(x) - u(x) = 0 \quad u(0) = 1 \quad \text{for } 0 \le x \le 1 \tag{11.117}$$

with the well known solution

$$u(x) = \mathrm{e}^x. \tag{11.118}$$

We choose a polynomial trial function with the proper initial value

$$\tilde{u}(x) = 1 + u_1 x + u_2 x^2. \tag{11.119}$$

The residual is

$$\begin{aligned} R(x) &= u_1 + 2u_2 x - \left(1 + u_1 x + u_2 x^2\right) \\ &= (u_1 - 1) + (2u_2 - u_1)x - u_2 x^2. \end{aligned} \tag{11.120}$$

11.5.2.1 Point Collocation Method

For our example we need two collocation points to obtain two equations for the two unknowns $u_{1,2}$. We choose $x_1 = 0, x_2 = \frac{1}{2}$. Then we have to solve the equations

$$R(x_1) = u_1 - 1 = 0 \tag{11.121}$$

$$R(x_2) = \frac{1}{2}u_1 + \frac{3}{4}u_2 - 1 = 0 \tag{11.122}$$

which gives

$$u_1 = 1 \quad u_2 = \frac{2}{3} \tag{11.123}$$

$$u_c = 1 + x + \frac{2}{3}x^2. \tag{11.124}$$

11.5.2.2 Sub-domain Method

We need two sub-domains to obtain two equations for the two unknowns $u_{1,2}$. We choose $V_1 = \{x, 0 < x < \frac{1}{2}\}$, $V_2 = \{x, \frac{1}{2} < x < 1\}$. Integration gives

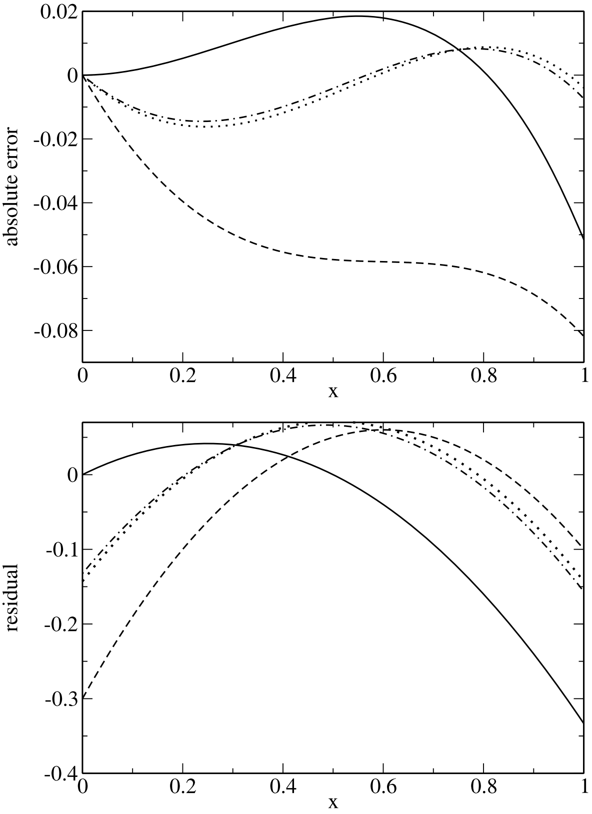


Fig. 11.5 (Approximate solution of a simple differential equation) The initial value problem $\frac{\mathrm{d}}{\mathrm{d}x}u(x) - u(x) = 0$ $u(0) = 1$ for $0 \le x \le 1$ is approximately solved with a polynomial trial function $\tilde{u}(x) = 1 + u_1 x + u_2 x^2$. The parameters $u_{1,2}$ are optimized with the method of weighted residuals using point collocation (*full curve*), sub-domain collocation (*dotted curve*), Galerkin's method (*dashed curve*) and least squares (*dash-dotted curve*). The absolute error $\tilde{u}(x) - \mathrm{e}^x$ (*top*) and the residual $R(x) = \frac{\mathrm{d}}{\mathrm{d}x}\tilde{u}(x) - \tilde{u}(x) = (u_1 - 1) + (2u_2 - u_1)x - u_2 x^2$ both are smallest for the least squares and sub-domain collocation methods

$$R_1 = \frac{3}{8}u_1 + \frac{5}{24}u_2 - \frac{1}{2} = 0 \tag{11.125}$$

$$R_2 = \frac{1}{8}u_1 + \frac{11}{24}u_2 - \frac{1}{2} = 0 \tag{11.126}$$

$$u_1 = u_2 = \frac{6}{7} \tag{11.127}$$

$$u_{sdc} = 1 + \frac{6}{7}x + \frac{6}{7}x^2. \tag{11.128}$$

11.5.2.3 Galerkin Method

Galerkin's method uses the weight functions $w_1(x) = x$, $w_2(x) = x^2$. The equations

$$\int_0^1 dx\, w_1(x)R(x) = \frac{1}{6}u_1 + \frac{5}{12}u_2 - \frac{1}{2} = 0 \tag{11.129}$$

$$\int_0^1 dx\, w_2(x)R(x) = \frac{1}{12}u_1 + \frac{3}{10}u_2 - \frac{1}{3} = 0 \tag{11.130}$$

have the solution

$$u_1 = \frac{8}{11} \quad u_2 = \frac{10}{11} \tag{11.131}$$

$$u_G = 1 + \frac{8}{11}x + \frac{10}{11}x^2. \tag{11.132}$$

11.5.2.4 Least Squares Method

The integral of the squared residual

$$S = \int_0^1 dx\, R(x)^2 = 1 - u_1 - \frac{4}{3}u_2 + \frac{1}{3}u_1^2 + \frac{1}{2}u_1 u_2 + \frac{8}{15}u_2^2 \tag{11.133}$$

is minimized by solving

$$\frac{\partial S}{\partial u_1} = \frac{2}{3}u_1 + \frac{1}{2}u_2 - 1 = 0 \tag{11.134}$$

$$\frac{\partial S}{\partial u_2} = \frac{1}{2}u_1 + \frac{16}{15}u_2 - \frac{4}{3} = 0 \tag{11.135}$$

which gives

$$u_1 = \frac{72}{83} \quad u_2 = \frac{70}{83} \tag{11.136}$$

$$u_{LS} = 1 + \frac{72}{83}x + \frac{70}{83}x^2. \tag{11.137}$$

11.6 Finite Elements

The method of finite elements (FEM) is a very flexible method to discretize partial differential equations [84, 210]. It is rather dominant in a variety of engineering sciences. Usually the expansion functions N_i are chosen to have compact support. The integration volume is divided into disjoint sub-volumes

$$V = \bigcup_{i=1}^{r} V_i \qquad V_i \cap V_{i'} = \emptyset \forall i \neq i'. \tag{11.138}$$

The $N_i(\mathbf{x})$ are piecewise continuous polynomials which are nonzero only inside V_i and a few neighbor cells.

11.6.1 One-Dimensional Elements

In one dimension the domain is an interval $V = \{x; a \leq x \leq b\}$ and the sub-volumes are small sub-intervals $V_i = \{x; x_i \leq x \leq x_{i+1}\}$. The one-dimensional mesh is the

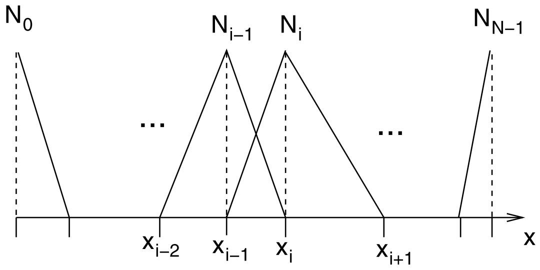


Fig. 11.6 (Finite elements in one dimension) The basis functions N_i are piecewise continuous polynomials and have compact support. In the simplest case they are composed of two linear functions over the sub-intervals $x_{i-1} \le x \le x_i$ and $x_i \le x \le x_{i+1}$

set of nodes $\{a = x_0, x_1 \ldots x_r = b\}$. Piecewise linear basis functions (Fig. 11.6) are in the 1-dimensional case given by

$$N_i(x) = \begin{cases} \frac{x_{i+1}-x}{x_{i+1}-x_i} & \text{for } x_i < x < x_{i+1} \\ \frac{x-x_{i-1}}{x_i-x_{i-1}} & \text{for } x_{i-1} < x < x_i \\ 0 & \text{else} \end{cases} \tag{11.139}$$

and the derivatives are (except at the nodes x_i)

$$N_i'(x) = \begin{cases} -\frac{1}{x_{i+1}-x_i} & \text{for } x_i < x < x_{i+1} \\ \frac{1}{x_i-x_{i-1}} & \text{for } x_{i-1} < x < x_i \\ 0 & \text{else.} \end{cases} \tag{11.140}$$

11.6.2 Two- and Three-Dimensional Elements

In two dimensions the mesh is defined by a finite number of points $(x_i, y_i) \in V$ (the nodes of the mesh). There is considerable freedom in the choice of these points and they need not be equally spaced.

11.6.2.1 Triangulation

The nodes can be regarded as forming the vertices of a triangulation[9] of the domain V (Fig. 11.7).

The piecewise linear basis function in one dimension (11.139) can be generalized to the two-dimensional case by constructing functions $N_i(x, y)$ which are zero at all nodes except (x_i, y_i)

$$N_i(x_j, y_j) = \delta_{i,j}. \tag{11.141}$$

[9] The triangulation is not determined uniquely by the nodes.

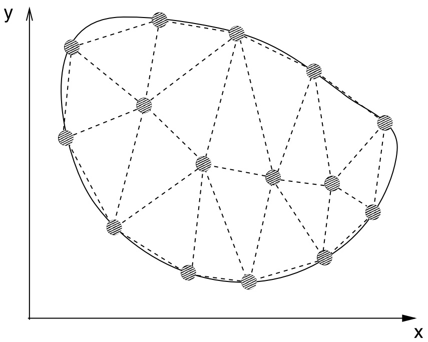


Fig. 11.7 (Triangulation of a two-dimensional domain) A two-dimensional mesh is defined by a set of node points which can be regarded to form the vertices of a triangulation

These functions are linear over each triangle which contains the vertex i and can be combined as the sum of small pyramids (Fig. 11.8). Let one of the triangles be denoted by its three vertices as T_{ijk}.[10] The corresponding linear function then is

$$n_{ijk}(x,y)=\alpha+\beta_x(x-x_i)+\beta_y(y-y_i) \tag{11.142}$$

where the coefficients follow from the conditions

$$n_{ijk}(x_i,y_i)=1 \quad n_{ijk}(x_j,y_j)=n_{ijk}(x_k,y_k)=0 \tag{11.143}$$

as

$$\alpha=1 \quad \beta_x=\frac{y_j-y_k}{2A_{ijk}} \quad \beta_y=\frac{x_k-x_j}{2A_{ijk}} \tag{11.144}$$

with

$$A_{ijk}=\frac{1}{2}\det\begin{vmatrix} x_j-x_i & x_k-x_i \\ y_j-y_i & y_k-y_i \end{vmatrix} \tag{11.145}$$

which, apart from sign, is the area of the triangle T_{ijk}. The basis function N_i now is given by

$$N_i(x,y)=\begin{cases} n_{ijk}(x,y) & (x,y)\in T_{ijk} \\ 0 & \text{else.} \end{cases}$$

In three dimensions we consider tetrahedrons (Fig. 11.9) instead of triangles. The corresponding linear function of three arguments has the form

$$n_{i,j,k,l}(x,y,z)=\alpha+\beta_x(x-x_i)+\beta_y(y-y_i)+\beta_z(z-z_i) \tag{11.146}$$

and from the conditions $n_{i,j,k,l}(x_i,y_i,z_i)=1$ and $n_{i,j,k,l}=0$ on all other nodes we find (an algebra program is helpful at that point)

[10] The order of the indices does matter.

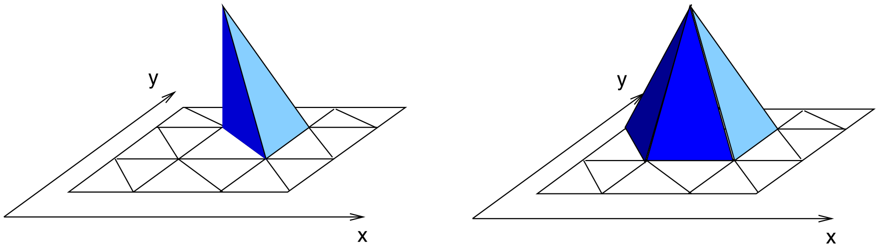
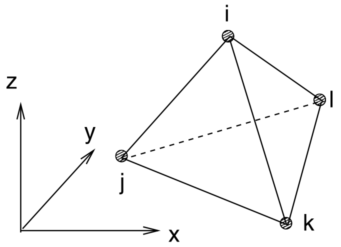


Fig. 11.8 (Finite elements in two dimensions) The simplest finite elements in two dimensions are piecewise linear functions $N_i(x, y)$ which are non-vanishing only at one node (x_i, y_i) (*right side*). They can be constructed from small pyramids built upon one of the triangles that contains this node (*left side*)

Fig. 11.9 (Tetrahedron) The tetrahedron is the three-dimensional case of a Euclidean simplex, i.e. the simplest polytope



$$\begin{aligned}
\alpha &= 1 \\
\beta_x &= \frac{1}{6V_{ijkl}} \det \begin{vmatrix} y_k - y_j & y_l - y_j \\ z_k - z_j & z_l - z_j \end{vmatrix} \\
\beta_y &= \frac{1}{6V_{ijkl}} \det \begin{vmatrix} z_k - z_j & z_l - z_j \\ x_k - x_j & x_l - x_j \end{vmatrix} \\
\beta_z &= \frac{1}{6V_{ijkl}} \det \begin{vmatrix} x_k - x_j & x_l - x_j \\ y_k - y_j & y_l - y_j \end{vmatrix}
\end{aligned} \tag{11.147}$$

where V_{ijkl} is, apart from sign, the volume of the tetrahedron

$$V_{ijkl} = \frac{1}{6} \det \begin{vmatrix} x_j - x_i & x_k - x_i & x_l - x_i \\ y_j - y_i & y_k - y_i & y_l - y_i \\ z_j - z_i & z_k - z_i & z_l - z_i \end{vmatrix} . \tag{11.148}$$

11.6.2.2 Rectangular Elements

For a rectangular grid rectangular elements offer a practical alternative to triangles. Since equations for four nodes have to be fulfilled, the basic element needs four parameters, which is the case for a bilinear expression. Let us denote one of the rectangles which contains the vertex i as $R_{i,j,k,l}$. The other three edges are

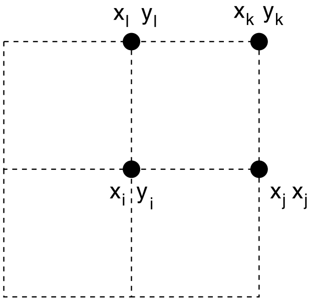


Fig. 11.10 (Rectangular elements around one vertex) The basis function N_i is a bilinear function on each of the four rectangles containing the vertex (x_i, y_i)

$$(x_j, y_j) = (x_i + b_x, y_i) \quad (x_k, y_k) = (x_i, y_i + b_y) \quad (x_l, y_l) = (x_i + b_x, y_i + b_y) \tag{11.149}$$

where $b_x = \pm h_x, b_y = \pm h_y$ corresponding to the four rectangles with the common vertex i (Fig. 11.10).

The bilinear function (Fig. 11.11) corresponding to R_{ijkl} is

$$n_{i,j,k,l}(x, y) = \alpha + \beta(x - x_i) + \gamma(y - y_i) + \eta(x - x_i)(y - y_i). \tag{11.150}$$

It has to fulfill

$$n_{i,j,k,l}(x_i, y_i) = 1 \quad n_{i,j,k,l}(x_j, y_j) = n_{i,j,k,l}(x_k, y_k) = n_{i,j,k,l}(x_l, y_l) = 0 \tag{11.151}$$

from which we find

$$\alpha = 1 \quad \beta = -\frac{1}{b_x} \quad \gamma = -\frac{1}{b_y} \quad \eta = \frac{1}{b_x b_y} \tag{11.152}$$

$$n_{i,j,k,l}(x, y) = 1 - \frac{x - x_i}{b_x} - \frac{y - y_i}{b_y} + \frac{(x - x_i)}{b_x}\frac{(y - y_i)}{b_y}. \tag{11.153}$$

The basis function centered at node i then is

$$N_i(x, y) = \begin{cases} n_{i,j,k,l}(x, y) & (x, y) \in R_{i,j,k,l} \\ 0 & \text{else.} \end{cases} \tag{11.154}$$

Generalization to a three-dimensional grid is straightforward (Fig. 11.12). We denote one of the eight cuboids containing the node (x_i, y_i, z_i) as $C_{i,j_1 \ldots j_7}$ with $(x_{j_1}, y_{j_1}, z_{j_1}) = (x_i + b_x, y_i, z_i) \ldots (x_{j_7}, y_{j_7}, z_{j_7}) = (x_i + b_x, y_i + b_y, z_i + b_z)$. The corresponding trilinear function is

$$\begin{aligned} n_{i,j_1 \ldots j_7} = 1 &- \frac{x - x_i}{b_x} - \frac{y - y_i}{b_y} - \frac{z - z_i}{b_z} \\ &+ \frac{(x - x_i)}{b_x}\frac{(y - y_i)}{b_y} + \frac{(x - x_i)}{b_x}\frac{(z - z_i)}{b_z} + \frac{(z - z_i)}{b_z}\frac{(y - y_i)}{b_y} \\ &- \frac{(x - x_i)}{b_x}\frac{(y - y_i)}{b_y}\frac{(z - z_i)}{b_z}. \end{aligned} \tag{11.155}$$

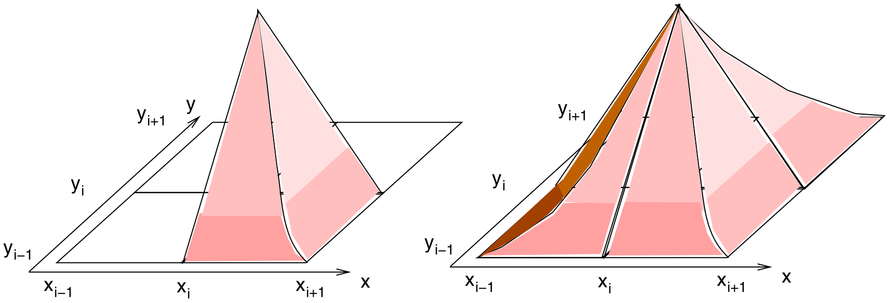
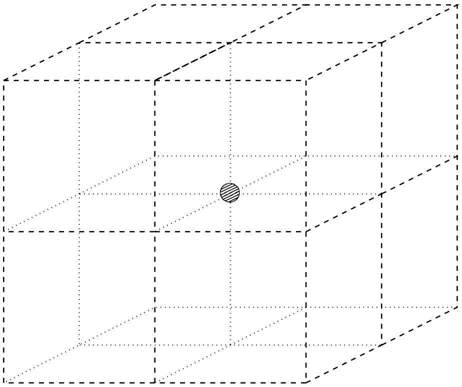


Fig. 11.11 (Bilinear elements on a rectangular grid) The basis functions $N_i(x, y)$ on a rectangular grid (*right side*) are piecewise bilinear functions (*left side*), which vanish at all nodes except (x_i, y_i) (*right side*)

Fig. 11.12 (Three-dimensional rectangular grid) The basis function N_i is trilinear on each of the eight cuboids containing the vertex i. It vanishes on all nodes except (x_i, y_i, z_i)



11.6.3 One-Dimensional Galerkin FEM

As an example we consider the one-dimensional linear differential equation (11.5)

$$\left(a\frac{\partial^2}{\partial x^2}+b\frac{\partial}{\partial x}+c\right)u(x)=f(x) \tag{11.156}$$

in the domain $0 \le x \le 1$ with boundary conditions

$$u(0)=u(1)=0. \tag{11.157}$$

We use the basis functions from (11.139) on a one-dimensional grid with

$$x_{i+1}-x_i=h_i \tag{11.158}$$

and apply the Galerkin method [88]. The boundary conditions require

$$u_0=u_{N-1}=0. \tag{11.159}$$

The weighted residual is

$$0 = R_j = \sum_i u_i \int_0^1 dx\, N_j(x)\left(a\frac{\partial^2}{\partial x^2} + b\frac{\partial}{\partial x} + c\right)N_i(x) - \int_0^1 dx\, N_j(x)f(x). \tag{11.160}$$

First we integrate

$$\int_0^1 N_j(x)N_i(x)\,dx = \int_{x_{i-1}}^{x_{i+1}} N_j(x)N_i(x)\,dx = \begin{cases} \frac{h_i+h_{i-1}}{3} & j=i \\ \frac{h_i}{6} & j=i+1 \\ \frac{h_{i-1}}{6} & j=i-1 \\ 0 & |i-j|>1. \end{cases} \tag{11.161}$$

Integration of the first derivative gives

$$\int_0^1 dx\, N_j(x)N_i'(x) = \begin{cases} 0 & j=i \\ \frac{1}{2} & j=i-1 \\ -\frac{1}{2} & j=i+1 \\ 0 & \text{else.} \end{cases} \tag{11.162}$$

For the second derivative partial integration gives

$$\int_0^1 dx\, N_j(x)\frac{\partial^2}{\partial x^2}N_i(x)$$
$$= N_j(1)N_i'(1-\varepsilon) - N_j(0)N_i'(0+\varepsilon) - \int_0^1 dx\, N_j'(x)N_i'(x) \tag{11.163}$$

where the first two summands are zero due to the boundary conditions. Since N_i and N_i' are nonzero only for $x_{i-1} < x < x_{i+1}$ we find

$$\int_0^1 dx\, N_j(x)\frac{\partial^2}{\partial x^2}N_i(x) = -\int_{x_{i-1}}^{x_{i+1}} dx\, N_j'(x)N_i'(x)$$
$$= \begin{cases} \frac{1}{h_{i-1}} & j=i-1 \\ -\frac{1}{h_i} - \frac{1}{h_{i-1}} & i=j \\ \frac{1}{h_i} & j=i+1 \\ 0 & \text{else.} \end{cases} \tag{11.164}$$

Integration of the last term in (11.160) gives

$$\int_0^1 dx\, N_j(x)f(x) = \int_{x_{i-1}}^{x_{i+1}} dx\, N_j(x)f(x)$$
$$= \int_{x_{j-1}}^{x_j} dx\, \frac{x-x_{j-1}}{x_j-x_{j-1}}f(x)$$
$$+ \int_{x_j}^{x_{j+1}} dx\, \frac{x_{j+1}-x}{x_{j+1}-x_j}f(x). \tag{11.165}$$

Applying the trapezoidal rule[11] for both integrals we find

$$\int_{x_{j-1}}^{x_{j+1}} dx\, N_j(x) f(x) \approx f(x_j)\frac{h_j + h_{j-1}}{2}. \tag{11.166}$$

The discretized equation finally reads

$$\begin{aligned} & a\left\{\frac{1}{h_{j-1}}u_{j-1} - \left(\frac{1}{h_j} + \frac{1}{h_{j-1}}\right)u_j + \frac{1}{h_j}u_{j+1}\right\} \\ & \quad + b\left\{-\frac{1}{2}u_{j-1} + \frac{1}{2}u_{j+1}\right\} \\ & \quad + c\left\{\frac{h_{j-1}}{6}u_{j-1} + \frac{h_j + h_{j-1}}{3}u_j + \frac{h_j}{6}u_{j+1}\right\} \\ & = f(x_j)\frac{h_j + h_{j-1}}{2} \end{aligned} \tag{11.167}$$

which can be written in matrix notation as

$$A\mathbf{u} = B\mathbf{f} \tag{11.168}$$

where the matrix A is tridiagonal as a consequence of the compact support of the basis functions

$$\begin{aligned} A = a & \begin{pmatrix} -\frac{1}{h_1} - \frac{1}{h_0}, & \frac{1}{h_1} & & & \\ & \ddots & & & \\ & \frac{1}{h_{j-1}}, & -\frac{1}{h_j} - \frac{1}{h_{j-1}}, & \frac{1}{h_j} & \\ & & & \ddots & \\ & & & \frac{1}{h_{N-3}}, & -\frac{1}{h_{N-2}} - \frac{1}{h_{N-3}} \end{pmatrix} \\ + b & \begin{pmatrix} 0, & \frac{1}{2} & & & \\ & \ddots & & & \\ & -\frac{1}{2}, & 0, & \frac{1}{2} & \\ & & & \ddots & \\ & & & -\frac{1}{2}, & 0 \end{pmatrix} \\ + c & \begin{pmatrix} \frac{(h_1+h_0)}{3}, & \frac{h_1}{6} & & & \\ & \ddots & & & \\ & \frac{h_{j-1}}{6}, & \frac{(h_j+h_{j-1})}{3}, & \frac{h_j}{6} & \\ & & & \ddots & \\ & & & \frac{h_{N-3}}{6}, & \frac{(h_{N-2}+h_{N-3})}{3} \end{pmatrix} \end{aligned} \tag{11.169}$$

[11] Higher accuracy can be achieved, for instance, by Gaussian integration.

$$B = \begin{pmatrix} \frac{h_0+h_1}{2} & & & & \\ & \ddots & & & \\ & & \frac{h_{j-1}+h_j}{2} & & \\ & & & \ddots & \\ & & & & \frac{h_{N-2}+h_{N-3}}{2} \end{pmatrix}.$$

For equally spaced nodes $h_i = h_{i-1} = h$ and after division by h (11.167) reduces to a system of equations where the derivatives are replaced by finite differences (11.20)

$$\begin{aligned} a & \left\{ \frac{1}{h^2} u_{j-1} - \frac{2}{h^2} u_j + \frac{1}{h^2} u_{j+1} \right\} \\ & + b \left\{ -\frac{1}{2h} u_{j-1} + \frac{1}{2h} u_{j+1} \right\} \\ & + c \left\{ \frac{1}{6} u_{j-1} + \frac{2}{3} u_j + \frac{1}{6} u_{j+1} \right\} \\ = {} & f(x_j) \end{aligned} \tag{11.170}$$

and the function u is replaced by a certain average

$$\frac{1}{6} u_{j-1} + \frac{2}{3} u_j + \frac{1}{6} u_{j+1} = u_j + \frac{1}{6}(u_{j-1} - 2u_j + u_{j+1}). \tag{11.171}$$

The corresponding matrix in (11.169) is the so called mass matrix. Within the framework of the finite differences method the last expression equals

$$u_j + \frac{1}{6}(u_{j-1} - 2u_j + u_{j+1}) = u_j + \frac{h^2}{6} \left(\frac{d^2u}{dx^2} \right)_j + O\big(h^4\big) \tag{11.172}$$

hence replacing it by u_j (this is called mass lumping) introduces an error of the order $O(h^2)$.

11.7 Boundary Element Method

The boundary element method (BEM) [18, 276] is a method for linear partial differential equations which can be brought into boundary integral form[12] like Laplace's equation (Chap. 17)[13]

$$-\Delta \Phi(\mathbf{r}) = 0 \tag{11.173}$$

for which the fundamental solution

$$\Delta G\big(\mathbf{r}, \mathbf{r}'\big) = -\delta\big(\mathbf{r} - \mathbf{r}'\big)$$

[12] This is only possible if the fundamental solution or Green's function is available.

[13] The minus sign is traditionally used.

is given by

$$G(\mathbf{r}-\mathbf{r}') = \frac{1}{4\pi|\mathbf{r}-\mathbf{r}'|} \quad \text{in three dimensions} \tag{11.174}$$

$$G(\mathbf{r}-\mathbf{r}') = \frac{1}{2\pi}\ln\frac{1}{|\mathbf{r}-\mathbf{r}'|} \quad \text{in two dimensions.} \tag{11.175}$$

We apply Gauss's theorem to the expression [277]

$$\begin{aligned}&\operatorname{div}\left[G(\mathbf{r}-\mathbf{r}')\operatorname{grad}(\Phi(\mathbf{r})) - \Phi(\mathbf{r})\operatorname{grad}(G(\mathbf{r}-\mathbf{r}'))\right]\\ &\quad = -\Phi(\mathbf{r})\Delta(G(\mathbf{r}-\mathbf{r}')).\end{aligned} \tag{11.176}$$

Integration over a volume V gives

$$\begin{aligned}&\oint_{\partial V} dA\left(G(\mathbf{r}-\mathbf{r}')\frac{\partial}{\partial n}(\Phi(\mathbf{r})) - \Phi(\mathbf{r})\frac{\partial}{\partial n}(G(\mathbf{r}-\mathbf{r}'))\right)\\ &\quad = -\int_V dV\,(\Phi(\mathbf{r})\Delta(G(\mathbf{r}-\mathbf{r}'))) = \Phi(\mathbf{r}').\end{aligned} \tag{11.177}$$

This integral equation determines the potential self-consistently by its value and normal derivative on the surface of the cavity. It can be solved numerically by dividing the surface into a finite number of boundary elements. The resulting system of linear equations often has smaller dimension than corresponding finite element approaches. However, the coefficient matrix is in general full and not necessarily symmetric.