1

Getting Started in One Space Variable

A journey of a thousand miles begins with a single step.

Lao Tzu

In this chapter, we give an initial presentation of the three numerical methods to be studied throughout the book, as the most popular to solve PDEs: the finite difference method (**FDM**), the finite element method (**FEM**) and the finite volume method (**FVM**). The first type of method acts directly on the differential equation in its standard form, also called the equation's **strong form**. In contrast, both the FEM and the FVM are applied to integral forms of the equation called a **weak form** or a **variational form** for the former, and a **conservative form** for the latter. From this point of view, the FVM is viewed by some authors as a variant of the FEM, since it corresponds to a particular way of solving the problem, once it is rewritten in a suitable integral form.

From the material provided in this chapter, the reader will certainly figure out that, as long as the points of the **discretisation lattice** at which solution approximations are determined coincide for the three approaches, and moreover the problem to solve is sufficiently simple in terms of geometry, data, nonlinearities and so on, the underlying SLAEs are very similar. As far as the FEM and the FVM are concerned, for example, this has been observed by several authors for years (see e.g. Idelsohn and Oñate [102]). Nevertheless, the three methods are conceptually different, and for this reason it is worth distinguishing them from each other whatever the case.

In the subsequent chapters, we shall elaborate a little more about the concept of discretisation lattices and the relationship of a numerical method to them in more general frameworks. For the moment, we will do this in the case where the solution depends on a single variable x. More specifically, we will set ourselves in the simplest possible framework of a model two-point boundary value ODE with zero (i.e. homogeneous) boundary conditions, which will serve as a model for introducing the principles that the FDM, FEM and FVM are based upon. In doing so, a first attempt will be made to provide some insight on the right choice of techniques for the computation of the numerical solution, in more challenging situations encountered in practice.

Chapter outline: In Section 1.1, the model ODE is introduced, together with features of its three different formulations to be exploited in the sequel. In Sections 1.2, 1.3 and 1.4, we describe the basic FDM, FEM and FVM, respectively, to solve this model problem. Section 1.4 is subdivided into three subsections. In Subsection 1.4.1, we describe a version of the FVM strongly connected to the two other methods; in Subsection 1.4.2, another version of the FVM is presented in the usual way for this method; and, in Subsection 1.4.3, some connections between the FVM and the FEM are highlighted. In Section 1.5, we extend the methods' application considered in the previous sections to the case of inhomogeneous (i.e. nonzero) boundary conditions. In Section 1.6, the solution of SLAEs resulting from the application of the three methods to the model problem is addressed, and a numerical example is given.

1.1 A Model Two-Point Boundary Value Problem

Let us consider the elongational deformations of an elastic bar of length L, having one of its ends fixed, say the left end, and the other end free. The bar is subject to a distribution of forces acting along its length, represented by a function $g(x), 0 \leq x \leq L$. Referring to Figure 1.1, we denote by $u(x)$ the resulting length variation of the elastic bar at a point x, or, equivalently, the displacement undergone in the longitudinal direction $\vec{e}_x$ at this point by $u(x)\vec{e}_x$.

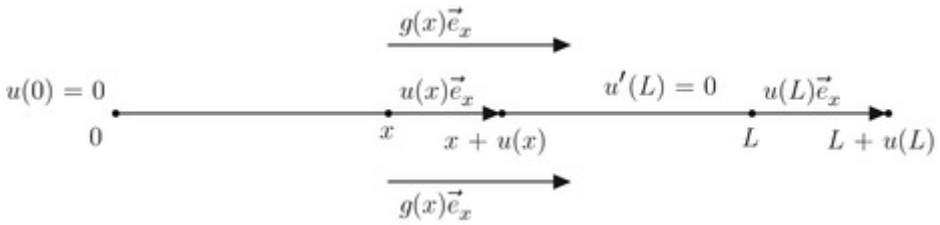

Figure 1.1 Elongation $u(x)$ for a longitudinal loading $g(x)$.

Provided the deformations of the elastic bar take place in the small strain regime[1], the physical problem just described can be set in the form of the following two-point boundary value ODE for the unknown function $u(x)$ (see e.g. [181]):

$$\begin{cases} -(pu')' = g & \text{in } (0, L) \\ u(0) = 0 & \text{(fixed left end)} \\ u'(L) = 0 & \text{(free right end)} \end{cases} \qquad (P'_1)$$

where p is a strictly positive function representing the local stiffness of the material the elastic bar is made of, and u' denotes $\frac{du}{dx}$.

In some situations, spring-like effects on the bar have to be taken into account. This means that a term of the form qu must be added to the left side of (P'_1) in order to properly model its deformation, q being a given non-negative function representing the spring rate. This in turn may be associated with another given force F. Representing by f the conjugate action of g and F, this leads to the following equation:

The model two-point boundary value problem (P_1)

$$\begin{array}{|ll|}\hline -(pu')' + qu = f & \text{in } (0,L) \\ u(0) = 0 & \text{(fixed left end)} \\ u'(L) = 0 & \text{(free right end)} \\ \hline \end{array}$$

We shall overlook a more elaborate presentation of this elastic bar model, for our goal here is only a comprehensive numerical study of equation (P_1). We refer to [186] for a fine description of the problem's physical modelling in more detail[2].

Actually equation (P_1), eventually with different boundary conditions, is a simplified mathematical model of several other phenomena or processes. In Chapter 7, we consider one of them (cf. Example 7.4).

Whatever real-life problem it models, the differential equation (P_1) is often employed by authors as a basic model to introduce numerical methods for PDEs (see. e.g. [119] and [186], among many others). As already pointed out, this is also what we do in the remainder of this chapter. (P_1) is called the equation's **strong form**. This means that the differential equation is to be regarded in the usual pointwise sense, which applies at least under the assumption that the functions p and q, representing the medium's physical characteristics, satisfy for certain constants α, β, A and B:

$$A \geq p(x) \geq \alpha > 0 \quad \text{and } B \geq q(x) \geq \beta \geq 0, \forall x \in [0, L] \tag{1.1}$$

Now without caring about regularity assumptions, neither on the data p, q and f, nor on the solution u of (P_1), we shall rewrite this problem in two different integral forms. We only assume that all the operations performed for this purpose are feasible and well defined, postponing or leaving to texts on analysis of differential equations the discussion about the conditions under which it is really so.

Let us first consider the **conservative form** of (P_1) relying upon the set $\mathcal{V}$ consisting of all the open subdomains of Ω, generically denoted by ω, that are connected and have a non-zero measure. More concretely, any $\omega \in \mathcal{V}$ is necessarily an open interval contained in $(0, L)$,

including of course this interval itself. Clearly enough, the integral in any $\omega \in \mathcal{V}$ of both sides of the differential equation in (P_1) coincides, that is, this problem implies that the following relations are satisfied:

$$\begin{cases} \int_{a_\omega}^{b_\omega} [-(pu')' + qu]dx = \int_{a_\omega}^{b_\omega} fdx \quad \forall \omega \in \mathcal{V} \\ u(0) = 0 \qquad\qquad\qquad\qquad\quad \text{(fixed left end)} \\ u'(L) = 0 \qquad\qquad\qquad\qquad\quad \text{(free right end)} \end{cases} \tag{P'_2}$$

where a_ω and b_ω denote the left and right end of ω, respectively.

Conversely, intuitively enough one can infer that problem (P'_2) implies (P_1). As a matter of fact, this can be rigorously established by means of the theory of Lebesgue integration (cf. [206] or [111]). In short, both problems are equivalent.

We may further develop the integral relation in (P'_2), in order to obtain a new integral-differential equation also equivalent to (P_1), namely:

The conservative form of model problem (P_1) (P_2)

$$\begin{array}{l} [pu'](a_\omega) - [pu'](b_\omega) + \int_{a_\omega}^{b_\omega} qudx = \int_{a_\omega}^{b_\omega} fdx \quad \forall \omega \in \mathcal{V}. \\ u(0) = 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{(fixed left end)} \\ u'(L) = 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{(free right end)} \end{array}$$

Problem (P_2) corresponds to the usual form the FVM is applied to.

Let us next consider the usual **weak form** of equation (P_1), also known as its **standard Galerkin variational form**. First, we multiply both sides of the differential equation with a **test function** v whose required properties will be specified in due course. For the moment, let us just say that the expression 'test-function' means that v is supposed to behave somehow like the solution u, in the sense that u itself could be one of such functions. In simpler terms, we say that v sweeps a set V – actually, a vector space consisting of functions – whose properties are specified hereafter.

Next, integrating the resulting relation over the interval $(0, L)$ and using integration by parts, we obtain

$$-pu'v\Big|_0^L + \int_0^L pu'v'dx + \int_0^L quvdx = \int_0^L fvdx$$

for every (test) function v.

First of all, we note that since $u'(L) = 0$, if we require that v vanish at the origin like u itself, we come up with

The (standard Galerkin) variational form of (P_1) (P3)

$$\boxed{\int_0^L pu'v'dx + \int_0^L quvdx = \int_0^L fvdx,\ \forall v \in V}$$

where u also belongs to V. Notice that all the integrals above must carry a meaning. Hence, it is necessary to require that f together with every function in V and its first-order derivative are square integrable in the sense of Lebesgue in the interval $(0, L)$, as we will see in Section 2.1. This implies that we are actually defining V to be

$$V = \{v \Big/ v, v' \in \mathcal{L}^2(0, L), v(0) = 0\} \tag{1.2}$$

where $\mathcal{L}^2(0, L)$ is the set of those functions f such that f^2 has a finite (Lebesgue) integral in $(0, L)$. We can assert that vector space V defined by equation (1.2) is the ideal choice in the framework of the standard Galerkin variational formulation (P_3), even though in many cases the solution u satisfies additional boundary conditions and has finer differentiability properties. This is because with such a choice we can deal with the widest possible class of problems carrying a physical meaning, modelled by equation (P_1), as seen in this chapter. A justification of the fact that this choice is also the right one from the mathematical point of view lies beyond the scope of this book. For a comprehensive discussion on this point, the author refers to [182].

Remark 1.1

As proven in classical books on Lebesgue integration (see e.g. [206]), every v such that $v' \in \mathcal{L}^2(0, L)$ is continuous[3]*. Therefore, there is no contradiction in prescribing $v(0) = 0$. For discussions on the conditions under which the above variational formulation (P_3) carries a precise mathematical sense, we refer for instance to reference [136]. For mathematical tools supporting this theory, we refer to reference [179].*

Now what have we gained by writing problem (P_1) in the conservative form (P_2) or in the variational form (P_3)?

First of all, we note that the highest derivative order that appears in both (P_2) and (P_3) is one, while it is two in (P_1). Thus if, for instance, we choose to approximate u by a function belonging to the class of piecewise polynomials like the FEM does, it is readily seen that ordinary continuous functions may be used for (P_3), whereas continuously differentiable ones would be required for (P_1). This is because it is not possible to differentiate twice a function that has discontinuous first-order derivatives at certain points in $(0, L)$.

Another advantage of formulation (P_3) over (P_1) is the fact that the boundary condition $u'(L) = 0$ may be disregarded in the former, for it is implicitly satisfied. Indeed, if we still assume that all the operations performed below are legitimate, we have from (P_3)

$$pu'v\Big|_0^L - \int_0^L (pu')'vdx + \int_0^L quvdx = \int_0^L fvdx \tag{1.3}$$

for every $v \in V$. Let us choose $v \in V$ such that $v(L) = 0$, but otherwise arbitrary. In doing so, we obtain

$$\int_0^L \left[-(pu')' + qu - f\right] vdx, \ \ \forall v \in V \text{ such that } v(L) = 0$$

Following reference [126], here the reader can take it for granted that the class of such functions v is wide enough for the above relation to imply that

$$-(pu')' + qu = f \text{almost everywhere in } (0, L)$$

From this result equation, (1.3) simply becomes

$$p(L)u'(L)v(L) - p(0)u'(0)v(0) = 0, \ \forall v \in V$$

Since $v(0) = 0$, choosing now $v \in V$ such that $v(L) = 1$, we immediately infer that $u'(L) = 0$ as required, as p is strictly positive by assumption.

Another clear advantage of (P_3) over (P_1) and (P_2) is related to the regularity of the datum p. Even assuming that the material of the bar is homogeneous, the function p varies with its cross section. Eventually, the latter could change shape abruptly in such a way that p is a discontinuous function. Although, on the one hand, this does not bring about any difficulty as far as problem (P_3) is concerned, more care is needed in handling the term $(pu')'$ in equation (P_1) or the term $[pu'](a_\omega) - [pu'](b_\omega)$ in equation (P_2) in this case.

Finally, a more tricky but fundamental point: formulation (P$_3$) (and to a lesser extent (P$_2$)) is more general than equation (P$_1$) from the physical point of view. In order to clarify this assertion, let us consider a distribution of forces f_ε that have a resultant P, uniformly acting on a very small length ε around the abscissa $x = L/2$. In current applications, such a system of forces is represented by $P\delta_{\frac{L}{2}}$, namely, a single force of modulus equal to P applied at the point given by $x = L/2$.[4] Of course, if we stick to the first definition of f_ε, all the problems (P$_1$), (P$_2$) and (P$_3$) carry a meaning. Actually, denoting by u_ε the corresponding solution, problem (P$_3$) writes

$$\int_0^L pu_\varepsilon' v' dx + \int_0^L qu_\varepsilon v dx = \int_{\frac{L}{2}-\frac{\varepsilon}{2}}^{\frac{L}{2}+\frac{\varepsilon}{2}} \frac{P}{\varepsilon} v dx, \quad \forall v \in V$$

Now, since v is a continuous function, the mean value theorem for integrals implies that the right side of the above equation tends to $Pv(L/2)$ as ε goes to zero. Otherwise stated, the variational problem (P$_3$) is perfectly well defined in the case of the single force applied to a particle located at point $x = L/2$, as

$$\int_0^L pu' v' dx + \int_0^L quv dx = Pv(L/2), \quad \forall v \in V, u \text{ itself in } V$$

As for problem (P$_2$), if we consider a subset $\mathcal{V}_\varepsilon$ of $\mathcal{V}$ consisting of intervals that either fully contain $(L - \varepsilon/2, L + \varepsilon/2)$ or have an empty intersection with it, we have

$$[pu'](a_\omega) - [pu'](b_\omega) + \int_{a_\omega}^{b_\omega} qu dx = \int_{a_\omega}^{b_\omega} f dx \quad \forall \omega \in \mathcal{V}_\varepsilon. \tag{1.4}$$

This means that the left side of equation (1.4) equals P if $\omega \in \mathcal{V}_\varepsilon$ is such that $\omega \cap (L/2 - \varepsilon/2, L/2 + \varepsilon/2) \neq \emptyset$ and equals 0 otherwise.

Since ε is bound to tend to zero, problem (1.4) will be as general as (P$_2$) for $\varepsilon = 0$. Indeed, it suffices to modify the conservative form in such a way that the integral $\int_\omega f dx$ is replaced by P if ω contains $L/2$ and by 0 otherwise.

On the other hand, the equation (P$_1$) in this case could only be based on an equality of the type $-(pu')' + qu = P\delta_{\frac{L}{2}}$, whose exact meaning is unclear for numerical purposes.

Since forces applied pointwise may act everywhere and may also be combined, we conclude that there are at least infinitely many systems of forces that are admissible for the models (P$_3$) and (P$_2$)

of the elongational deformation of a bar, although they do not correspond to a differential equation of the form (P_1), at least as far as functions defined in accordance with the classical concept are concerned [179].

Summarising, we have exhibited four advantages of formulation (P_3) over (P_1), among which at least two also apply to (P_2), in spite of the fact that a rather simple linear ODE was used as a model for the purpose of this comparison.

1.2 The Basic FDM

In this section, we introduce one of the oldest and simplest numerical methods to solve equation (P_1): the **Three-point FDM**. The basic idea is to search for values of the unknown function u only at the points in $[0, L]$ belonging to a certain (finite) discretisation lattice $\mathcal{G} := \{x_0, x_1, x_2, \ldots, x_{n-1}, x_n\}$, which is calledan **FD grid** in this case. An FD grid is illustrated in Figure 1.2.

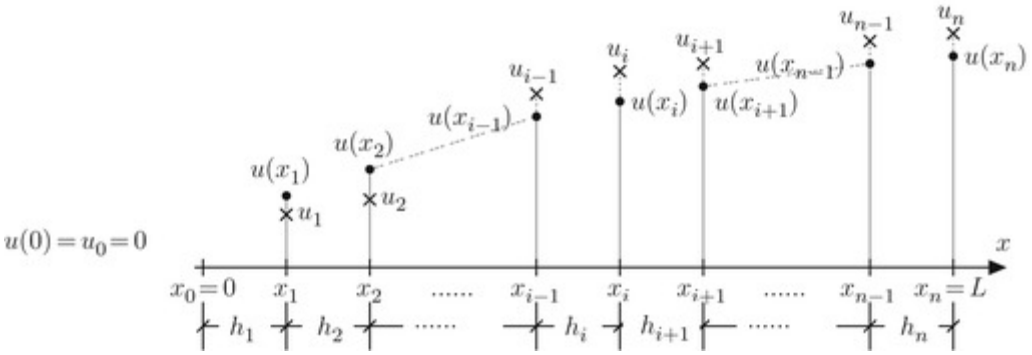

Figure 1.2 An FD grid with exact and approximate values of $u(x)$ at grid points x_i.

For the sake of simplicity, we first consider that both p and q are constant. In this case, problem (P_1) becomes

$$\begin{cases} -pu'' + qu = f \ \text{ in } (0, L) \\ u(0) = 0 \qquad \text{(fixed left end)} \\ u'(L) = 0 \qquad \text{(free right end)} \end{cases} \tag{1.5}$$

Of course, such a problem can be solved by hand provided f is sufficiently simple. However, in order to attain our goal, pretending this is not the case, we head for a numerical solution.

Let the set of $n+1$ points of $\mathcal{G}$ be such that $x_0 = 0$, $x_n = L$ and $x_i > x_{i-1}$ for every $i \in \{1, 2, \ldots, n-1, n\}$. For the moment, we assume that these points are equally spaced. Thus,

setting $h = L/n$, we have $x_i = ih$, for $i \in \{0, 1, 2, \ldots, n-1, n\}$, so that $x_i - x_{i-1} = h$ for every $i \in \{1, 2, \ldots, n-1, n\}$. If the function f is twice continuously differentiable, the solution u of (P_1) will be four times continuously differentiable. So, if $i \neq 0$ and $i \neq n$, using a standard Taylor expansion we have

$$\begin{cases} u(x_{i+1}) = u(x_i) + hu'(x_i) + h^2u''(x_i)/2 + h^3u'''(x_i)/6 + h^4u^{iv}(\xi_i^+)/24 \\ \text{and} \\ u(x_{i-1}) = u(x_i) - hu'(x_i) + h^2u''(x_i)/2 - h^3u'''(x_i)/6 + h^4u^{iv}(\xi_i^-)/24, \end{cases} \tag{1.6}$$

where ξ_i^+ and ξ_i^- are suitable points belonging to the intervals $[x_i, x_{i+1}]$ and $[x_{i-1}, x_i]$, respectively. Adding up the two relations of equation (1.6), after simple manipulations we come up with

$$\begin{cases} -u''(x_i) = \dfrac{2u(x_i) - u(x_{i+1}) - u(x_{i-1})}{h^2} + R_i(u)h^2 \\ \text{where} \\ R_i(u) = [u^{iv}(\xi_i^+) + u^{iv}(\xi_i^-)]/24, \end{cases} \tag{1.7}$$

On the other hand, from equation (1.5) it holds that

$$-pu''(x_i) + qu(x_i) = f(x_i) \text{ for } i = 1, 2, \ldots, n-1. \tag{1.8}$$

Hence, equation (1.7) means that, up to the term $R_i(u)h^2$, we may replace in equation (1.8) the second derivative of u at x_i by the **FD**, that is, the **finite difference** $[2u(x_i) - u(x_{i+1}) - u(x_{i-1}]/h^2$. Otherwise stated, provided u is four times differentiable in $[0, L]$, the following relations hold among the values of u at the grid points formed by the x_is:

$$\begin{cases} \text{For } i = 1, 2, \ldots, n-1, \\ p\dfrac{2u(x_i) - u(x_{i+1}) - u(x_{i-1})}{h^2} + qu(x_i) = f_h(x_i) \\ \text{where} \\ f_h(x_i) = f(x_i) - pR_i(u)h^2. \end{cases} \tag{1.9}$$

Our assumptions on u allow us to verify quite easily that $R_i(u)$ can be bounded independently of n. Thus, provided n is large enough, or equivalently h is sufficiently small, the **finite difference equation** $p\frac{2u(x_i)-u(x_{i+1})-u(x_{i-1})}{h^2} + qu(x_i) = f(x_i)$ is satisfied at every point x_i for $i = 1, \ldots, n-1$ up to a small additive term whose magnitude goes to zero as fast as h^2 when n increases indefinitely. However, though certainly convincing, such an argument is not satisfactory from both the practical and the mathematical points of view. Indeed, we are looking for values of the unknown u at least at the grid points, and relations (1.9) cannot be exploited for this purpose

since the function f_h, in spite of being very close to f, is not exactly known. Nevertheless, the above development naturally leads to the conceptual construction of the FDM: if the values $u(x_i)$ do not satisfy an exploitable relation, there should exist some values u_i close to them (i.e., **approximations** of theirs) that do (cf. Figure 1.2). More precisely, the u_is are meant to fulfil

$$\begin{cases} \text{For } i = 1, 2, \ldots, n-1, \\ p\dfrac{2u_i - u_{i+1} - u_{i-1}}{h^2} + qu_i = f(x_i) \end{cases} \tag{1.10}$$

However, there are other features of problem (P_1) which have not been incorporated into our description of the FDM yet, such as the boundary conditions $u(0) = 0$ and $u'(L) = 0$. Whereas the former is naturally taken into account by simply setting $u_0 = 0$, the latter requires somewhat more careful considerations. A possibility that seems quite natural comes from the first (resp. second) relation of equation (1.6) for $i = n-1$ (resp. $i = n$). Indeed, this tells us that $u'(L) = [u(x_n) - u(x_{n-1})]/h + T_n(u)h$, where $T_n(u)$ accounts for the terms involving derivatives of u of order higher than one. This suggests that the approximate values u_{n-1} and u_n must satisfy $u_n - u_{n-1} = 0$. However, as we will see later on, such a choice is not compatible with the quality of approximation provided by the way we approximate the differential equation itself. The introduction of the concept of a fictitious point $x_{n+1} := (n+1)h = L + h$ lying outside the domain $(0, L)$ gives rise to a better approximation of the boundary condition at $x = L$. We associate with this point a fictitious exact value of u, $u(x_{n+1})$ satisfying a relation of the form $p\frac{2u(x_n) - u(x_{n+1}) - u(x_{n-1})}{h^2} + qu(x_n) = f_h(x_n)$, where $f_h(x_n)$ stands for a natural extension of f_h to $x = L$, which we decline to specify.

Then we mimic as fictitiously the first relation of equation (1.6), taking $i = n$ this time, and subtract the second one from the resulting relation to conclude that a (fictitious) value $u(x_{n+1})$ would satisfy $[u(x_{n+1}) - u(x_{n-1})]/h = S_n(u)h^2$, where $S_n(u)$ accounts for terms involving derivatives of u of order higher than two. This looks like a more reasonable way to deal with the boundary condition $u'(L) = 0$ in our FD analog of (P_1): we introduce an additional unknown u_{n+1} and then eliminate it by simply making $u_{n+1} - u_{n-1} = 0$. Recalling equation (1.10), this leads as naturally to the following extension of the approximation of the differential equation at point $x = x_n = L$:

$$\begin{cases} p\dfrac{2u_n - u_{n+1} - u_{n-1}}{h^2} + qu_n = f(x_n) \\ \text{with } u_{n+1} = u_{n-1}. \end{cases} \tag{1.11}$$

Finally combining equations (1.10) and (1.11), and using the term **uniform grid** to qualify an equally spaced grid, we come up with

The three-point FD analog of (P_1) with a uniform grid (1.12)

$$\boxed{\begin{array}{l} \text{Find } u_i, i = 1, 2, \ldots, n \text{ satisfying} \\ p\dfrac{2u_i - u_{i+1} - u_{i-1}}{h^2} + qu_i = f(x_i) \\ \text{with } u_0 = 0 \text{ and } u_{n+1} = u_{n-1}. \end{array}}$$

where u_i is an **approximation** of $u(x_i)$.

A question that immediately arises concerns the quality of such an approximation. In principle we can assert that everything is fine, but the choice $u_{n+1} = u_{n-1}$ might not be the best possible in this respect. For the moment we refrain from specifying what an optimal way to define u_{n+1} should be, among other quality considerations. These are left to Chapter 2, which is specifically devoted to this subject.

Going back to equation (1.12), let us take a closer look at this problem's nature. First of all, we note that equation (1.12) is nothing but a SLAE with n unknowns and n equations, whose right side is the vector $[f(x_1), f(x_2), \ldots, f(x_n)]^T$. Therefore, it possesses a unique solution if and only if $f(x_i) = 0$ for all $i \in \{1, 2, \ldots, n\}$, implies that $u_i = 0$ for $i = 1, 2, \ldots, n$. Actually this implication holds true, according to the following argument.

Let M be the subscript of (one of) the point(s) at which $|u_i|$ attains its maximum value. If $M = 1$, since $f_1 = 0$, taking $i = 1$ in equation (1.12) we have $|u_1| = p|u_2|/(2p + qh^2) \leq |u_2|/2$, since by assumption $p > 0$ and $q \geq 0$. Then, noting that necessarily $|u_2| \leq |u_1|$, we conclude that $|u_1| = 0$ and the result is thus established. If $M > 1$, recalling that $f(x_M) = 0$, we infer from equation (1.12) that $|u_M|$ is at most equal to the mean value of $|u_{M-1}|$ and $|u_{M+1}|$. Since none of these values can be greater than $|u_M|$ by assumption, both are necessarily equal to $|u_M|$. Thus, the maximum value is also attained for $i = M - 1$. If $M - 1 = 1$ we are finished, but if not we may apply the same argument to $|u_{M-1}|$, thereby concluding that the maximum value of $|u_i|$ is also equal to $|u_{M-2}|$. Then step by step we will reach the equation $|u_1| = p|u_2|/(2p + qh^2)$, knowing that both $|u_1|$ and $|u_2|$

must take the maximum absolute value of all the u_is. This implies that $|u_1| = |u_2| = 0$, and hence all the u_is must equal zero.

Another issue to address at this stage is the form of the SLAE (equation (1.12)). By inspection and after division by a factor of two of the nth equation, we easily infer from equation (1.12) that the vector $\vec{u}_h := [u_1, u_2, \ldots, u_n]^T$, consisting of the n approximate values of u, satisfies a SLAE with matrix $A_h = \{a_{i,j}\}$ and right-side vector $\vec{b}_h = \{b_i\}$ of the form

SLAE for the Three-point FD scheme with a uniform grid (1.13)

$$
\begin{array}{|ll|}
\hline
A_h\vec{u}_h = \vec{b}_h & \\
\text{where} & \\
a_{i,i} = 2p/h^2 + q \text{ if } i \neq n; & a_{n,n} = p/h^2 + q/2; \\
a_{i,i+1} = a_{i+1,i} = -p/h^2 \ \forall i < n; & a_{i,j} = 0 \text{ if } |i-j| > 1; \\
\text{and} & \\
b_i = f(x_i) \text{ if } i \neq n; & b_n = f(x_n)/2. \\
\hline
\end{array}
$$

The reader might observe that, thanks to the multiplication by $1/2$ of the nth equation resulting from equation (1.12), A_h became a symmetric matrix.

Before pursuing our presentation of the FDM, it seems instructive to give an idea of the aspect of both matrix A_h and vector $\vec{b}_h$, considering the particular case where p and q are constant. Taking $n = 6$, for instance, we display in Array 1.1 the matrix A_h in this case, whereas vector $\vec{b}_h$ is exhibited in equation (1.14).

$$
A_h = \begin{bmatrix}
\frac{2p}{h^2} + q & -\frac{p}{h^2} & 0 & 0 & 0 & 0 \\
-\frac{p}{h^2} & \frac{2p}{h^2} + q & -\frac{p}{h^2} & 0 & 0 & 0 \\
0 & -\frac{p}{h^2} & \frac{2p}{h^2} + q & -\frac{p}{h^2} & 0 & 0 \\
0 & 0 & -\frac{p}{h^2} & \frac{2p}{h^2} + q & -\frac{p}{h^2} & 0 \\
0 & 0 & 0 & -\frac{p}{h^2} & \frac{2p}{h^2} + q & -\frac{p}{h^2} \\
0 & 0 & 0 & 0 & -\frac{p}{h^2} & \frac{p}{h^2} + \frac{q}{2}
\end{bmatrix}
$$

Array 1.1

$$
\vec{b}_h = \left[f(x_1),\ f(x_2),\ f(x_3),\ f(x_4),\ f(x_5),\ f(x_6)/2\right]^T. \tag{1.14}
$$

Notice that most coefficients of A_h are zero (at least for $n > 5$). This kind of matrix is called a **sparse matrix**. Actually matrix A_h, besides being sparse, is also a **tridiagonal matrix** since $a_{i,j} = 0$ whenever $|i-j| > 1$.

In order to extend the FDM to more general cases, it is wise to regard the FD used to approximate u'' at x_i in the light of a different interpretation:

If we introduce intermediate points $x_{i-1/2}$ given by $(i-1/2)h$, for $i = 1, \ldots, n$, we may define approximations of the first-order derivative of u at those points by the FD $D_u^1(x_{i-1/2}) := [u(x_i) - u(x_{i-1})]/h$. Then, naturally enough, we define an approximation of the second-order derivative of u at x_i by another FD using such approximations of the first-order derivatives at $x_{i-1/2}$ and $x_{i+1/2}$, thereby obtaining $D_u^2(x_i) := [D_u^1(x_{i+1/2}) - D_u^1(x_{i-1/2})]/h$, for $i = 1, 2, \ldots, n-1$. Of course, if we consider the context of an FD scheme, we must replace by u_i the value of $u(x_i)$ in the expression of $D_u^1(x_{i-1/2})$, $i = 1, 2, \ldots, n$, thereby obtaining another approximation $\delta_{h,i-1/2}$ of $u'(x_{i-1/2})$. Then, the approximation of the second-order derivative denoted by $\Delta_{h,i}$ instead of $D_u^2(x_i)$, will be given by $\Delta_{h,i} := [\delta_{h,i+1/2} - \delta_{h,i-1/2}]/h$, $i = 1, 2, \ldots, n-1$. In this case, using the concept of a fictitious point, we may extend to $i = n$ the definition of $\delta_{h,i+1/2}$, and hence the one of $\Delta_{h,i}$.

Now for several reasons, one might be interested in using a non–equally spaced grid – or **non-uniform grid** – to solve problem (P$_1$). For instance, this could happen if the datum f was very irregular in a certain region, in which it would be advisable to place more grid points in an attempt to improve the local accuracy of the method. In this case, using a fictitious point which we define to be $x_{n+1} := L + h_n$, we set $h_i := x_i - x_{i-1}$, for $i = 1, 2, \ldots, n+1$, and approximate the first derivative of u at $x_{i-1/2} := (x_{i-1} + x_i)/2$ by the incremental ratio

$$D_u^1(x_{i-1/2}) := [u(x_i) - u(x_{i-1})]/h_i.$$

Then quite naturally, for $i = 1, 2, \ldots, n$ the second-order derivative of u at x_i is approximated by **the divided difference** using the local **intermediate grid size**, i.e., by

$$D_u^2(x_i) := \frac{D_u^1(x_{i+1/2}) - D_u^1(x_{i-1/2})}{(h_{i+1} + h_i)/2}.$$

Noticing that the values of u at the grid points x_i are not available, and using instead corresponding approximations u_i, we are led to another approximation of $u''(x_i)$, still denoted by $\Delta_{h,i}$, for $i = 1, 2 \ldots, n$, given by

$$\Delta_{h,i} := \frac{\delta_{h,i+1/2} - \delta_{h,i-1/2}}{(h_{i+1} + h_i)/2}$$

with

$$\delta_{h,i-1/2} := [u_i - u_{i-1}]/h_i,$$

where $\delta_{h,i-1/2}$ is an approximation of $u'(x_{i-1/2})$ for $i = 1, 2, \ldots, n$, with the obvious fictitious extension $\delta_{h,n+1/2}$.

All this leads to the following problem to approximate equation (P_1) in the case of constant coefficients p and q:

The three-point FD analog of (P_1) with a non-uniform grid (1.15)

Find $u_i, i = 1, 2, \ldots, n$ satisfying
$-p\Delta_{h,i} + qu_i = f(x_i)$
with $u_0 = 0$ and $u_{n+1} = u_{n-1}$
where
$$\Delta_{h,i} := \frac{\delta_{h,i+1/2} - \delta_{h,i-1/2}}{(h_{i+1} + h_i)/2}$$
with $\delta_{h,i-1/2} := [u_i - u_{i-1}]/h_i$.

Like equation (1.12), equation (1.15) is a SLAE with an $n \times n$ matrix $A_h = \{a_{i,j}\}$ and right-side vector $\vec{b}_h = \{b_i\}$, and unknown vector $\vec{u}_h = [u_1, u_2, \ldots, u_n]^T$. We obtain a SLAE with a non symmetric matrix, **symmetrisable** through the multiplication of each equation by the corresponding intermdiate grid size, namely,

The SLAE for the Three-point FD scheme with a non-uniform grid (1.16)

$A_h\vec{u}_h = \vec{b}_h$
where

$a_{i,j} = 0$	if $\lvert i - j\rvert > 1$;
$a_{i,i} = 2p/(h_i h_{i+1}) + q$	for $1 \le i < n$; $a_{n,n} = p/h_n^2 + q/2$;
$a_{i,i+1} = -2p/[h_{i+1}(h_i + h_{i+1})]$	for $1 \le i < n$;
$a_{i,i-1} = -2p/[h_i(h_i + h_{i+1})]$	for $1 < i < n$; $a_{n,n-1} = -p/h_n^2$;
$b_i = f(x_i)$	for $1 \le i < n$; $b_n = f(x_n)/2$.

The reader might prove that system (1.16) has a unique solution (cf. Exercise 1.1).

While, on one hand, the extension of either equation (1.12) or (1.15) to the case where q is not constant is straightforward, on the other hand this task requires some care as far as p is concerned. For this reason we postpone such an extension, since it can be carried out more easily in the light of the other two discretisation methods studied in this chapter. This is more especially the case of the FEM, which is next presented.

1.3 The Piecewise Linear FEM ($\mathcal{P}_1$ FEM)

As already pointed out, the FEM applies to the variational form of problem (P_1), namely, equation (P_3). Here, we no longer assume any differentiability property of p in $(0, L)$. We keep only the assumption that this function is bounded above and below away from zero in $[0, L]$. However,although this is by no means necessary, we make the more than reasonable hypothesis from the physical point of view that p is discontinuous only at a certain finite number of points y_i, $i = 1, 2, \ldots, M$, with $y_1 < y_2 < \ldots < y_M$. Here, the method's discretisation lattice is called a **mesh** denoted by $\mathcal{T}_h$, consisting of n closed intervals $T_i := [x_{i-1}, x_i]$ called **elements**, with $0 = x_0 < x_1 < \ldots < x_{n-1} < x_n = L$. The mesh points S_i whose abscissae are the x_i s are called **nodes**. Notice that the nodes are the grid points of a non-uniform FD grid. Recalling the definition of h_i in Section 1.2, we further define a characteristic length of the mesh $\mathcal{T}_h$ called

The mesh step size h

$$h := \max_{i \in \{1,2,\ldots,n\}} h_i. \tag{1.17}$$

If all the elements of the mesh have the same length h, we say that $\mathcal{T}_h$ is a **uniform mesh**. Now, for every element $T_j \in \mathcal{T}_h$, we denote by $\mathcal{P}_1(T_j)$ the set of polynomials of degree less than or equal to one defined in T_j. Using the Kronecker symbol δ_{ij}[5], we introduce a set of **shape functions** φ_i for $i = 0, 1, \ldots, n-1, n$ through the

Shape function definition (1.18)

$$\boxed{\begin{array}{l}\varphi_i \in \mathcal{P}_1(T_j)\ \forall T_j \in \mathcal{T}_h; \\ \varphi_i(x_j) = \delta_{ij}\end{array}}$$

This definition is illustrated in Figure 1.2, which displays the shape function φ_i for a generic i different from 0 and n, together with φ_0 and φ_n. Actually, equation (1.18) allows us to derive quite easily the following analytic expression of φ_i, for $1 \leq i < n$:

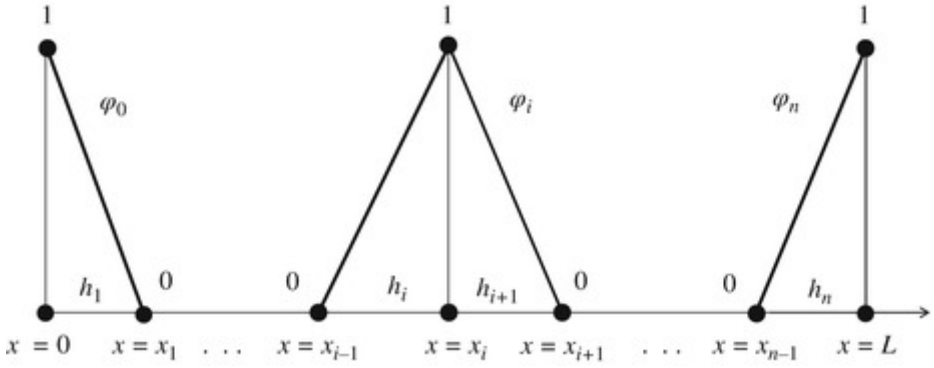

Figure 1.2 Shape functions φ_i for $0 < i < n$, φ_0, and φ_n.

$$\varphi_i(x) = \begin{cases} 0 & \text{if } x \notin [x_{i-1}, x_{i+1}] \\ (x - x_{i-1})/h_i & \text{if } x \in [x_{i-1}, x_i] \\ (x_{i+1} - x)/h_{i+1} & \text{if } x \in [x_i, x_{i+1}]. \end{cases} \quad (1.19)$$

The reader should complete the definition of the shape functions by determining the analytic expressions of φ_0 and φ_n. The space spanned by these n+1 shape functions is denoted by W_h.

Now, instead of u, we shall search for an approximation u_h associated with $\mathcal{T}_h$ of the form:

$$u_h = \sum_{j=1}^{n} u_j \varphi_j. \quad (1.20)$$

The coefficients u_j in equation (1.20) are nothing but the values $u_h(x_j)$, as one can easily check using property (1.18). Actually, they will play the role of approximations of $u(x_j)$ in the same way as in the FDM. Now, instead of sweeping the whole space V, we restrict the variational problem to the space V_h consisting of functions of the form $\sum_{i=1}^{n} v_i \varphi_i$, where the v_i s are arbitrary real numbers. In other words, V_h is the space spanned by the φ_i s for $i > 0$, and in this sense u_h itself belongs to V_h.

In Figure 1.3, we show the general aspect of a function in V_h, such as u_h. Since by construction $\varphi_i(0) = 0$ for every i, except for $i = 0$, we observe that all the functions v in the space V_h vanish at $x = 0$. For a FEM based on the standard Galerkin formulation, the solution u_h

together with test functions $v \in V_h$ should belong to the space V defined in Section 1.1. The fact that all of them vanish at $x = 0$ is half a way to fulfil this condition. Another related question is why the shape functions φ_i, and consequently any $v \in V_h$, must be continuous. The reader is encouraged to think about it to find the right answer. She or he is referred to Subsection 7.1.2 for a hint.

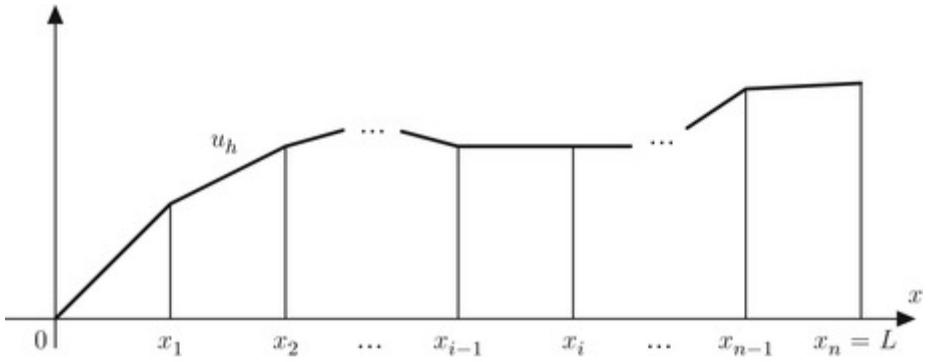

Figure 1.3 An illustration of a typical function belonging to FE space V_h.

Now we set

The $\mathcal{P}_1$ FE approximation of (P_1) (1.21)

Find $u_h \in V_h$ such that
$\int_0^L [p u_h' v' + q u_h v] dx = \int_0^L f v dx \ \forall v \in V_h$

In particular, equation (1.21) must hold for $v = \varphi_i$ with $i = 1, 2, \ldots, n$. Then, replacing v with φ_i and using equation (1.20) together with well-known properties of integrals, we come up with the **Piecewise linear FE** scheme given by equation (1.22), whose unknowns u_j are approximations of $u(x_j)$, u being the solution of (P_3). This scheme is also called

The $\mathcal{P}_1$ FE scheme to solve (P_1) (1.22)

Find u_j for $j = 1, 2, \ldots, n$ such that
$$\sum_{j=1}^{n} u_j \int_0^L [p \varphi_j' \varphi_i' + q \varphi_j \varphi_i] dx = \int_0^L f \varphi_i dx \ \text{for } i = 1, 2, \ldots, n.$$

Conversely, multiplying both sides of the above equation by an arbitrary real number v_i and adding up from $i = 1$ up to $i = n$, we conclude quite easily that equation (1.22) implies (1.21). As a consequence, both problems can be viewed as equivalent, if we identify u_h with its coefficients in equation (1.20), even if rigorously they are different entities, since u_h is a function defined at every point of $[0, L]$, and the u_j s are only the values of u_h at a finite set of points x_j.

Here again, equation (1.22) is nothing but a SLAE with an $n \times n$ matrix $A_h = \{a_{i,j}\}$ and right-side vector $\vec{b}_h = \{b_i\}$, whose unknown vector $\vec{u}_h$ is $[u_1, u_2, \ldots, u_n]^T$. More concretely, we have to solve

The SLAE for the $\mathcal{P}_1$ FE scheme (1.23)

$$A_h \vec{u}_h = \vec{b}_h$$
where
$$a_{i,j} = \int_0^L [p\varphi_j' \varphi_i' + q\varphi_j \varphi_i] dx$$
and
$$b_i = \int_0^L f\varphi_i dx.$$

This means that equation (1.22) can be recast in the same form as equation (1.13). Notice that matrix A_h is necessarily symmetric.

Next, we examine the well-posedness of equation (1.22). As we know, this is ensured provided $\vec{b}_h = \vec{0}$ implies $\vec{u}_h = \vec{0}$. Since systems (1.22) and (1.21) are equivalent, if $\vec{b}_h = \vec{0}$ we have

$$\int_0^L [pu_h' v' + qu_h v] dx = 0 \; \forall v \in V_h.$$

Hence, taking $v = u_h$ we conclude that

$$\int_0^L [p(u_h')^2 + q(u_h)^2] dx = 0.$$

On the other hand, according to our assumptions, $p(x) \geq \alpha > 0$ and $q(x) \geq 0$ for every $x \in [0, L]$. Therefore, the left side of the above expression is strictly positive, unless u_h' vanishes identically, which is thus the only possibility for the above relation to hold. Hence, u_h must be constant in $[0, L]$, and since by construction $u_h(0) = 0$, we must have $u_h \equiv 0$. This implies that $\vec{u}_h = \vec{0}$.

Similarly to the case of the FDM with an equally spaced grid, let us exhibit the matrix A_h in the particular case where p and q are constant. Referring to Figure 1.2 and to equation (1.19), first of all we note that the product of two functions φ_i and φ_j, or of their derivatives, is identically zero whenever $|i - j| > 1$. On the other hand, from equation (1.19) the product of φ_i and φ_{i-1} (or of its derivatives) vanishes identically in every element but T_i. It immediately follows that, like in the case of the FDM for a uniform grid, A_h besides being symmetric is a tridiagonal matrix.

Let us first compute $a_{i+1,i}$ for $1 \leq i < n$. Using equation (1.19), we trivially obtain

$$a_{i+1,i} = \int_{x_i}^{x_{i+1}} h_{i+1}^{-2}[-p + q(x - x_i)(x_{i+1} - x)]dx$$

Integrating the quadratic function above in (x_i, x_{i+1}), we readily derive

$$a_{i+1,i} = -ph_{i+1}^{-1} + qh_{i+1}/6.$$

Similarly, we can compute $a_{i,i}$ for $1 \leq i < n$ by

$$a_{i,i} = \int_{x_{i-1}}^{x_i} h_i^{-2}[p + q(x - x_{i-1})^2]dx + \int_{x_i}^{x_{i+1}} h_{i+1}^{-2}[p + q(x_{i+1} - x)^2]dx$$

which yields

$$a_{i,i} = p(h_i^{-1} + h_{i+1}^{-1}) + q(h_i + h_{i+1})/3.$$

Without any difficulty, we infer from the above calculation of $a_{i,i}$ that

$$a_{n,n} = ph_n^{-1} + qh_n/3.$$

An illustration of this matrix divided by h is supplied in Array 1.2 for $n = 6$ and a uniform mesh with size h.

$$\frac{A_h}{h} = \begin{bmatrix} \frac{2p}{h^2} + \frac{2q}{3} & -\frac{p}{h^2} + \frac{q}{6} & 0 & 0 & 0 & 0 \\ -\frac{p}{h^2} + \frac{q}{6} & \frac{2p}{h^2} + \frac{2q}{3} & -\frac{p}{h^2} + \frac{q}{6} & 0 & 0 & 0 \\ 0 & -\frac{p}{h^2} + \frac{q}{6} & \frac{2p}{h^2} + \frac{2q}{3} & -\frac{p}{h^2} + \frac{q}{6} & 0 & 0 \\ 0 & 0 & -\frac{p}{h^2} + \frac{q}{6} & \frac{2p}{h^2} + \frac{2q}{3} & -\frac{p}{h^2} + \frac{q}{6} & 0 \\ 0 & 0 & 0 & -\frac{p}{h^2} + \frac{q}{6} & \frac{2p}{h^2} + \frac{2q}{3} & -\frac{p}{h^2} + \frac{q}{6} \\ 0 & 0 & 0 & 0 & -\frac{p}{h^2} + \frac{q}{6} & \frac{p}{h^2} + \frac{q}{3}. \end{bmatrix}$$

Array 1.2

On the other hand, in general the calculation of $\vec{b}_h$ is not as simple as in the case of the FDM, unless we use numerical quadrature. This is possible if f is continuous in $[0, L]$. In this case, since b_i is the sum of theintegrals of $f\varphi_i$ over elements T_i and T_{i+1}, without any essential loss in accuracy, we may apply the trapezoidal rule (see e.g. [105]) in each of them, thereby obtaining an approximation $\tilde{b}_i$ of b_i given by

$$\tilde{b}_i = [f(x_{i-1})\varphi_i(x_{i-1}) + f(x_i)\varphi_i(x_i)]h_i/2 + [f(x_{i+1})\varphi_i(x_{i+1}) + f(x_i)\varphi_i(x_i)]h_{i+1}/2.$$

Since $\varphi_i(x_j) = 0$ if $i \neq j$, we derive

$$\tilde{b}_i = f(x_i)(h_i + h_{i+1})/2 \text{ for } i < n$$

and, quite easily,

$$\tilde{b}_n = f(x_n)h_n/2.$$

Notice that this is the same right side as in the case of the FD scheme (1.15) if we multiply the ith equation by $(h_i + h_{i+1})/2$ for $i < n$ and the nth equation by h_n. In particular, if the mesh is uniform, we obtain the same right side as equation (1.14) multiplied by h.

Of course, we can also use the trapezoidal rule to integrate the terms $\int_{x_{i-1}}^{x_i} qg(x)dx$ in the expression of $a_{i,i}$ or $a_{i,i-1}$, where $g(x) = [(x - x_{i-1})/h_i]^2$ or $g(x) = (x - x_{i-1})(x_i - x)/h_i^2$. Then, quite easily, we establish that the resulting approximate values $\tilde{a}_{i,j}$ of the nonzero entries of matrix A_h correspond to the same entries given by equation (1.16) multiplied by $(h_i + h_{i+1})/2$ for $i < n$ and otherwise by $h_n/2$. In particular in the case of a uniform mesh, the matrix A_h/h is of the form displayed in Array 1.1 (for $n = 6$). Otherwise stated, at least if both p and q are constant, the use of the trapezoidal rule to integrate all the functions involved in the calculation of matrix A_h and right-side vector $\vec{b}_h$ results in a SLAE identical to the one of the FDM for a non-uniform grid, whose points coincide with the nodes of $\mathcal{T}_h$. Actually, even in the case where p and q are not constant, but arbitrary continuous functions, the use of the same trapezoidal rule to compute the **FE matrix** A_h induces an identical **FD analog** of (P_1) based on the same non-uniform lattice. The reader could derive such an **FD scheme** as Exercise 1.2, and verify that it yields relations (1.16) whenever p and q are constant.

The reader has certainly observed that if the trapezoidal rule is used to approximate the integral $\int_0^L q\varphi_i\varphi_{i\pm1}dx$, the corresponding terms of the coefficient $a_{i,i\pm1}$ will be zero. The resulting diagonalisation procedure in connection with the $\mathcal{P}_1$ FEM is known as **mass lumping**. In this book this technique will serve several times as a tool to obtain interesting properties of FE-based numerical schemes.

Another similar situation of practical interest is the one where p is only piecewise continuous in $[0, L]$, q being piecewise continuous or not. In this case, it may be practical to construct the FE mesh (resp. FD grid) in such a way that the set of discontinuity points $\{y_1, y_2, \ldots, y_M\}$ forms a

subset of $\{x_1, x_2, \ldots, x_{n-1}\}$. In other words, we assume that the discontinuity points of p coincide with mesh nodes (resp. grid points). Then, here again, at least for a continuous q, the FE scheme resulting from the use of the trapezoidal rule as an approximate quadrature method in each element corresponds to an efficient Three-point FD scheme with a non-uniform grid, to solve equation (P_1) for a discontinuous p.

Remark 1.2

Such an assertion is based on the fact that here pu' is a differentiable function everywhere, even if neither p nor u' is differentiable at the discontinuity points of p. Then, it is reasonable to apply FDs to approximate directly $[pu']'$ and not u'' like we did in Section 1.2. The reader will be able to figure this out when deriving the corresponding FE scheme.

Finally, if q too is only piecewise continuous, the trapezoidal rule may be applied to integrals in the intersection of the elements with intervals in which q is continuous. An equivalent and natural FD scheme can also be thus derived, in the case where the discontinuities of q coincide with mesh nodes or grid points, as the reader may quite easily check taking the case of a single discontinuity.

1.4 The Basic FVM

The FVM applies to formulation (P_2) of the differential equation for a particular type of subsets ω called **control volumes** (**CV**s). Although there is only one standard formulation of the basic FVM, for any set of disjoint CVs, in the literature two approaches are commonly distinguished: **Vertex-centred FVM** and **Cell-centred FVM**. In the former, the CVs are intervals specifiedhereafter, containing the nodes of the mesh $\mathcal{T}_h$; it yields discrete problems quite close to those considered in Sections 1.1 and 1.2 for the FDM and the FEM. In the latter, the CVs are the elements of $\mathcal{T}_h$ themselves. Of course, since they are based on the same principles, both methods are very similar. However, even in the case of the one-dimensional problem (P_1), the Cell-centred FVM does not resemble the Three-point FDM or the $\mathcal{P}_1$ FEM.

1.4.1 The Vertex-centred FVM

Starting from the FE mesh $\mathcal{T}_h$ (or the associated non-uniform grid), in this version of the FVM the CVs are the intervals $V_j := (x_{j-1/2}, x_{j+1/2})$ where $x_{j-1/2} := (x_{j-1} + x_j)/2$ for $j = 1, 2, \ldots, n-1$, which we complete with $V_n := (x_{n-1/2}, x_n)$ and $V_0 := (x_0, x_{1/2})$. The measure of each CV for $1 \leq j \leq n-1$ denoted by $h_{j-1/2}$ is given by $(h_{j-1} + h_j)/2$ for $j < n$ and by $h_1/2$ and $h_n/2$ for V_0 and V_n, respectively. Now, we assume that in each V_j, u is approximated by a constant function whose value is denoted here again by u_j. In order to take into account the essential boundary condition $u(0) = 0$, we set $u_0 = 0$. In Figure 1.4, illustrations of these definitions and notations are supplied.

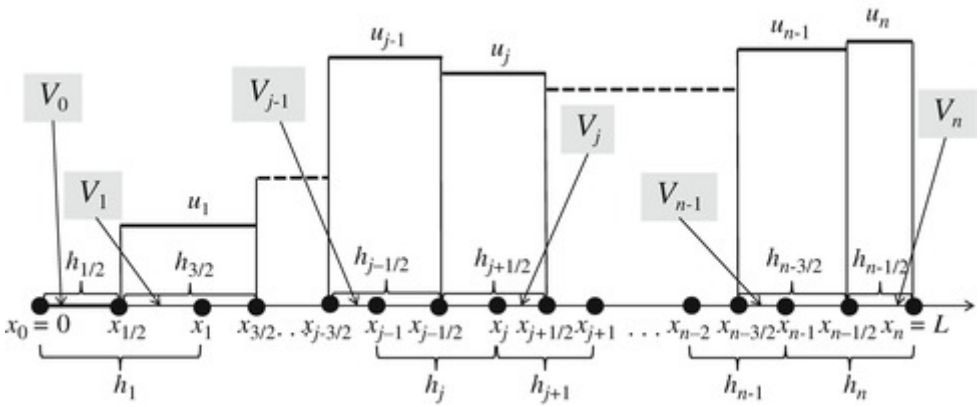

Figure 1.4 CVs V_j for the Vertex-centred FVM and approximations u_j of u.

Now we apply the conservation principle in (P_2) to each V_j. For the sake of simplicity, we first assume that p is constant and approximate u' at the left end of V_j by $(u_j - u_{j-1})/h_j$ for $j = 1, 2, \ldots, n$ and at the right end of V_j by $(u_{j+1} - u_j)/h_{j+1}$ if $1 \leq j < n$. The latter expression is replaced by zero if $j = n$. This gives rise to the following:

Discrete balance equation in V_j for a constant p (1.24)

For $j < n$:

$$p\left[\frac{(u_j - u_{j-1})}{h_j} - \frac{u_{j+1} - u_j}{h_{j+1}}\right] + \int_{x_{j-1/2}}^{x_{j+1/2}} [qu_j - f]dx = 0;$$

For $j = n$:

$$p\frac{u_j - u_{j-1}}{h_j} + \int_{x_{j-1/2}}^{x_j} [qu_j - f]dx = 0;$$

with $u_0 = 0$.

This is again a SLAE with n equations and unknown vector $\vec{u}_h = [u_1, u_2, \ldots, u_n]^T$. Assuming that both f and q are continuous at the mesh nodes, let us apply a modified Gaussian mid-point

rule (see e.g. [29] or [208]) to approximate their respective integrals on the left and right side of the first equation of (1.24). More precisely, the mid-point of the interval $(x_{j-1/2}, x_{j+1/2})$ for $j < n$ (resp. $(x_{n-1/2}, x_n)$ for $j = n$) is shifted to x_j (resp. x_n) to obtain $q(x_j)h_{j-1/2}$ and $f(x_j)h_{j-1/2}$ for $j < n$ (resp. $q(x_n)h_n/2$ and $f(x_n)h_n/2$ for $j = n$). In doing so, we come up with a scheme derived from equation (1.24) that can be written in the form of a SLAE with an $n \times n$ matrix $A_h = \{a_{j,i}\}$ and right-side vector $\vec{b}_h = \{b_j\}$ derived from equation (1.24). A_h is a symmetric matrix whose entries vanish except $a_{j,j}$ and $a_{j,j-1} = a_{j-1,j}$, for $j = 1, 2, \ldots, n$. These nonzero entries together with the components b_j of $\vec{b}_h$ are the same as for the FEM in the same case, provided the trapezoidal rule is employed to approximate the integrals in eachelement of the mesh, assuming of course that both q and f are continuous (cf. Section 1.2). This assertion can be verified as Exercise 1.3.

On the other hand, if p is an arbitrary continuous function, in the FV analog of (P_2) we approximate $[pu']'(x_{j-1/2})$ by $p(x_{j-1/2})(u_j - u_{j-1})/h_j$. Then, instead of equation (1.24), we have the

Discrete balance equation in V_j for a continuous p (1.25)

For $1 \leq j < n$:

$$\frac{p(x_{j-1/2})(u_j - u_{j-1})}{h_j} - \frac{p(x_{j+1/2})(u_{j+1} - u_j)}{h_{j+1}} + \int_{x_{j-1/2}}^{x_{j+1/2}} [qu_j - f]dx = 0;$$

For $j = n$:

$$\frac{p(x_{j-1/2})(u_j - u_{j-1})}{h_n} + \int_{x_{j-1/2}}^{x_j} [qu_j - f]dx = 0$$

with $u_0 = 0$.

Referring to Figure 1.4, and as seen in Section 4.4, in the FVM approach it is generally considered that unknown values of u are approximated at one point per CV, which is called the CV's **representative point**. The corresponding approximations u_j are extended to the whole CV, thereby giving rise to the piecewise constant approximating function u_h. In the case of the Vertex-centred method, the representative point of CV V_j is x_j, for $j = 1, 2, \ldots, n$.

Now, if the same quadrature rules as in the case of a constant p are employed to compute the integrals involving q and f, that is, the trapezoidal rule for the FEM and the appropriate one-point rules for the FVM, then equation (1.25) corresponds to the $\mathcal{P}_1$ FE scheme, provided the mid-point rule is used to approximate the integrals $\int_{x_{j-1}}^{x_j} p[\varphi_j']^2 dx$ or $\int_{x_{j-1}}^{x_j} p[\varphi_j'\varphi_{j-1}']dx$ for $1 \leq j \leq n$. The reader may easily check that this assertion holds true as Exercise 1.4. Moreover,

the resulting modification of equation (1.25) suggests a possible way to discretise equation (P_1) by the FDM with a non-uniform grid for a non-constant p. Summarising, in this case, all three schemes reduce to the following:

Unified FD–FE–FV scheme to solve (P_1) for continuous p and q (1.26)

For $1 \leq j < n$:

$$\frac{p(x_{j-1/2})(u_j - u_{j-1})}{h_j h_{j+1/2}} - \frac{p(x_{j+1/2})(u_{j+1} - u_j)}{h_{j+1} h_{j+1/2}} + u_j q(x_j) = f(x_j);$$

For $j = n$:

$$\frac{p(x_{j-1/2})(u_j - u_{j-1})}{h_j^2} + u_j \frac{q(x_j)}{2} = \frac{f(x_j)}{2};$$

with $u_0 = 0$.

As a conclusion, at least in the one-dimensional case, the Vertex-centred FVM is practically equivalent to the $\mathcal{P}_1$ FE scheme with the same mesh, or to the Three-point FD scheme with a non-uniform grid, whose points are the mesh nodes.

Incidentally, such an equivalence also applies to the case of a discontinuous p, as long as $\mathcal{T}_h$ is constructed in such a way that the set $\{y_1, y_2, \ldots, y_M\}$ of discontinuity points of p is a subset of the set of mesh nodes (i.e. of grid points). In doing so, p is necessarily uniquely defined at the mid-points of each element, and we may apply the balance equation (1.25). Then, here again, if we approximate the integrals of $f\varphi_i$ and $q\varphi_j\varphi_i$ by the trapezoidal rule; approximate $\int_{x_{j-1/2}}^{x_{j+1/2}} q dx$, $\int_{x_{j-1/2}}^{x_{j+1/2}} f dx$, $\int_{x_{n-1/2}}^{x_n} q dx$ and $\int_{x_{n-1/2}}^{x_n} f dx$ by the 'shifted' mid-point quadrature rule specified above; and use the mid-point quadrature rule to approximate the integral of $p\varphi_j'\varphi_i'$ in each element, we realise that in this case, too, the resulting FE scheme and the Vertex-centred FV scheme derived from equation (1.25) coincide. It is interesting to note again that this scheme suggests an appropriate way to discretise (P_1) by the FDM in the case where p has discontinuities at grid points.

1.4.2 The Cell-centred FVM

We recall that, in this case, the CVs are the intervals T_j of the mesh $\mathcal{T}_h$ (also called **cells**). Then, in each cell, the function u is approximated by a constant function equal to $u_{j-1/2}$. In Figure 1.5, these definitions and notations are illustrated.

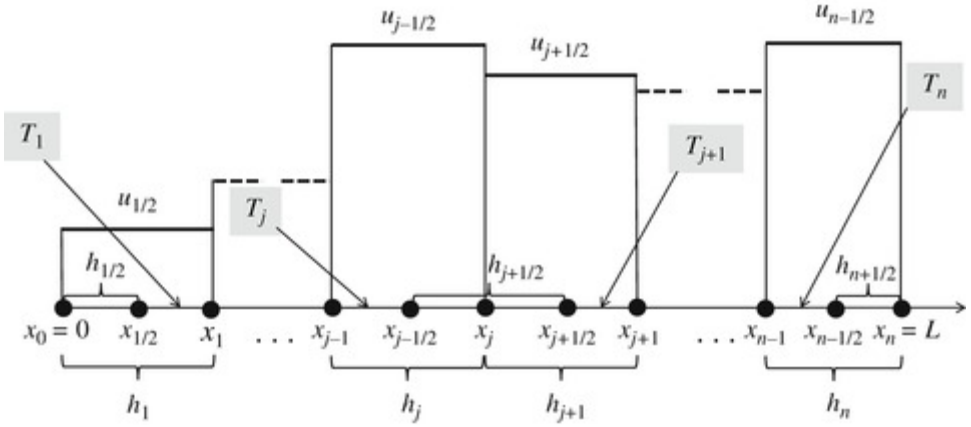

Figure 1.5 CVs T_j for the Cell-centred FVM and approximations $u_{j-1/2}$ of u.

By definition, the representative point of CV T_j is $x_{j-1/2}$ for $j = 1, \ldots, n$. We also set $x_{-1/2} = x_0$ and $x_{n+1/2} = x_n$ because, for convenience, we will consider an extended set of representative points by adding $x_{-1/2}$ and $x_{n+1/2}$, respectively, for fictitious CVs $T_0 := \{0\}$ and $T_{n+1} : [L, L + h_n]$ lying outside $(0, L)$. The value of u_h equals zero in the former and coincides with $u_{n-1/2}$ in the latter, as will be discussed here. Notice that, in the scheme, $\frac{u_{j+1/2}-u_{j-1/2}}{h_{j+1/2}}$ stands for the first-order derivatives of u at the cell end-point x_j, for $j = 1, 2, \ldots, n-1$; at $x = x_0$, this derivative is approximated by $\frac{u_{1/2}-u_0}{h_{1/2}}$, with $h_{1/2} := h_1/2$ and $u_0 = 0$, which is a way to take into account the boundary condition $u(x_0) = 0$. Finally, we have to set to zero a suitable approximation of $u'(L)$. Like in the FDM, we could add an approximation $u_{n+1/2}$ of u at the fictitious representative point $L + h_n/2$ of T_{n+1} and then enforce $\frac{u_{n+1/2}-u_{n-1/2}}{h_n} = 0$. However, here this is needless for it suffices to take a zero flux on the right end of CV T_n, and this boundary condition will be automatically incorporated into the discrete analog of (P$_2$) (cf. (1.27)). Thus, assuming that p is continuous, this gives rise to the following:

Discrete balance equation in CV T_j for a non-constant p (1.27)

For $j = 1$:

$$p(x_{j-1})\frac{u_{j-1/2}-u_{j-1}}{h_{j-1/2}} - p(x_j)\frac{u_{j+1/2}-u_{j-1/2}}{h_{j+1/2}} + \int_{x_{j-1}}^{x_j}[qu_{j-1/2} - f]dx = 0;$$

For $1 < j < n$:

$$p(x_{j-1})\frac{u_{j-1/2}-u_{j-3/2}}{h_{j-1/2}} - p(x_j)\frac{u_{j+1/2}-u_{j-1/2}}{h_{j+1/2}} + \int_{x_{j-1}}^{x_j}[qu_{j-1/2} - f]dx = 0;$$

For $j = n$:

$$p(x_{j-1})\frac{u_{j-1/2}-u_{j-3/2}}{h_{j-1/2}} + \int_{x_{j-1}}^{x_j}[qu_{j-1/2} - f]dx = 0;$$

with $u_0 = 0$.

It is a common practice to use the mid-point quadrature rule in order to approximate the integrals of both q and f, if these functions are continuous in each cell. On the other hand, in situations of practical interest, p is often discontinuous at M points of a set $\{y_1, y_2, \ldots, y_M\}$. However, in general, physics of the phenomenon being modelled requires that the **flux** pu' across those discontinuity points of p be continuous. One of the features of the FVM is to mimic conservation properties satisfied by the exact solution of the model equation. In the case under study, this can be achieved if we proceed as follows. Assume that the mesh is constructed in such a way that the intersection of the set of discontinuity points of p with the set $\{x_1, x_2, \ldots, x_{n-1}\}$ is not empty. Let us denote by p_j^- and p_j^+ the values of p at mesh node x_j from the left and from the right, respectively, for $j = 1, 2, \ldots, n-1$. We set $\tilde{p}_j := (p_j^- + p_j^+)/2$ completed with $\tilde{p}_0 = p(x_0)$ and $\tilde{p}_n = p(x_n)$. Of course, if p is continuous at x_j, then $\tilde{p}_j = p(x_j)$.

Then, we may set up the following Cell-centred FV approximation of (P_1), which obviously extends to functions p everywhere continuous in $[0, L]$, namely:

Cell-centred FV scheme to solve (P_1) accommodating discontinuous p (1.28)

For $j = 1$:

$$\tilde{p}_{j-1}\frac{u_{j-1/2}-u_{j-1}}{h_j h_{j-1/2}} - \tilde{p}_j\frac{u_{j+1/2}-u_{j-1/2}}{h_j h_{j+1/2}} + q(x_{j-1/2})u_{j-1/2} = f(x_{j-1/2});$$

For $1 < j < n$:

$$\tilde{p}_{j-1}\frac{u_{j-1/2}-u_{j-3/2}}{h_j h_{j-1/2}} - \tilde{p}_j\frac{u_{j+1/2}-u_{j-1/2}}{h_j h_{j+1/2}} + q(x_{j-1/2})u_{j-1/2} = f(x_{j-1/2});$$

For $j = n$:

$$\tilde{p}_{j-1}\frac{u_{j-1/2}-u_{j-3/2}}{h_j h_{j-1/2}} + q(x_{j-1/2})u_{j-1/2} = f(x_{j-1/2});$$

with $u_0 = 0$.

Remark 1.3

In reference [68], a more elaborated treatment of discontinuous p is proposed, which supposedly has better approximation properties.

The reader may exhibit the entries of the $n \times n$ matrix A_h and of the n-component right-side vector $\vec{b}_h$ of the SLAE with n equations corresponding to equation (1.28), whose vector of n unknowns is $[u_{1/2}, u_{3/2}, u_{5/2}, \ldots, u_{n-1/2}]^T$ (Exercise 1.5). The fact that this system has a unique solution can be established by means of a simple variant of the argument employed in the case of the FDM for an equally spaced grid, as long as the CVs have a fixed measure h and p is constant. Indeed, it suffices to let $u_{j-1/2}$ play the role of u_i in that case. If the cells have a variable measure and p is not constant and is eventually discontinuous, at the price of some additional though straightforward calculations, the existence and uniqueness of a solution can also be shown to hold (cf. Exercise 1.6).

Remark 1.4

*The conservation of the **numerical fluxes** from both sides of the cell end-points enforced in scheme (1.28) is a key property to establish its reliability. This means that it is capable of yielding meaningful approximations of the exact solution. In Chapters 5 and 7, we will see this in more detail for related FV schemes.*

1.4.3 Connections to the Other Methods

To conclude this section, we observe that the FVMs, we have studied so far can be viewed as an FEM of a particular type, in the sense that both are based on a variational formulation. The only real difference is that such a variational formulation applies only to the approximate problem (i.e. the discrete problem) and has no real counterpart in the framework of the exact problem (i.e. the continuous problem). Let us see how this works.

Starting from the Cell-centred FVM, we consider a space U_h associated with the mesh $\mathcal{T}_h$, consisting of functions whose value is a constant w_j in each element T_j. Then, we introduce the concept of **discrete derivative** of a function $w \in U_h$ in each interval $(x_{j-3/2}, x_{j-1/2})$. This quantity is defined by $w'^h := (w_j - w_{j-1})/h_{j-1/2}$, for $j = 2, \ldots, n$ completed with $w'^h(x) := w_1/h_{1/2}$ for $x \in (0, x_{1/2})$ and $w'^h = 0$ for $x \in (x_{n-1/2}, L)$. In doing so, we endeavour to search for a function $u_h \in U_h$ such that

$$\int_0^L [pu_h'^h w'^h + qu_h w]dx = \int_0^L fwdx \; \forall w \in U_h. \tag{1.29}$$

Let now ψ_i be a *shape function* of U_h, namely, the function whose value is onein T_i and zero elsewhere. Clearly enough, every function in $w \in U_h$ can be expanded as $w(x) = \sum_{i=1}^n w_i \psi_i(x) \;\; \forall x \in (0, L)$. Hence, using the description of the FEM as a guide, we take in equation (1.29) w successively equal to ψ_i. Clearly, for $i = 2, \ldots, n-1$, $\psi_i'^h(x) = 1/h_{i-1/2}$ if $x \in (x_{i-3/2}, x_{i-1/2})$, $\psi_i'^h(x) = -1/h_{i+1/2}$ if $x \in (x_{i-1/2}, x_{i+1/2})$ and $\psi_i'^h(x) = 0$ elsewhere. Moreover, $\psi_1'^h(x) = 1/h_{1/2}$ if $x \in (0, x_{1/2})$, $\psi_1'^h(x) = -1/h_{3/2}$ if $x \in (x_{1/2}, x_{3/2})$ and $\psi_1'^h(x) = 0$ otherwise. Finally, $\psi_n'^h(x) = 1/h_{n-1/2}$ if $x \in (x_{n-3/2}, x_{n-1/2})$ and $\psi_n'^h(x) = 0$ otherwise.

Therefore, from equation (1.29), we derive

$$\begin{cases} \int_0^{x_{1/2}} p\, u_h'^h/h_{1/2}dx - \int_{x_{1/2}}^{x_{3/2}} p\, u_h'^h/h_{3/2}dx + \int_0^{x_1} q\, u_h dx = \int_0^{x_1} f dx; \\ \text{For } i = 2, \ldots, n-1: \\ \int_{x_{i-3/2}}^{x_{i-1/2}} p\, u_h'^h/h_{i-1/2}dx - \int_{x_{i-1/2}}^{x_{i+1/2}} p\, u_h'^h/h_{i+1/2}dx + \int_{x_{i-1}}^{x_i} q\, u_h dx = \int_{x_{i-1}}^{x_i} f dx \\ \int_{x_{n-3/2}}^{x_{n-1/2}} p\, u_h'^h/h_{n-1/2}dx + \int_{x_{n-1}}^{x_n} q\, u_h dx = \int_{x_{n-1}}^{x_n} f dx. \end{cases} \tag{1.30}$$

Now, noting that $u_h = \sum_{j=1}^{n} u_j \psi_j$ and recalling the definition of $u_h'^h$, from equation 1.30, we obtain

$$\begin{cases} \int_0^{x_{1/2}} \frac{p u_1}{h_{1/2}^2} dx - \int_{x_{1/2}}^{x_{3/2}} \frac{p(u_2 - u_1)}{h_{3/2}^2} dx + \int_0^{x_1} q\, u_h dx = \int_0^{x_1} f dx; \\ \text{For } 1 < i < n; \\ \int_{x_{i-3/2}}^{x_{i-1/2}} \frac{p(u_i - u_{i-1})}{h_{i-1/2}^2} dx - \int_{x_{i-1/2}}^{x_{i+1/2}} \frac{p(u_{i+1} - u_i)}{h_{i+1/2}^2} dx + \int_{x_{i-1}}^{x_i} q\, u_h dx = \int_{x_{i-1}}^{x_i} f dx \\ \int_{x_{n-3/2}}^{x_{n-1/2}} \frac{p(u_n - u_{n-1})}{h_{n-1/2}^2} dx + \int_{x_{n-1}}^{x_n} q\, u_h dx = \int_{x_{n-1}}^{x_n} f dx. \end{cases} \tag{1.31}$$

Further assuming that p is continuous in $[0, x_{1/2})$ and in $(x_{i-3/2}, x_{i-1/2})$ for $i = 2, \ldots, n$, if we use a one-point quadrature formula to integrate p in these intervals, and recalling that $u_h = u_i$ in T_i, we obtain

$$\begin{cases} \frac{p(0)u_1}{h_{1/2}} - \frac{p(x_1)(u_2 - u_1)}{h_{3/2}} + u_1 \int_0^{x_1} q\, dx = \int_0^{x_1} f dx; \\ \text{For } 1 < i < n; \\ \frac{p(x_{i-1})(u_i - u_{i-1})}{h_{i-1/2}} - \frac{p(x_i)(u_{i+1} - u_i)}{h_{i+1/2}} + u_i \int_{x_{i-1}}^{x_i} q dx = \int_{x_{i-1}}^{x_i} f dx \\ \frac{p(x_{n-1})(u_n - u_{n-1})}{h_{n-1/2}} + u_n \int_{x_{n-1}}^{x_n} q dx = \int_{x_{n-1}}^{x_n} f dx. \end{cases} \tag{1.32}$$

Now, if we renumber the u_is in such a manner that u_i becomes $u_{i-1/2}$ and further apply one-point numerical quadrature to compute the integrals of q and f, we find out that equation (1.32) is nothing but the Cell-centred FV scheme (1.28). As a conclusion, the Cell-centred FVM is derived from a suitable weak formulation of equation (P_1) at a discrete level, using the simplest possible type of representation of the solution, namely, piecewise constant functions.

As for the Vertex-centred FVM, a similar conclusion applies, but the argument is a little more tricky. In this case, we may consider a piecewise linear approximate solution $u_h := \sum_{j=1}^{n} u_j \varphi_j$, to be determined by solving a discrete variational problem similar to that of equation (1.22), in which φ_i is replaced by $\bar{\varphi}_i$, where $\bar{\varphi}_i$ is a piecewise constant function whose value is one in the CV $(x_{i-1/2}, x_{i+1/2})$ and zero elsewhere for $i = 1, 2, \ldots, n-1$; and $\bar{\varphi}_n$ equals one in $(x_{n-1/2}, L)$ and zero elsewhere. For such functions, we also define a discrete derivative analogous to the one of functions in U_h by $\bar{\varphi}_i^{'h}(x) = 1/h_i$ if $x \in (x_{i-1}, x_i)$ and $\bar{\varphi}_i^{'h} = -1/h_{i+1}$ if $x \in (x_i, x_{i+1})$; $\bar{\varphi}_i^{'h}(x) = 0$ elsewhere in $(0, L)$ for $i = 1, 2, \ldots, n-1$; and $\bar{\varphi}_n^{'h}(x) = 1/h_n$ for $x \in (x_{n-1/2}, L)$ and $\bar{\varphi}_n^{'h}(x) = 0$ elsewhere. After some rather fastidious calculations, which we voluntarily skip for the sake of conciseness, we conclude that the discrete variational problem analogous to equation (1.22) is

$$\begin{cases} \text{Find } u_j \text{for } j = 1, 2, \ldots, n \text{ such that} \\ \sum_{j=1}^{n} u_j \int_0^L [p\varphi_j^{'} \bar{\varphi}_i^{'h} + q\varphi_j \bar{\varphi}_i] dx = \int_0^L f\bar{\varphi}_i dx \text{ for } i = 1, 2, \ldots, n. \end{cases} \tag{1.33}$$

Equation (1.33) is nothing but the Vertex-centred FV scheme (1.25), provided suitable one-point quadrature rules are employed to evaluate the integrals in equation (1.33). For this reason, the resulting scheme is also called the finite volume–finite element scheme (see e.g. [68]).

1.5 Handling Nonzero Boundary Conditions

The model boundary value problem we have studied in this chapter so far has only **homogeneous boundary conditions** (i.e. zero boundary conditions). Although such a situation effectively occurs in the case of the elongational deformations of a bar, in many practical applications the boundary conditions are **inhomogeneous**. In the case of equation (P_1), this corresponds to

$$\begin{cases} -[pu^{'}]^{'} + qu = f \text{ in } (0, L); \\ u(0) = a; \\ p(L)u^{'}(L) = b, \end{cases} \tag{1.34}$$

where a and b are given real numbers. In physical terms, this means that a displacement equal to a is prescribed at the bar's left end, and that a force of intensity equal to b acts on its right end.

Let us briefly examine the modifications that have to be introduced in the numerical schemes studied in this chapter, in order to take into account such data. We expect that, as a by-product,

this will give the necessary insight on how to deal with other types of boundary conditions, homogeneous or not.

To begin with, we consider the case where only the **Dirichlet boundary condition** is inhomogeneous, namely, $u(0) = a$. As far as the FDM and both types of FVM are concerned, this is simply a matter of setting $u_0 = a$ instead of $u_0 = 0$ in the corresponding schemes. Whatever the case, this implies that only the right side of the first equation of the corresponding scheme will have to be adjusted by the addition of a coefficient multiplied by a. The reader will face no difficulty to identify such a coefficient in each case.

As for the FEM, the necessary modification is a little more subtle, although in the end the same kind of conclusion applies: setting again $u_0 = a$ and recalling the function φ_0, we first rewrite the approximate solution as

$$u_h = \sum_{j=0}^{n} u_j \varphi_j.$$

Now plugging this expression of u_h into equation (1.21) and taking $v = \varphi_i$, we note that for $i = 1$ only, a new term $b_0 := u_0 \int_0^{x_1} [p\varphi_0' \varphi_1' + q\varphi_0\varphi_1]dx$ will appear on the left side of the resulting equation. Hence, such an inhomogeneous boundary condition is taken into account by simply adding to the original right-side b_1 the value $-b_0$. Using quadrature formulae as prescribed in Section 1.3, we conclude that the correction $-b_0$ plays the same role as in the FDM and the FVM.

Next, we switch to the case of the inhomogeneous **Neumann boundary condition** $[pu']'(L) = b$. In the case of the FDM, this is only a matter of redefining the fictitious approximate value u_{n+1} by means of the natural relation:

$$p(L)\frac{u_{n+1} - u_{n-1}}{2h_n} = b.$$

Then, the value $u_{n+1} = 2bh_n/p(L) + u_{n-1}$ is plugged into the nth equation of scheme (1.12) or (1.15). As we should point out, in the case of uniform grids, at least for academic purposes, it is advisable to improve the accuracy of this approximation. For instance, one might adapt to an inhomogeneous Neumann boundary condition at $x = L$ the trick to be used in Chapter 2, in the homogeneous case.

In the case of the FVM, we proceed similarly by modifying only the last equation of the Vertex-centred or the Cell-centred scheme (1.25) or (1.28) (cf. Exercise 1.8). The natural thing to do here

is to subtract (resp. add) to the left (resp. right) side of this equation the quantity b, in order to take into account the term stemming from the integration of $-[pu']'$ in the rightmost CV. The reader might easily determine in Exercise 1.7 the final form of the thus-modified right-side vector $\vec{b}_h$ in the case of each method (FDM or FVM).

Finally, we consider the FEM. First of all, if we want the inhomogeneous boundary condition $p(L)u'(L) = b$ to be implicitly satisfied in the variational formulation (P_3), we have to add to the right side the term $bv(L)$. Indeed, when we integrate by parts the first term of the integral on the left side of this equation in order to recover (P_1), we first obtain $-[pu']' + qu = f$ by choosing v such that $v(L) = 0$, like in the homogeneous case. Then, using this equation and taking an arbitrary v in V such that $v(L) \neq 0$, we obtain $[pu'v](L) = bv(L)$. This obviously implies that $[pu'](L) = b$. Such a modification is carried to the discrete analog (1.21) of (P_3), which leads to a correction in the last equation of (1.22) only. This is because on the one hand $b\varphi_i(L) = 0$ for every i less than n, and on the other hand $b\varphi_n(L) = b$. In short, in order to take into account the inhomogeneous Neumann boundary condition $[pu'](L) = b$, here again it is necessary to correct only the last equation of (1.22) by adding b to former b_n.

Several variants of (P_1) aimed at completing the above presentation are considered in Section 6.1. More precisely, boundary conditions other than Dirichlet at $x = 0$ and Neumann at $x = L$ are prescribed, among which lie Dirichlet or Neumann boundary conditions at both ends. Robin boundary conditions, that is, conditions of the type $u'(0) + cu(0) = a$ and/or $u'(L) + du(L) = b$ for given real numbers a, b, c, d, are also treated.

1.6 Effective Resolution

So far, we have described numerical solution schemes without caring about their practical implementation. We know that all of them lead to a SLAE, whose solution for particular field data generates the numerical values the schemes are designed to provide. To close this chapter, we consider some relevant aspects of this process in a twofold manner. First of all, we address some generalities on the form of the SLAEs to be solved, thereby drawing some conclusions on right choices of solution methods. Next, we give some numerical examples by solving problems of the form (P_1) with the Cell-centred FVM.

1.6.1 Solving SLAEs for one-dimensional problems

As we saw in this chapter, all the three discretisation methods under study reduce to the solution of a SLAE $A_h\vec{u}_h = \vec{b}_h$ with n unknowns and n equations. Since we could assert beforehand that all these systems have a unique solution, in principle solving them is no problem. However, in practical terms, it is mandatory to take into account the particular form of the matrix system; otherwise, the resolution cost could become excessive if n is very large. Indeed, A_h is a **sparse matrix**, akin to the case of most numerical methods for solving boundary value differential equations. Actually, here matrix A_h is a symmetric tridiagonal matrix since $a_{i,j} = 0$ whenever $|i - j| > 1$. In any event, a SLAE whose matrix is sparse should preferably be solved by methods that preserve its sparsity structure. This means that, in the underlying matrix manipulations, one should create the least possible, amount of new nonzero entries. Possibly, one should restrict the positions of nonzero entries to the original ones all the way. In this manner, storage requirements are reduced to a minimum, and computational costs do not increase unnecessarily.

In the class of iterative methods for solving SLAEs, in most cases it is possible to keep the matrix as it was at the beginning of the solution process. However, these methods are subject to the convergence (or not) of the successive approximations to the solution vector $\vec{u}_h$. In the framework of the FD, FE and FV schemes for the problem we considered, classical iterative methods for solving SLAEs such as **Gauss–Seidel**, **successive overrelaxation** and **conjugate gradient** do converge, since matrix A_h besides being symmetric is **positive-definite** and **strictly diagonally dominant** if $q > 0$. As we know, the former property means that all the eigenvalues of A_h are strictly positive, and the latter that $a_{i,i} > \sum_{j=1, j\neq i}^{n} |a_{i,j}|$, for every i in $\{1, 2, \ldots, n\}$. In both cases, the spectral radius of the matrix B_h corresponding to those iterative methods is strictly less than one, and hence the iterations converge. Referring to the preliminary section on linear algebra, we recall that B_h is the matrix such that $\vec{u}_h^k = B_h\vec{u}_h^{k-1} + \vec{c}_h$, where $\vec{u}_h^k$ are successive approximations of $\vec{u}_h$ for $k = 1, 2, \cdots$, and $\vec{c}_h$ is a suitable vector of $\Re^n$ derived from $\vec{b}_h$.

On the other hand, one might prefer the use of direct methods, such as **Crout's method** or **Cholesky's method**. As we know, they always lead to the exact solution vector, except for round-off errors. This is because direct methods are based on the decomposition of the original matrix A_h into the product of two matrices S_h and R_h, S_h being lower triangular and R_h upper

triangular. In the case of **banded matrices**, such as tridiagonal matrices, the nonzero entries of both factor matrices also lie within a band. In the absence of **pivoting** in **Gaussian elimination**, the latter has the same width as the band of the original matrix. This means that, for a tridiagonal matrix A_h, both S_h and R_h are bi-diagonal matrices[6]. The reader can easily check this assertion by applying the formulae supplied in the preliminary section on linear algebra. In short, the system is rewritten as $S_h R_h \vec{u}_h = \vec{b}_h$, or into the form of two systems $R_h \vec{u}_h = \vec{z}$ and $S_h \vec{z} = \vec{b}_h$, to be solved one after the other in a straightforward manner, since both S_h and R_h are triangular matrices.

Unfortunately, there is little room here for being more specific about all the relevant aspects of methods for solving the SLAE resulting from the application of FD, FE or FV discretisations to solve a boundary value ODE like (P_1). That is why we refer to the vast literature on the subject, including classical books such as references [199], [63] and [121] (volumes 1 and 2), among many others. Nevertheless, we shall go back to large SLAE solving in connection with numerical methods for PDEs in Section 4.4.

1.6.2 Example 1.1: Numerical Experiments with the Cell-centred FVM

In Chapter 2, we shall carry out a numerical study of the Three-point FDM, the $\mathcal{P}_1$ FEM and the Vertex-centred FVM, in the light of the reliability results that we will formally establish for the three methods. Here, we check the numerical behaviour of the Cell-centred FVM (equation (1.28)) in the solution of two test problems for the same ODE as (P_1) taking $L = 1$, eventually assorted with inhomogeneous boundary conditions (equation (1.34)). We consider only integrable f, because otherwise the integral of f in some CV may not be finite, and thus the method could not be applied as such.

The main point in the numerical results presented here is the observation of the maximum absolute error decay as the meshes vary, in such a way that the maximum of the mesh steps h_j decreases as n increases. The absolute errors are computed at the method's representative points (i.e. at the cell centres). We use two families of non-uniform meshes with $n = 2^m \times 10$ CVs, for $m = 0, 1, 2, 3, 4$. In the first family, there are only two different values of h_j, namely, $h_j = 0.8/n$ for odd values of j and $h_j = 1.2/n$ for even values of j. In the second family of meshes, we took h_j randomly equal to one out of three possible values proportional to 1, 2 and 3, for at least one CV, and thus for $(n-2)$ CVs at the most. We denote by F_1 the first family of

meshes and by F_2 the second family. Gaussian elimination was used to solve the underlying SLAEs. In Tables 1.1 and 1.2, the displayed values were rounded to the fifth significant figure.

Table 1.1 Maximum absolute errors at CV representative points for test problem 1

	10	20	40	80	160

Table 1.2 Maximum absolute errors at CV representative points for test problem 2

	10	20	40	80	160

Test-problem 1: Continuous p and q are considered, namely, $p(x) = q(x) = e^x$. Taking $f \equiv 1$, $a = 1$ and $b = -1$, the exact solution is given by $u(x) = e^{-x}$.

As one can infer from Table 1.1 for both families of meshes, the maximum absolute errors are roughly divided by four as n is multiplied by two. According to the concepts to be introduced in Chapter 2, this behaviour suggests that even for non-uniform meshes the Cell-centred FVM is *second-order convergent* in the pointwise sense, although such a scheme's property will not be formally established in this book.

Test-problem 2: We consider discontinuous p and q, namely, $p(x) = 1/2$ if $0 \leq x \leq 1/2$ and $p(x) = 1$ if $1/2 < x \leq 1$, and $q(x) = x/u(x)$ where u is the manufactured solution given by $u(x) = x(2 - x)$ if $0 \leq x \leq 1/2$ and $u(x) = -x^2/2 + x - 3/8$ if $1/2 < x \leq 1$. Notice that in the case of F_1, the discontinuity point of p is always an end-point of two neighbouring cells. Incidentally, according to equation (1.28), whenever p is discontinuous at x_j we take $\tilde{p}_j = [p_j^- + p_j^+]/2$ instead of $p(x_j)$ in the scheme. Watching Table 1.2, we realise that the errors in the case of non-uniform meshes with CVs having systematically $x = 1/2$ as an end-point decrease by a factor of two from a level to the other. On the other hand, surprisingly enough, this factor doubles in the case of the randomly generated meshes, for which the discontinuity point of p is never a CV end-point x_j. This suggests that the latter is a better approach to handle a discontinuous p. However, results of this nature must be the object of more careful and rigorous analyses, such a those conducted in Chapter 2.

1.7 Exercises

1.1 Prove that system (1.16) has a unique solution.

1.2 Derive the FD scheme to solve equation (P_1) with continuously varying p, q and f based on a non-uniform grid, equivalent to a $\mathcal{P}_1$ FE scheme derived as follows: the trapezoidal rule is employed to compute all the underlying integrals in the intervals between two neighbouring grid points.

1.3 Assume that in the $\mathcal{P}_1$ FE formulation, the trapezoidal quadrature rule is employed to compute the integrals involving q and f. Find the one-point quadrature formula to compute the integrals of q and f in the formulation (1.24) of the Vertex-centred FVM, for which both methods correspond to the same SLAE.

1.4 In the case of a continuously varying p, assume that the integrals $\int_{x_{j-1}}^{x_j} p[\varphi_j']^2 dx$, $\int_{x_j}^{x_{j+1}} p[\varphi_j']^2 dx$ and $\int_{x_{j+1}}^{x_j} p[\varphi_j'\varphi_{j+1}']dx$ are approximated by the mid-point quadrature rule for all appropriate j. Check that the equivalence under the other conditions specified in the previous exercise, between the resulting $\mathcal{P}_1$ FEM and thus-modified FVM (equation (1.25)), which is equation (1.26), still holds true.

1.5 Exhibit the entries of the $n \times n$ matrix A_h and of the n-component right-side vector $\vec{b}_h$ of the SLAE with n equations corresponding to equation (1.28).

1.6 Prove the existence and uniqueness of a solution to equation (1.28).

1.7 Give the final form of the right-side vector $\vec{b}_h$ for the Three-point FDM (equation (1.15)) and the Vertex-centred FVM in case the boundary condition $pu'(L) = b$ holds.

1.8 Modify the FV scheme (equation (1.28)) in order to accommodate the inhomogeneous boundary conditions in equation (1.34).

1.9 For $L = 1$, consider a right side of the differential equation (P_1) equal to 1 plus the Dirac distribution δ_c, where $c \in (0, 1]$. Take $p \equiv 1$, $u(0) = a > 0$, $u'(1) = 0$ and $q = 1/s(x)$, where s is the strictly positive function in $[0, 1]$ defined by $s(x) = a + x$ if $0 \leq x < c$ and $s(x) = a + c$ if $1 \geq x \geq c$. Replace u by s in the variational formulation of problem (1.34), and show that s is the solution to this problem for $c < 1$. What happens if $c = 1$? Give an interpretation of this case in terms of inhomogeneous Neumann boundary conditions. Assume that the problem is being solved by the $\mathcal{P}_1$ FEM with a mesh constructed in such a way that c is the abscissa of a mesh node. Check that whatever the mesh satisfying such a condition the corresponding FE solution u_h equals s.[7]

Notes

1 In the present case, this means that the absolute value of the first-order derivative of u is small, say, nowhere greater than 1%.

2 The differential equation in (P_1) is the *Sturm equation* $a_0 u'' + a_1 u' + a_2 u = a_3$ for given functions a_0, a_1, a_2 and a_3 satisfying certain properties. The Sturm equation can be recast in the form appearing in (P_1) called the *Liouville form*. For this reason, it is sometimes referred to as a *Sturm–Liouville problem*. However, classically the latter stems from a PDE, that governs the free or the forced vibrations of a string (see e.g. [57]). This PDE known as the *equation of the vibrating string*, will be studied in Chapter 3. If the applied forces are periodic in time, it can be reduced to the ODE (P_1). However, in this case q is negative, which makes the problem fundamentally different from the one studied here.

3 More precisely, v is absolutely continuous.

4 In old days, this distribution of forces used to be called the 'Dirac function' $\delta_{\frac{L}{2}}$ multiplied by P, where $\delta_{\frac{L}{2}}$ would be such that $\delta_{\frac{L}{2}}(x) = 0, \ \forall x \in [0, L]$, $x \neq L/2$, $\delta_{\frac{L}{2}}(L/2) = +\infty$, and "$\int_0^L \delta_{\frac{L}{2}} dx" = 1$.

5 $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ if $i \neq j$.

6 As long as no partial pivoting is necessary in Gaussian eliminations, which in principle is the case of the matrices considered in this chapter.

7 As seen in Chapter 2, this property extended to any piecewise linear function establishes the method's consistency: if the solution to the differential equation is a continuous piecewise linear function and the mesh matches the transition points of its different linear pieces, then the underlying FEM yields the exact solution.