2

Qualitative Reliability Analysis

nothing at all takes place in the universe in which some rule of maximum or minimum does not appear.

Leonhard Euler

Since we are dealing with techniques providing approximations of the exact solution, the following fundamental question arises: How reliable are they in terms of accuracy? In this chapter, we endeavour to give an appropriate answer to this question in the framework of the model problem studied in the previous chapter. As a by product, the basic steps to be followed in order to handle more complex situations in the subsequent chapters will be disclosed.

Generally speaking, the user of numerical methods for differential equations should possibly know in advance what can be expected from the values obtained through the solution of the underlying discrete problems, as compared to the corresponding values of the exact solution. If we stick to a given discretisation lattice, it is not always easy to give a satisfactory answer to the above question. That is why, as a rule, the best way to ensure that a numerical method is capable of generating accurate approximations is to establish that the finer the discretisation lattice, the closer we come to the exact values we are searching for. Basically, this means that the number of calculation points attached to the grids or meshes increases indefinitely. However, it is advisable (and in some cases mandatory) that this process be accompanied by a certain uniformity criterion, at least for enhanced efficiency. Here we consider that, unless there is a specific reason to do otherwise, the points at which approximations are calculated, though increasing in number, do not become too dense in a certain region of the problem definition domain, to the detriment of any other. Mathematically, this is translated by a certain **uniformity** of a discretisation lattice. For instance, in case the one-dimensional problem (p1) is solved by the FDM with non-uniform grids, we will say we are working with a **quasi-uniform family of grids** if all the grids $\mathcal{G}_n = \{x_0, x_1, \cdots, x_{n-1}, x_n\}$ under consideration are such that the ratio between the minimum ρ and the maximum h over i of $h_i = |x_i - x_{i-1}|$, commonly known as the **grid size**, is bounded below by a constant $c > 0$ independent of the grid. c is called the **uniformity constant** of the family. In particular, this bound must hold as h goes to zero, that is, as n goes to infinity. In case

the FVM or the FEM is employed with meshes $\mathcal{T}_h$, taking the same definitions of h and ρ, we come up with the similar concept of a **quasi-uniform family of meshes** if $\rho/h \geq c > 0$ for every mesh in the family. Here again, the parameter h (called the **mesh size** for such a family) may become as small as we wish.

A very important quantity in any reliability study of a discretisation method is the **error of the numerical solution**. Basically, the error is a positive measure of the difference between the exact solution and the approximate solution. However, there are infinitely many ways to define such a measure. In order to be objective, it is advisable to choose a type of measure which is practical while being sufficiently representative of the approximation process as a whole. For instance, if only specific grid points are considered to measure the error, one might be disregarding relevant behaviors taking place elsewhere. That is why it is usually preferable to use global error measures, in the sense that somehow the contribution of the errors in every region of the problem definition domain is taken into account. This leads to the concept of **norm**, which can be viewed as a measure of distance between two elements in a given set. In the case of FD errors, the norm will be a measure of the distance between the vector $\vec{u}_h$ of approximate values of the solution at the grid points and the vector of exact values at the same points. In the case of the FEM, the norm will measure the distance between two functions, namely, the exact solution u and its approximation u_h. In the case of the FVM, one can use either the former or the latter, by defining the underlying piecewise constant approximating functions u_h.

Once we are equipped with an appropriate norm, the main purpose of the reliability analysis of a numerical method to solve differential equations of a certain type is to study the conditions under which the error of the approximate solution measured in this norm goes to zero, as the grid or mesh sizes tend to zero. This property is commonly called the **convergence of a numerical method**.

Chapter outline: In Section 2.1, we extend to more general vector spaces the concept of norm recalled in the preliminary section on linear algebra. The same is done therein for inner products. According to the celebrated **Lax–Richtmyer equivalence theorem** [123], the convergence in a certain norm of a numerical method to solve a linear differential equations occurs if the associated scheme is **stable** in the sense of the same norm and **consistent** with a **strictly positive order** in terms of the discretisation parameter(s). Section 2.2 is devoted to the presentation of the concept of **stability in norm**. Consistency and corresponding order are concepts exemplified by relation (1.9). They are formalised in Section 2.3. In Section 2.4, we explain why the combination of

stability and consistency with an strictly positive order results in convergence. Numerical examples given in Subsection 2.4.4 using the FDM, the FEM and the FVM introduced in Chapter 1 closes this chapter.

2.1 Norms and Inner Products

Before going into reliability analyses of discretisation methods of problem (p1), it is useful to formalise the concepts of norm and inner product in vector spaces, like the one of m-component real vectors, or of functions of a certain class, for example the set V defined in Section 1.3.

In abstract terms, we consider here that E is a real vector space with null element 0_E.

2.1.1 Normed Vector Spaces

E is said to be a **normed vector space**, or equivalently a space equipped with a **norm**, if E is a real vector space; and with every $e \in E$ we can associate a real number $\| e\|_E$ having the following properties:

i. $\|e\|_E \geq 0$, $\forall e \in E$ and $\|e\|_E = 0$ if and only if $e = 0_E$;

ii. $\|\alpha e\|_E = |\alpha| \|e\|_E$, $\forall \alpha \in \Re$, $\forall e \in E$;

iii. $\|e_1 + e_2\|_E \leq \|e_1\|_E + \|e_2\|_E$, $\forall e_1, e_2 \in E$,

where property (i) means positive-definiteness of the norm, (ii) is the homogeneity property and (iii) is known as the **triangle inequality**. Let us give some examples.

Three norms over the space $\Re^n$ of n-component real vectors $\vec{c} = [c_1, c_2, \cdots, c_n]^T$ were defined in the preliminary section on linear algebra, namely, the norms

$$\|\vec{c}\|_\infty \stackrel{def}{=} \max_{i\in\{1,2,\cdots,n\}} |c_i| \qquad \text{(maximum norm or } l^\infty - \text{norm)}$$

$$\|\vec{c}\|_1 \stackrel{def}{=} \sum_{i=1}^{n} |c_i| \qquad (l^1 - \text{norm})$$

$$\|\vec{c}\|_2 \stackrel{def}{=} \left[\sum_{i=1}^{n} |c_i|^2\right]^{\frac{1}{2}} \qquad \text{(Euclidean norm or } l^2\text{-norm).}$$

The Euclidean norm $\|\vec{c}\|_2$ is also known as the **modulus** of $\vec{c}$ and corresponds to the length of this vector in $\Re^2$ or $\Re^3$. Actually, this fact explains the denomination of property (iii), for in the case of $\Re^2$ the length of $\vec{c}+\vec{d}$ is necessarily bounded by the sum of the lengths of $\vec{c}$ and $\vec{d}$, as illustrated in Figure 2.1.

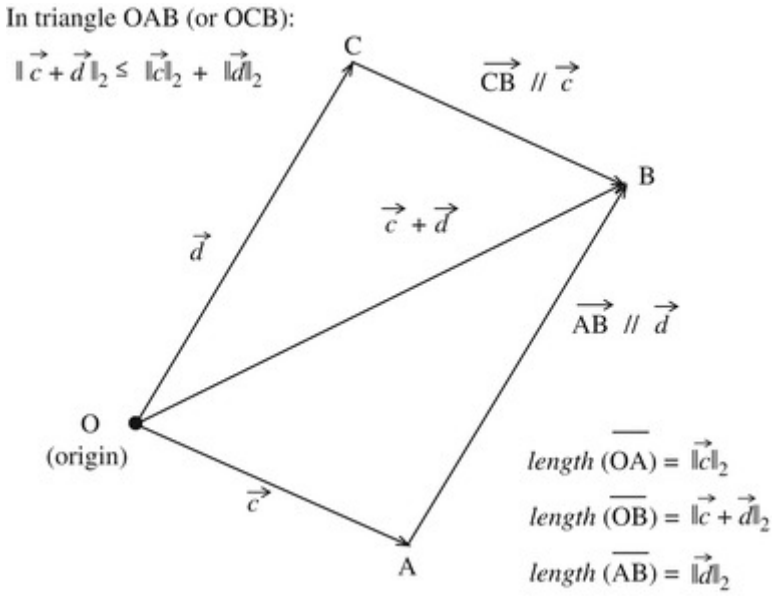


Figure 2.1 The triangle inequality for the Euclidean norm in $\Re^2$.

For coherence, we employ in this subsection the notation $\|\vec{c}\|_2$ for the modulus of a vector $\vec{c}\in\Re^n$. However, in the remainder of this text, we shall rather use the more popular and shorter notation $|\vec{c}|$ to represent such a quantity.

The following norms can be defined over the space of continuous functions f in the closed interval $[0,L]$, denoted by $C[0,L]$:

$$\|f\|_{0,\infty} \stackrel{def}{=} \max_{x\in[0,L]} |f(x)| \qquad \text{(maximum norm or } \mathcal{L}^\infty\text{-norm)}$$

$$\|f\|_{0,1} \stackrel{def}{=} \int_0^L |f(x)|dx \qquad (\mathcal{L}^1\text{-norm})$$

$$\|f\|_{0,2} \stackrel{def}{=} \left[\int_0^L |f(x)|^2 dx\right]^{\frac{1}{2}} \qquad (\mathcal{L}^2\text{-norm})$$

Remark 2.1

It is well-known that all six expressions above fulfil the conditions (i)–(iii). Actually, only property (i) for the norms $\|\cdot\|_{0,1}$ *and* $\|\cdot\|_{0,2}$ *cannot be established without fully exploiting the continuity of functions in* $C[0,L]$*. This is because such expressions do not really define norms for spaces containing discontinuous functions such as* $\mathcal{L}^2(0,L)$*. Indeed, if* $f \in \mathcal{L}^2(0,L)$ *and* $\|f\|_{0,2} = 0$*, then* f *is a function that vanishes almost everywhere in* $(0,L)$ *i.e., everywhere in this interval, except at most at points that form a set whose total measure is equal to zero–in short, a null set (e.g., a countable set of points). But since the norm* $\|\cdot\|_{0,2}$ *is quite natural and handy for square integrable functions, in order to overcome this difficulty, instead of* $\mathcal{L}^2(0,L)$ *itself, one usually works with a suitable associated quotient space* $L^2(0,L)$ *(see e.g. [206]). Nevertheless, as most authors do, one may identify* $\mathcal{L}^2(0,L)$ *and* $L^2(0,L)$ *by abusively regarding the latter as a space of functions not to be distinguished from each other, as long as they are identical almost everywhere in* $(0,L)$*.*

In complement to this brief introduction to the concept of norm, we observe that a real valued expression $|e|_E$ defined for elements e in a vector space E that fulfils (ii) and (iii) and such that $|e|_E \geq 0, \forall e \in E$ is called a **semi-norm** of E.

2.1.2 Inner Product Spaces

We shall be particularly concerned about a subclass of normed spaces whose norm is defined by means of an **inner product**. A real vector space E is said to be an **inner product space**, or, equivalently, to be equipped with an inner product, if with every pair $(d;e)$ of elements belonging to E we associate a real number $(d|e)$ called the inner product of d and e, satisfying the following properties:

I. $(e|e) \geq 0 \ \forall e \in E$ and $(e|e) = 0$ if and only if $e = 0_E$ (positive-definiteness);

II. $(d|e) = (e|d)\ \forall d, e \in E$ (symmetry);

III. $(\alpha_1 d_1 + \alpha_2 d_2|e) = \alpha_1(d_1|e) + \alpha_2(d_2|e)\ \forall \alpha_1, \alpha_2 \in \mathfrak{R}$, $\forall d_1, d_2, e \in E$ (linearity).

For every inner product, the so-called **Cauchy–Schwarz inequality** holds, namely:

$$|(d|e)| \leq \sqrt{(d|d)}\sqrt{(e|e)}\ \forall d, e \in E.$$

Indeed, owing to properties (I)–(III), we have $\forall d, e \in E$ and $\forall \alpha \in \mathfrak{R}$:

$$(d + \alpha e|d + \alpha e) = (d|d) + 2\alpha(d|e) + \alpha^2(e|e) \geq 0$$

Since the quadratic function of α on the left-hand side of the above inequality can only be non-negative for every α if

$$|2(d|e)|^2 - 4(d|d)(e|e) \leq 0,$$

the result follows.

To any given inner product $(\cdot|\cdot)$ corresponds a norm defined by

$$\|e\| \stackrel{def}{=} \sqrt{(e|e)}\ \forall e \in E$$

On the one hand, properties (i) and (ii) trivially hold for such a norm as a consequence of (I) and (II). On the other hand, it can be checked rather easily (cf. Exercise 2.1) that (iii) is also satisfied by virtue of the Cauchy–Schwarz inequality. For example, the norm $\|\vec{v}\|_2$ of $\vec{v} = [v_1, v_2, \cdots, v_n]^T \in \mathfrak{R}^n$ is derived from the inner product $(\vec{u}|\vec{v}) := \sum_{i=1}^{n} u_i v_i$. Moreover, we know from elementary geometry that for $n = 2$ or $n = 3$, if we denote by θ the angle between $\vec{u}$ and $\vec{v}$, it holds

$$(\vec{u}|\vec{v}) = \|\vec{u}\|_2\ \|\vec{v}\|_2 \cos\ \theta. \tag{2.1}$$

The norm $\|\cdot\|_{0,2}$ of $C[0, L]$, in turn, is associated with the following inner product:

$$(f|g)_0 \stackrel{def}{=} \int_0^L f(x)g(x)dx$$

Notice that the expression $(\cdot|\cdot)_0$ is not an inner product for $\mathcal{L}^2(0, L)$, since rigorously property (I) does not hold in this case. However, like in the case of the norm $\|\cdot\|_{0,2}$ (and also of $\|\cdot\|_{0,1}$), we may replace $\mathcal{L}^2(0, L)$ with $L^2(0, L)$, over which $(\cdot|\cdot)_0$ does become an inner product, provided we regard this space as a function space in the sense considered in the previous subsection.

As a by-product of the Cauchy–Schwarz inequality applied to $L^2(0,L)$, we can assert that whenever two functions f and g belong to $L^2(0,L)$, their product belongs to the space $L^1(0,L)$ of (Lebesgue) integrable functions in $(0,L)$. Indeed,

$$\pm\int_0^L f(x)g(x)dx \leq \sqrt{\int_0^L f^2(x)dx}\sqrt{\int_0^L g^2(x)dx}, \forall f,g \in L^2(0,L).$$

Every norm $\|\cdot\|$ associated with an inner product $(\cdot|\cdot)$ obeys the so-called *Parallelogram Law* (cf. [132]), namely:

$$\|d+e\|^2 + \|d-e\|^2 = 2(\|d\|^2 + \|e\|^2)\ \forall d,e \in E.$$

Indeed, the above identity can be verified in a straightforward manner by developing the two terms on its left side and making use of properties (II) and (III).

However, there are norms that correspond to no inner product. For instance, this is the case of the maximum norm. In order to justify this assertion, it suffices to consider the following counterexample:

Let $f(x) = x/L$ and $g(x) = (L-x)/L$ both belonging to $C[0,L]$.

We have $\|f\|_{0,\infty} = \|g\|_{0,\infty} = 1$. Moreover, $\|f+g\|_{0,\infty} = 1$ and $\|f-g\|_{0,\infty} = 1$. Therefore, the Parallelogram Law is violated.

Remark 2.2

Actually, the converse of the Parallelogram Law is also true: *If a norm* $\|\cdot\|$ *defined on a vector space* E *satisfies the Parallelogram Law for every* $e,d \in E$, *then it is necessarily associated with an inner product* $(\cdot|\cdot)$ *on the same space. More specifically, the latter is given by* $(d|e) = \frac{1}{4}(\|d+e\|^2 - \|d-e\|^2)$ *(see e.g. [28]).*

Remark 2.3

Some norms or inner products may not be well suited to a certain space, as the corresponding expression could be unbounded or undefined for subclasses of elements within this space. For instance, the expression $(f|g)_0$ *is not defined for all the functions* f *and* g *in* $\mathcal{L}^1(0,L)$ *(take e.g.* $f(x) = g(x) = x^{-\frac{1}{2}}$*), and the maximum norm cannot be a norm over* $\mathcal{L}^2(0,L)$ *(take e.g.* $f(x) = x^{-\frac{1}{4}}$*).*

2.2 Stability of a Numerical Method

We say that a numerical method is stable in terms of a certain norm, if the approximate solution measured in this norm is bounded independently of the corresponding discretisation parameter(s) by a constant multiplied by a suitable measure of the problem data. In mathematical terms, let $\|U_d\|_A$ be a certain norm of the solution U_d obtained by the discretisation of a linear problem, and $\|D\|_B$ be another norm of the problem data D. If we can assert that there exists a constant $C_{AB} > 0$, independent of any particular discretisation level under consideration, such that

$$\|U_d\|_A \leq C_{AB} \|D\|_B,$$

then we say that the discretisation method is stable in the sense of the norm $\|\cdot\|_A$. Notice that in all the cases studied in this book (and also in the case of most numerical methods for solving linear PDEs), the solution U_d is completely defined by a linear combination with a finite number n_D of real coefficients, to be determined by solving a SLAE with a square $n_D \times n_D$ matrix. This kind of system has to be solved just once if the problem is independent of time, like the model problem of Chapter 1. If one is dealing with a time-dependent problem like those to be considered in Chapter 3 onwards, in some cases a SLAE has to be solved at every time step.

Regardless, **stability** in the above sense implies **existence and uniqueness** of the solution U_d (eventually at every time step). This is because a SLAE has a unique solution if and only if the

only possible solution for a null right-side vector is the null vector. Since this implies that $U_d \equiv 0$ as a linear combination of zero coefficients, there can be no other solution U_d' corresponding to the same data set D, for in this case by linearity we would have $\|U_d - U_d'\|_A \leq C_{AB} \|D - D\|_B = 0$, which implies that $U_d = U_d'$. To complete the argument, it suffices to observe that uniqueness implies existence in the case of a SLAE.

It is important to note that, strictly speaking, there is no need for the constant C_{AB} to be independent of discretisation parameters. However, for the purpose of a convergence study based on the underlying stability result, it is usually required that this constant be independent of the discretisation level, characterised by grid or mesh sizes in most cases. Two examples of stability in the above sense are given right away.

2.2.1 Stability in the Maximum Norm

As a first study of stability in norm, let us consider the counterpart of scheme (1.12) in the case of an inhomogeneous Dirichlet boundary condition, namely,

$$\begin{cases} \text{Find } u_i, i = 1, 2, \cdots, n \text{ satisfying} \\ p\frac{2u_i - u_{i+1} - u_{i-1}}{h^2} + qu_i = f(x_i) \\ \text{with } u_0 = a \text{ and } u_{n+1} = u_{n-1}. \end{cases} \tag{2.2}$$

This scheme is also equivalent to a SLAE with the same matrix as in the case of equation (1.12). Its right-side vector in turn is the same as in the homogeneous case, except for the first component, whose value is now $f(x_1) + pa/h^2$.

For example, if we use the maximum norm of $\mathfrak{R}^n$ for both the numerical solution and $\vec{f}_h := [f(x_1), f(x_2), \cdots, f(x_n)]^T$ assuming that f is bounded in $[0, L]$, we shall prove that the scheme is stable in terms of this norm. We use $\vec{f}_h$ here instead of $\vec{b}_h$ defined in Section 1.2 for convenience. Actually, the components of both vectors coincide, except $b_n (= f(x_n)/2)$. More precisely, we will establish the validity of the

Stability inequality for the Three-point FD scheme

$$\boxed{\begin{array}{l}\|\vec{u}_h\|_\infty \leq |a| + CF \\ \text{where } F = \|\vec{f}_h\|_\infty \text{ and } C = L^2/(2p).\end{array}} \tag{2.3}$$

We recall that $\vec{u}_h$ is the vector of approximate values of u at the grid points x_i for $1 \leq i \leq n$. Hence, this means in particular that the numerical solution cannot grow without control as we compute with finer and finer grids, since the norm of $\vec{f}_h$ depends only on f, and hence not on h.

The stability of scheme (2.2) in the above sense is a consequence of the following result:

The discrete maximum principle (DMP): If $f(x_i) \leq 0$ for all i, then $\max_{1\leq i\leq n} u_i \leq a$ if $q = 0$ and $\max_{1\leq i\leq n} u_i \leq \max[0, a]$ if $q > 0$.

We can verify that the DMP holds by an argument quite similar to the one employed in Chapter 1 in order to establish that equation (1.12) has a unique solution: Let us first treat the case $q > 0$. Setting $\nu = (qh^2 + 2p)/(2p)$, by assumption we have $\nu > 1$. Let M be such that $u_M = \max_{1\leq i\leq n} u_i$. If $u_M \leq 0$, then the maximum is non-positive and half a way is done. Let us then assume that $u_M > 0$. If $M = n$, then from the nth equation of (2.2) we infer that $\nu u_n \leq u_{n-1}$. Thus, the only possibility is $u_M = u_n = 0$, which is now discarded. Hence, $M \neq n$ and if $M > 0$, equation (2.2) implies that $\nu u_M \leq (u_{M+1} + u_{M-1})/2$. In this case, at least one out of u_{M-1} and u_{M+1} must be strictly greater than u_M, which is impossible by assumption. Hence, if $u_M > 0$, M must be zero, and thus $u_M = a$. As a result, the remaining half way is also done.

The case where $q = 0$ can be treated in a similar manner and is left as Exercise 2.2.

Observing that $\min_{1\leq i\leq n} u_i = -\max_{1\leq i\leq n}[-u_i]$, the validity of the DMP implies that the following **discrete minimum principle** is also true: If $f(x_i) \geq 0$ for all i, then $\min_{1\leq i\leq n} u_i \geq a$ if $q = 0$ and $\min_{1\leq i\leq n} u_i \geq \min[0, a]$ if $q > 0$.

Notice that if the distribution of forces is positive, the bar will continuously stretch away from its left end. However, its minimum (negative) elongation will take place precisely there if $a < 0$, as both physics and mathematics indicate. On the other hand, if only a displacement $a > 0$ is applied at the section $x = 0$, physics also tells us that the elongation process will be alleviated along the bar. However, since it is subject to a tension $f \geq 0$, the elongation can nowhere be negative. The case $a = 0$ is even more eloquent, for whenever f is positive, the physical solution

certainly corresponds to increasing positive section displacement as one approaches the bar's right end. This means again that the minimum value of the displacement is 0, precisely at $x = 0$. The discrete minimum principle tells us that the numerical scheme reproduces these properties at the discrete level. Otherwise stated, in practical terms, the DMP (and consequently the stability), means that somehow the numerical solution is able to mimic a physically admissible solution. In mathematical terms, this property is illustrated in Figure 2.2, where the aspect of the displacement function u in the case of a positive f is shown for different values of a both positive and negative and for $a = 0$. The graph of u takes the approximate shape of a parabola, with a horizontal tangent at $x = L$ and a minimum value a guaranteed at $x = 0$, if $a \leq 0$. On the other hand, a strictly positive a is not necessarily the minimum value of u. However, according to the minimum principle, the curve representing u cannot cut the x axis.

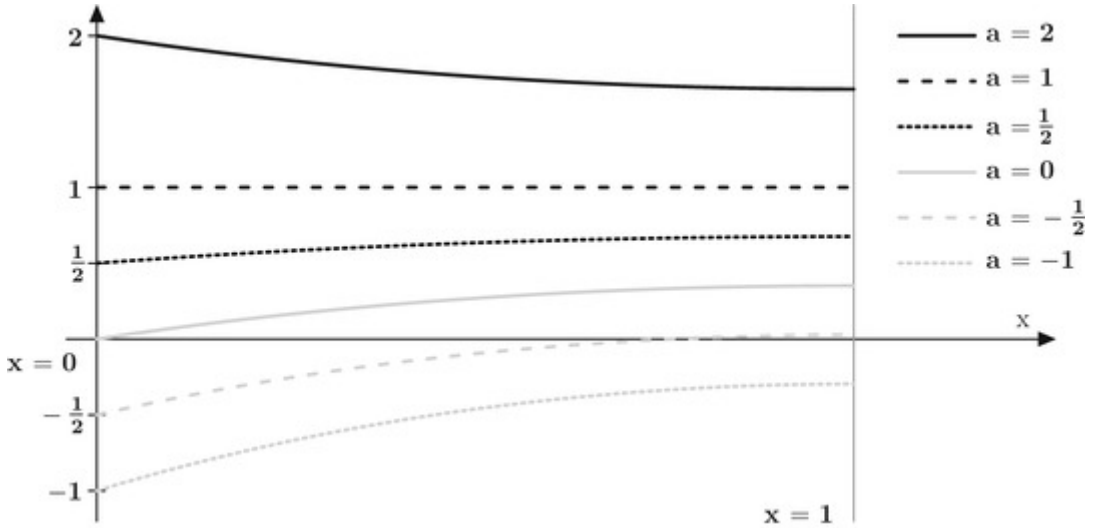


Figure 2.2 An illustration of the solution u for $f \geq 0$ and different values of $a = u(0)$.

Now, in order to apply the DMP to the stability study, let us assume that $\phi(x) = [(x - L)/L]^2/(2p)$ is the solution of equation (p1). Notice that this corresponds to a right-hand side f of the differential equation equal to $-[p\phi']' + q\phi$ supplemented with the inhomogeneous Dirichlet boundary condition $\phi(0) = 1/(2p)$, which happens to be the maximum value of ϕ in $[0, L]$. Applying the operations on the left side of scheme (2.2) to the exact values of ϕ and taking into account that $\phi[(n + 1)h] = \phi[(n - 1)h]$, we obtain after straightforward calculations:

$$\frac{p[2\phi(ih)-\phi(ih-h)-\phi(ih+h)]}{h^2}+q\phi(ih)=-\frac{1}{L^2}+q\phi(ih) \tag{2.4}$$

Let us define the n-component vectors $\vec{w}^+$ and $\vec{w}^-$ by $w_i^+=+u_i+FL^2[\phi(ih)-1/(2p)]$ and $w_i^-=-u_i+FL^2[\phi(ih)-1/(2p)]$, $i=1,2,,\cdots,n$. Then, owing to equations (2.4) and (2.2), we have

$$\begin{cases}\dfrac{p[2w_i^+-w_{i-1}^+-w_{i+1}^+]}{h^2}+qw_i^+=f(ih)-F+qFL^2\left[\phi(ih)-\dfrac{1}{2p}\right]\le 0,\\ \dfrac{p[2w_i^--w_{i-1}^--w_{i+1}^-]}{h^2}+qw_i^-=-f(ih)-F+qFL^2\left[\phi(ih)-\dfrac{1}{2p}\right]\le 0.\end{cases} \tag{2.5}$$

Hence, by virtue of the DMP, $w_i^+\le\max\{0,a+FL^2[\phi(0)-1/(2p)]\}=\max[0,a]\le|a|$ and $w_i^-\le\max\{0,-a+FL^2[\phi(0)-1/(2p)]\}=\max[0,-a]\le|a|$ for every i. However, $\pm u_i=w_i^\pm+FL^2[1/(2p)-\phi(ih)]\le w_i^\pm+FL^2/(2p)\le|a|+FL^2/(2p)$, since $\phi(x)\ge 0$ for every x and $w_i^\pm\le|a|$. It immediately follows that we have stability in the sense of equation (2.3) with $C=L^2/(2p)$.

The DMP can be used in an analogous manner to study the stability in the sense of equation (2.3) of the variant equation (1.15) for a non-uniform grid, and also to several other schemes for constant p and q or not. This comment also extends to all those schemes duly adapted to the case of inhomogeneous boundary conditions, such as the equation (1.25) and (1.28) FV schemes and the linear FE scheme. However, for at least one reason, the bound (2.3) may be unsatisfactory. This happens, for instance, if p is very small as compared to other data and more especially to q. That is why we conclude this subsection by giving another stability result of a wider scope, as long as $q>0$, since it applies to all those schemes for different types of boundary conditions and nonzero functions q and p. In particular, it will be very useful in Chapter 3. Denoting the vector $[u_0,u_1,u_2,\cdots,u_{n-1},u_n]^T\in\mathfrak{R}^{n+1}$ by $\vec{u}$, let q_i, p_i, p_i^+ and p_i^- be given non-negative coefficients for $i=0,\cdots,n$, satisfying $q_i\ge\beta>0$, for every i, together with $2p_i=p_i^++p_i^-$ for $0\le i\le n$ and $p_0^-=p_n^+=0$. Assume that for every $\vec{g}=[g_0,g_1,\cdots,g_n]^T$ and $\vec{f}=[f_0,f_1,\cdots,f_n]^T$ in $\mathfrak{R}^{n+1}$, we have

$$(q_i+2p_i)u_i-p_i^+u_{i+1}-p_i^-u_{i-1}=q_ig_i+f_i \text{ for } i=0,1,2,\cdots,n, \tag{2.6}$$

where u_{-1} and u_{n+1} are meaningless. Then, $\vec{u}$ is bounded as follows:

$$\|\vec{u}\|_\infty \leq \|\vec{g}\|_\infty + \|\vec{f}\|_\infty/\beta. \tag{2.7}$$

Indeed, since $q_i > 0$, we can rewrite equation (2.6) as

$$(1 + 2p_i/q_i)u_i = (p_i^+ u_{i+1} + p_i^- u_{i-1})/q_i + g_i + f_i/q_i \text{ for } i = 0, 1, 2, \cdots, n.$$

Using three times the triangle inequality for absolute values, we obtain

$$(1 + 2p_i/q_i)|u_i| \leq (p_i^+ |u_{i+1}| + p_i^- |u_{i-1}|)/q_i + |g_i| + |f_i|/q_i \text{ for } i = 0, 1, 2, \cdots, n.$$

Therefore, we trivially have

$$(1 + 2p_i/q_i)|u_i| \leq \|\vec{u}\|_\infty (p_{i+1}^+ + p_{i-1}^-)/q_i + \|\vec{g}\|_\infty + \|\vec{f}\|_\infty/\beta \text{ for } i = 0, 1, 2, \cdots, n.$$

Hence, using the properties of the coefficients p_i^- and p_i^+, we derive

$$(1 + 2p_i/q_i)|u_i| \leq 2p_i/q_i \, \|\vec{u}\|_\infty + \|\vec{g}\|_\infty + \|\vec{f}\|_\infty/\beta \text{ for } i = 0, 1, 2, \cdots, n.$$

Letting u_M be a component of $\vec{u}$ such that $|u_M| = \|\vec{u}\|_\infty$, we have

$$(1 + 2p_M/q_M) \, \|\vec{u}\|_\infty \leq 2p_M/q_M \, \|\vec{u}\|_\infty + \|\vec{g}\|_\infty + \|\vec{f}\|_\infty/\beta \text{ for } i = 0, 1, 2, \cdots, n.$$

which yields equation (2.7).

Notice that this result applies in particular to equation (2.2) if $q > 0$. Indeed, in this case, $p_i^+ = p_i^- = p/h^2$, $p_i = p/h^2$ for $i = 1, 2, \cdots, n-1$, $p_0 = p_0^+ = 0$, $p_n^- = 2p_n = 2p/h^2$, $q_0 = 1$, $g_0 = a$, $f_0 = 0$ and $g_i = 0$, $q_i = q$, $f_i = f(x_i)$ for $i = 1, 2, \cdots, n$. Hence, we have established the following:

Stability inequality for the Three-point FD scheme with $q > 0$

$$\boxed{\max_{1 \leq i \leq n} |u_i| \leq |a| + \max_{1 \leq i \leq n} |f(x_i)|/\min[q, 1],} \tag{2.8}$$

Equation (2.8) is fairly equivalent to equation (2.3), except if $p << \min[q, 1]$ or $\min[q, 1] << p$.

2.2.2 Stability in the Mean-square Sense

Now, we switch to an example of stability in terms of norms of the mean square type, such as L^2 norms. Since this is more natural in the context of variational forms, we will only carry out this

kind of study for the FE scheme (equation (1.22)). However, for the sake of generality, we will consider its modification in order to incorporate inhomogeneous boundary conditions at both ends of $(0, L)$, with arbitrary p and q satisfying our initial assumptions. This requires a working space consisting of functions which are sums of functions in V with constants. This space is known in the literature as the **Sobolev space** $H^1(0, L)$ [1].

First, we write down the problem to solve followed by the corresponding FE approximate problem, namely,

$$\begin{cases} \text{Find } u \in H^1(0, L) \text{ such that } u(0) = a \text{ and} \\ \int_0^L (pu'v' + quv)dx = \int_0^L fvdx + bv(L) \ \forall v \in V \end{cases} \tag{2.9}$$

$$\begin{cases} \text{Find } u_h \in W_h \text{ such that } u_h(0) = a \text{ and} \\ \int_0^L (pu_h'v' + qu_hv)dx = \int_0^L fvdx + bv(L) \ \forall v \in V_h. \end{cases} \tag{2.10}$$

Let us take $v = u_h - a$. Since $v \in V_h$ because $v(0) = 0$ by construction, from equation (2.10) we obtain

$$\int_0^L [(pu_h')^2 + qu_h^2]dx = \int_0^L [aqu_h + f(u_h - a)]dx + b[u_h(L) - a]. \tag{2.11}$$

Recalling the definition of the L^2-norm and the bounds of p and q, we derive

$$\alpha \int_0^L [u_h']^2 dx \leq \int_0^L [B|a||u_h| + |f(u_h - a)|]dx + |b|(|u_h(L)| + |a|). \tag{2.12}$$

On the other hand, for every function in $v \in V$ and $\forall y \in (0, L]$, we have

$$v^2(y) = \int_0^y [v^2]'(x)dx.$$

Therefore, $\forall y \in (0, L]$

$$v^2(y) = 2\int_0^y [vv'](x)dx.$$

Then, by virtue of the Cauchy–Schwarz inequality, this implies that

$$v^2(y) \leq 2\left\{\int_0^y v(x)^2 dx\right\}^{1/2} \left\{\int_0^y [v'(x)]^2 dx\right\}^{1/2} \ \forall y \in [0, L]. \tag{2.13}$$

or

$$v^2(y) \leq 2 \parallel v\parallel_{0,2} \parallel v^{'}\parallel_{0,2} \forall y \in [0, L]. \tag{2.14}$$

Two useful results follow from equation (2.14), namely,

$$|v(L)| \leq [2 \parallel v\parallel_{0,2} \parallel v^{'}\parallel_{0,2}]^{1/2} \forall v \in V, \tag{2.15}$$
$$\text{and } \int_0^L v^2(y)dy \leq \int_0^L [2 \parallel v\parallel_{0,2} \parallel v^{'}\parallel_{0,2}]dy \forall v \in V,$$

or, equivalently,

$$\parallel v \parallel_{0,2}^2 \leq 2L \parallel v\parallel_{0,2} \parallel v^{'}\parallel_{0,2} \forall v \in V,$$

which finally yields

The Friedrichs–Poincaré inequality

$$\boxed{\parallel v\parallel_{0,2} \leq 2L \parallel v^{'}\parallel_{0,2} \forall v \in V.} \tag{2.16}$$

Now, we plug into equation (2.12) the inequality $\int_0^L |u_h|dx \leq L^{1/2} \parallel u_h\parallel_{0,2}$, which holds thanks to the Cauchy–Schwarz inequality. It follows that

$$\alpha \parallel u_h^{'} \parallel_{0,2}^2 \leq 2L^{3/2}B|a| \parallel u_h^{'}\parallel_{0,2} + \int_0^L |fu_h|dx + |a| \int_0^L |f|dx + |b|(|u_h(L)| + |a|).$$

Next, we manipulate the right side above by using both equations (2.15) and (2.16). More precisely, the Cauchy–Schwarz inequality, combined with the obvious relation $u_h(L) = \int_0^L u_h^{'}(x)dx + a$, leads successively to

$$\alpha \parallel u_h^{'} \parallel_{0,2}^2 \leq 2L(B|a|L^{1/2} + \parallel f\parallel_{0,2}) \parallel u_h^{'}\parallel_{0,2} + L^{1/2}|a| \parallel f\parallel_{0,2} + |b|(|u_h(L)| + |a|),$$
$$\alpha \parallel u_h^{'} \parallel_{0,2}^2 \leq 2\, L(B|a|\, L^{1/2} + \parallel f\parallel_{0,2}\,) \parallel u_h^{'}\parallel_{0,2} + L^{1/2}|a| \parallel f\parallel_{0,2} + |b|$$
$$(2L^{1/2} \parallel u_h^{'}\parallel_{0,2} + |a|),$$

or, in more compact form:

$$\begin{cases} \alpha \parallel u_h^{'} \parallel_{0,2}^2 \leq D \parallel u_h^{'}\parallel_{0,2} + |a|(L^{1/2} \parallel f\parallel_{0,2} + |b|) \\ \text{where } D = 2L(B|a|L^{1/2} + \parallel f\parallel_{0,2}) + 2|b|L^{1/2}. \end{cases} \tag{2.17}$$

Now we use **Young's inequality**,[1]

$$|sr| \leq (s^2\delta + r^2/\delta)/2 \; \forall s, r \in \Re \text{ and } \forall \delta > 0.$$

Then, taking $s = \| u_h' \|_{0,2}$, $r = D$ and $\delta = \alpha$, we readily derive from equation (2.17):

$$\alpha \| u_h' \|_{0,2}^2 \leq D^2/\alpha + 2|a|(L^{1/2} \| f \|_{0,2} + |b|). \tag{2.18}$$

After straightforward calculations, equation (2.18) yields

A first mean-square stability inequality for the $\mathcal{P}_1$ FEM

$$\boxed{\| u_h' \|_{0,2} \leq C[|a| + \| f \|_{0,2} + |b|],} \tag{2.19}$$

for a suitable constant C depending only on α, B and L (determination of a fine expression for C from equations (2.17) and (2.18) is proposed to the reader as Exercise 2.3). Now, it is clear that a stability result similar to equation (2.19) also holds for the L^2-norm of u_h, owing to the Friedrichs–Poincaré inequality, that is,

A second mean-square stability inequality for the $\mathcal{P}_1$ FEM

$$\boxed{\| u_h \|_{0,2} \leq C'[|a| + \| f \|_{0,2} + |b|].} \tag{2.20}$$

Indeed, equation (2.16) implies that $\| u_h - a \|_{0,2} \leq 2L \| (u_h - a)' \|_{0,2} = 2L \| u_h' \|_{0,2}$. On the other hand, using property (iii) of a norm, we easily obtain $\| u_h - a \|_{0,2} \geq \| u_h \|_{0,2} - |a| L^{1/2}$. Therefore, $\| u_h \|_{0,2} \leq 2L \| u_h' \|_{0,2} + L^{1/2}|a|$. Then, combining this with equation (2.19), we obtain equation (2.20) with $C' = 2LC + L^{1/2}$. Under an additional condition on q specified below, combining equations (2.19) and (2.20), one can easily establish that a similar result also holds for

The energy norm of the bar

$$\boxed{\| v \|_e := \left\{ \tfrac{1}{2} \int_0^L [p(v')^2 + qv^2](x) dx \right\}^{1/2} \; \forall v \in V.} \tag{2.21}$$

The fact that $\| \cdot \|_e$ is effectively a norm on space V is an immediate consequence of the properties of p and q. Notice that the energy norm is only a semi-norm on $H^1(0, L)$, and in the case of an inhomogeneous boundary condition $u_h(0) = a$. Thus, $u_h \in H^1(0, L)$, but $u_h \notin V$. Nevertheless, if there exists $\beta > 0$ such that

$$q(x) \geq \beta \, \forall x \in [0, L]$$

the energy norm is a norm on $H^1(0, L)$, and also on V even in case not. This is due to the fact that this norm is associated with the **energy inner product** given by

$$(u|v)_e := \frac{1}{2} \int_0^L [pu'v' + quv](x)dx. \tag{2.22}$$

Actually, the energy norm carries this name because it expresses the **total internal energy** of the bar given by $I_e(v) := (v|v)_e$ for a given state of longitudinal displacement v. The term $\frac{1}{2} \int_0^L [(pv')(x)]^2 dx/2$ of I_e accounts for the elongational stiffness of the bar, and the term $\frac{1}{2} \int_0^L [(qv)(x)]^2 dx/2$ stands for the stored energy due to a spring effect on the bar subject to a longitudinal displacement v.

Summarising what we saw in this subsection, for a suitable mesh independent constant C_e, we have

A stability inequality for the $\mathcal{P}_1$ FEM in the energy norm

$$\boxed{\| u_h \|_e \leq C_e[|a| + \| f \|_{0,2} + |b|].} \tag{2.23}$$

It is interesting to note that all the steps leading to equation (2.23) apply as well to the exact solution u, and thus we also have

$$\| u \|_e \leq C_e[|a| + \| f \|_{0,2} + |b|]. \tag{2.24}$$

The stability result (equation (2.23)) states that, similarly to the longitudinal displacement u, the total internal energy of the bar for its FE counterpart u_h remains controlled by the applied loads represented by a, b and f, whatever the mesh. In short, here again the stability of the discrete model means that somehow it mimics the physical behavior of the exact model.

In complement to the above study, the reader may derive as Exercise 2.4 the precise expression of C_e resulting from a suitable combination of equations (2.19) and (2.20).

Remark 2.4

The square of the energy norm minus the ***work of external forces*** $\int_0^L fv\,dx$ *is the* ***total potential energy*** *of the bar subject to a density of forces* f*, for a given* ***admissible longitudinal displacement*** v*. The term 'admissible' means that, besides satisfying the kinematic condition* $v(0) = 0$*, the total potential energy in terms of* v *is finite. The latter condition requires that* $v \in H^1(0, L)$*, which justifies the choice of space* V *in the variational formulation* (P_3)*, from the point of view of classical continuum mechanics. Actually, the solution* u *is the displacement function that minimises the total potential energy (see e.g. [131]). Incidentally, we observe that the boundary condition* $u'(L)$ *is satisfied by the minimising displacement* u*, but not necessarily by an arbitrary admissible displacement* v*. In this sense, the different nature of boundary conditions* $u(0) = 0$ *and* $u'(L) = 0$ *is not merely mathematical. In solid mechanics, they are called* ***essential boundary conditions*** *and* ***natural boundary conditions****. These expressions characterise that the former are prescribed to any admissible displacement, while the latter is a balance of force condition to be satisfied by the minimising displacement.*

To conclude, we observe that the stability of both the Vertex-centred and the Cell-centred FVM can be studied following the same principles as those applying to the FDM. It suffices to start from the corresponding schemes, and apply the maximum principle that holds for equations (1.24) or (1.27) in both cases. Checking this assertion is left to the reader as Exercise 2.5.

2.3 Scheme Consistency

In this section, we will be concerned about determining the order of the so-called **local truncation error** inherent to a given numerical method in terms of relevant discretisation parameters, such as grid or mesh sizes. By definition, this error is the difference between the left and the right side of the numerical scheme if the approximate values are replaced by the corresponding exact values. This difference is also commonly called the **scheme's residual**. In case the residuals tend to zero as the discretisation parameters go to zero, we say that the method is **consistent**. The **order of consistency** of the method will result from a global evaluation of the local truncation errors as seen below. Let us consider two examples.

2.3.1 Consistency of the Three-point FD Scheme

For simplicity, we study only the case of a constant p. Moreover, for convenience, we use $n \times n$ matrix B_h instead of A_h defined in Section 1.2, both matrices being identical except for the last row of B_h, which equals the last row of A_h multiplied by two. Recalling the presentation of the FDM in Section 1.2, let us replace u_i by $u(x_i)$ in the FD scheme (1.12) for $i = 1, 2, \cdots, n$. Denoting by $\vec{\bar{u}}_h$ the vector $[u(x_1), u(x_2), \cdots, u(x_n)]^T$ of $\Re^n$, we define the **residual vector** $\vec{r}_h(u) = [r_{h,1}(u), r_{h,2}(u), \cdots, r_{h,n}(u)]^T$ by

$$\vec{r}_h(u) = B_h \vec{\bar{u}}_h - \vec{f}_h.$$

Recalling equations (1.7) and (1.9), we easily infer that for $i < n$,

$$r_{h,i}(u) = p\frac{2u(x_i) - u(x_{i-1}) - u(x_{i+1})}{h^2} + qu(x_i) - f(x_i) = -pR_i(u). \tag{2.25}$$

On the other hand, since $u^{'}(L) = 0$, using a standard Taylor expansion we have

$$2u(x_{n-1}) = 2u(x_n) + h^2 u^{''}(x_n) - h^3 u^{'''}(x_n)/3 + h^4 u^{(iv)}(\xi_n)/12$$

where ξ_n is a suitable abscissa in the interval $[x_{n-1}, x_n]$. Therefore, we have

$$r_{h,n}(u) = 2p\frac{u(x_n) - u(x_{n-1})}{h^2} + qu(x_n) - f(x_n) = -pR_n(u), \tag{2.26}$$

where $R_n(u) = hu^{'''}(x_n)/3 - h^2 u^{(iv)}(\xi_n)/12$. Hence, provided the derivatives of u up to the fourth order are continuous in $[0, L]$, the **local truncation errors** for the FDM with a uniform grid can be bounded as follows:

$$\begin{aligned} |r_{h,i}(u)| &\le p \parallel u^{(iv)} \parallel_{0,\infty} h^2/12 \ \forall i < n \\ |r_{h,n}(u)| &\le p(\parallel u^{'''} \parallel_{0,\infty} h/3 + \parallel u^{(iv)} \parallel_{0,\infty} h^2/12). \end{aligned} \tag{2.27}$$

Then, by definition, we say that the local truncation error for the FDM with an equally spaced grid is of order two at the grid points x_i for $i < n$, and of order one at the grid point $x_n = L$. As we will see in Section 2.3.2, the truncation errors are directly linked to the accuracy of the discretisation method. Therefore, the above conclusion is not satisfactory, since for a small h, the error at grid point x_n might pollute the accuracy elsewhere. In other words, in spite of the fact that all the local truncation errors but one are of order two, the order of consistency of the scheme will be one, since the order of the local truncation error at x_n is one, that is, the dominant order.

That is why, for explanatory purposes, it is advisable to modify the definition of the fictitious unknown u_{n+1} in the following fashion.

First of all, we observe that f must be continuously differentiable in $[0, L]$, otherwise, there is no way for u to be more than two times differentiable. Thus, we have $-pu'''(L) + qu'(L) = f'(L)$, and taking into account the boundary condition at $x = L$, we have: $u'''(L) = -f'(L)/p$. Therefore, if we replace u_{n+1} with $u_{n-1} - h^3 f'(L)/3p$, the local truncation error at x_n turns to $\underline{r}_{h,n}(u)$ given by

$$\begin{cases} \underline{r}_{h,n}(u) = 2p\dfrac{u(x_n) - u(x_{n-1})}{h^2} + qu(x_n) - f(x_n) + hf'(x_n)/3 = -p\underline{R}_n(u) \\ \text{where } \underline{R}_n(u) := -h^2 u^{(iv)}(\xi_n)/12, \end{cases} \tag{2.28}$$

as one can easily check. The above correction in the last equation gives rise to another Three-point FD scheme, for which a better consistency result holds, namely,

A modified Three-point FD scheme

$$\boxed{\begin{array}{l} p\dfrac{2\underline{u}_i - \underline{u}_{i-1} - \underline{u}_{i+1}}{h^2} + q\underline{u}_i = \underline{f}_i \text{ for } i = 1, 2, \cdots, n, \text{ with } \underline{u}_0 = a \text{ and } \underline{u}_{n+1} = \underline{u}_{n-1}, \\ \underline{f}_i = f(x_i) \text{ for } i = 1, 2, \cdots, n-1 \text{ and } \underline{f}_n = f(x_n) - hf'(x_n)/3. \end{array}} \tag{2.29}$$

Let us denote by $\vec{\underline{u}}_h$ the corresponding solution vector, $\vec{\underline{u}}_h = [\underline{u}_1, \underline{u}_2, \cdots, \underline{u}_n]^T$, which is the solution of the SLAE

$$B_h \vec{\underline{u}}_h = \vec{\underline{f}}_h \tag{2.30}$$

where $\vec{\underline{f}}_h = [\underline{f}_1, \underline{f}_2, \cdots, \underline{f}_n]^T$. Accordingly, we denote the underlying residual vector by $\vec{\underline{r}}_h(u) = [\underline{r}_{h,1}(u), \underline{r}_{h,2}(u), \cdots, \underline{r}_{h,n}(u)]^T$, satisfying $\vec{\underline{r}}_h(u) := B_h \vec{\underline{u}}_h - \vec{\underline{f}}_h$. Noticing that $\underline{r}_{h,i}(u) := r_{h,i}(u)$ for $i = 1, 2, \cdots, n-1$, combining equations (2.27) and (2.28), we establish the

Consistency of the modified Three-point FD scheme 2.29

$$\boxed{|\underline{r}_{h,i}(u)| \leq p \parallel u^{(iv)} \parallel_{0,\infty} h^2/12 \text{ for } 1 \leq i \leq n.} \tag{2.31}$$

The other way around, we can assert that, at least for constant p and q and $b = 0$, the Three-point FD scheme (1.12) with an equally spaced grid is **second-order consistent**, provided we set

$u_{n+1} = u_{n-1} - h^3 f'(L)/3p$. Moreover, taking this modification into account, the reader can easily verify that a stability inequality of the same type as equation (2.3) holds for scheme (2.29), provided F is replaced by $\underline{F} := \max_{1 \le i \le n} |\underline{f}_i|$.

2.3.2 Consistency of the $\mathcal{P}_1$ FE Scheme

As a model, we consider the case of homogeneous Neumann boundary conditions at $x = L$ and inhomogeneous Dirichlet boundary conditions at $x = 0$. In the study of the FEM, it is more natural to treat consistency and truncation errors in a variational framework. This means that now we plug the exact values of u at the nodal points on the left side of the approximate variational problem (1.22), and then subtract from the result its right side. Let us denote by V_h^+ the subset of W_h consisting of functions whose value at $x = 0$ equals a. We are actually replacing u_h with a function $\tilde{u}_h \in V_h^+$, whose nodal value at x_i for $i > 0$ is $u(x_i)$. Using the inner product notation for the integrals in equation (1.22), it is convenient to define two **residual functions** $r_h(u)$ and $s_h(u)$ such that

$$(r_h(u)|v)_0 + (s_h(u)|v')_0 = (p\tilde{u}_h'|v')_0 + (q\tilde{u}_h|v)_0 - (f|v)_0 \text{ for any } v \in V_h.$$

Recalling that the above right side vanishes $\forall v \in V_h$, if we replace $\tilde{u}_h$ by the exact solution u, we obtain

$$(r_h(u)|v)_0 + (s_h(u)|v')_0 = (p[\tilde{u}_h - u]'|v')_0 + (q[\tilde{u}_h - u]|v)_0 \text{ for any } v \in V_h. \qquad (2.32)$$

In order to appropriately express the functions $[\tilde{u}_h - u]'$ and $\tilde{u}_h - u$, we resort to the elementary interpolation theory. Referring to Figure 2.3, we first notice that in each element $T_i = [x_{i-1}, x_i]$, $\tilde{u}_h$ is the linear interpolating function of u at x_{i-1} and x_i, whose expression is

$$\tilde{u}_h(x) = \frac{(x - x_{i-1})u(x_i) + (x_i - x)u(x_{i-1})}{x_i - x_{i-1}}.$$

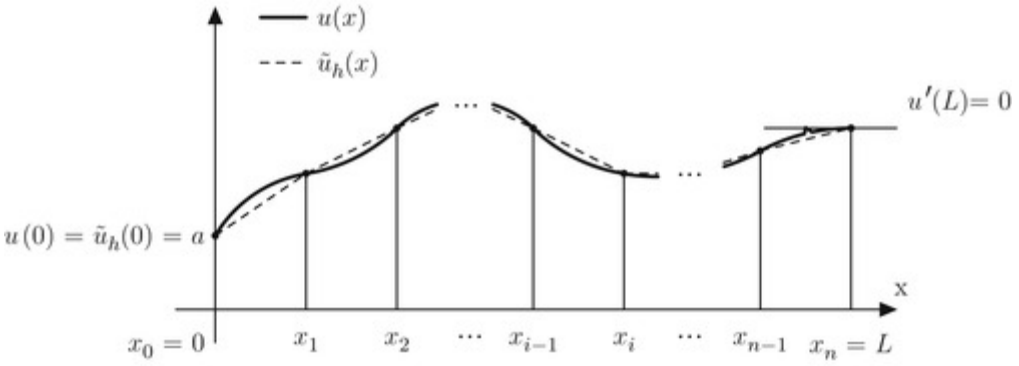


Figure 2.3 The function u and its interpolate $\tilde{u}_h$ in the FE space V_h^+.

Let us assume that u'' belongs to $L^2(0, L)$, or equivalently that $u \in H^2(0, L)$, using a **Sobolev space** notation (cf. [1]). Since this implies that both u'' and $\varphi u''$ are integrable functions in $(0, L)$ for every linear function φ, the following identities hold (please check!):

$$u(x) = u(x_i) - (x_i - x)u'(x) - \int_x^{x_i} (x_i - s)u''(s)ds \text{ and}$$

$$u(x) = u(x_{i-1}) + (x - x_{i-1})u'(x) - \int_{x_{i-1}}^{x} (s - x_{i-1})u''(s)ds.$$

Recalling that $h_i = x_i - x_{i-1}$, suitable combinations of both identities allow us to derive quite easily

$$[\tilde{u}_h - u](x) = \frac{(x - x_{i-1}) \int_x^{x_i} (x_i - s)u''(s)ds + (x_i - x) \int_{x_{i-1}}^{x} (s - x_{i-1})u''(s)ds}{h_i}, \quad (2.33)$$

$$\text{and } [\tilde{u}_h - u]'(x) = \frac{\int_x^{x_i} (x_i - s)u''(s)ds - \int_{x_{i-1}}^{x} (s - x_{i-1})u''(s)ds}{h_i}. \quad (2.34)$$

Plugging equations (2.33) and (2.34) into (2.32), we readily obtain for $x \in (x_{i-1}, x_i)$:

$$\begin{cases} [r_h(u)](x) = \dfrac{q(x)}{h_i}\left[(x-x_{i-1})\displaystyle\int_x^{x_i}(x_i-s)u''(s)ds \right. \\ \qquad\qquad \left. +(x_i-x)\displaystyle\int_{x_{i-1}}^{x}(s-x_{i-1})u''(s)ds\right] \\ \text{and} \\ [s_h(u)](x) = \dfrac{p(x)}{h_i}\left[\displaystyle\int_x^{x_i}(x_i-s)u''(s)ds - \int_{x_{i-1}}^{x}(s-x_{i-1})u''(s)ds\right]. \end{cases} \tag{2.35}$$

Rather fastidious though straightforward calculations reported below lead to the following estimates for the L^2-norm of both residual functions, thereby characterising the

Consistency of the $\mathcal{P}_1$ FE scheme

$$\boxed{\begin{array}{l} \| s_h(u)\|_{0,2} \leq C_s h \| u''\|_{0,2} \\ \| r_h(u)\|_{0,2} \leq C_r h^2 \| u''\|_{0,2}. \end{array}} \tag{2.36}$$

C_s is a mesh-independent constant proportional to A. C_r in turn is estimated below as proportional to B, following basically the same arguments that apply to C_s.

Let us prove equation (2.36) from (2.35). First, we consider the case of $s_h(u)$. We have

$$\| s_h(u) \|_{0,2}^2 \leq \left[\frac{A}{ch}\right]^2 \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left[\int_x^{x_i}(x_i-s)u''(s)ds - \int_{x_{i-1}}^{x}(s-x_{i-1})u''(s)ds\right]^2 dx.$$

c being the **uniformity constant** of the family of meshes in use (notice that $\min\limits_{1\leq i\leq n} h_i / \max\limits_{1\leq i\leq n} h_i \geq c$ for every n).

Since $(s\pm r)^2 \leq 2(s^2+r^2)$ and both (x,x_i) and (x_{i-1},x) are subsets of (x_{i-1},x_i), this yields

$$\| s_h(u) \|_{0,2}^2 \leq \frac{2A^2}{(ch)^2} \times$$
$$\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left\{\left[\int_{x_{i-1}}^{x_i}(x_i-s)|u''(s)|ds\right]^2 + \left[\int_{x_{i-1}}^{x_i}(s-x_{i-1})|u''(s)|ds\right]^2\right\}dx.$$

or, noting that $\forall s\in(x_{i-1},x_i)$, $|s-x_{i-1}|\leq h$ and $|x_i-s|\leq h$,

$$\| s_h(u) \|_{0,2}^2 \leq \frac{4A^2}{c^2}\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left[\int_{x_{i-1}}^{x_i}|u''(s)|ds\right]^2 dx,$$

Thus,

$$\| s_h(u) \|_{0,2}^2 \leq \frac{4A^2h}{c^2} \sum_{i=1}^{n} \left[\int_{x_{i-1}}^{x_i} |u''(x)| dx \right]^2.$$

On the other hand, by virtue of the Cauchy–Schwarz inequality applied to the L^2-inner product in (x_{i-1}, x_i), for any square integrable function g in this interval, we have $|\int_{x_{i-1}}^{x_i} g(x)dx|^2 \leq \int_{x_{i-1}}^{x_i} dx \int_{x_{i-1}}^{x_i} |g(x)|^2 dx = h_i \int_{x_{i-1}}^{x_i} |g(x)|^2 dx$. Hence, we further obtain

$$\| s_h(u) \|_{0,2}^2 \leq \frac{4A^2h}{c^2} \sum_{i=1}^{n} h_i \int_{x_{i-1}}^{x_i} |u''(x)|^2 dx.$$

This finally leads to

$$\| s_h(u) \|_{0,2} \leq \frac{2Ah}{c} \left[\int_0^L |u''(x)|^2 dx \right]^{1/2} = C_s h \| u'' \|_{0,2}$$

with $C_s = 2A/c$.

The case of $r_h(u)$ can be dealt with in a similar manner. That is why we skip some details which call on the same arguments as above. We have

$$\| r_h(u) \|_{0,2}^2 \leq \left[\frac{B}{ch} \right]^2 \times$$
$$\sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left[(x - x_{i-1}) \int_x^{x_i} (x_i - s) u''(s) ds + (x_i - x) \int_{x_{i-1}}^{x} (s - x_{i-1}) u''(s) ds \right]^2 dx,$$

or

$$\| r_h(u) \|_{0,2}^2 \leq \frac{2B^2}{c^2} \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left\{ \left[\int_x^{x_i} (x_i - s) u''(s) ds \right]^2 + \left[\int_{x_{i-1}}^{x} (s - x_{i-1}) u''(s) ds \right]^2 \right\} dx.$$

This yields

$$\| r_h(u) \|_{0,2}^2 \leq 2 \left[\frac{Bh}{c} \right]^2 \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left\{ \left[\int_x^{x_i} |u''(s)| ds \right]^2 + \left[\int_{x_{i-1}}^{x} |u''(s)| ds \right]^2 \right\} dx,$$

and, further,

$$\| r_h(u) \|_{0,2}^2 \leq h\left[\frac{2Bh}{c}\right]^2 \sum_{i=1}^{n}\left[\int_{x_{i-1}}^{x_i} |u''(s)|ds\right]^2.$$

Then, using the Cauchy–Schwarz inequality applied to integrals in (x_{i-1}, x_i), we obtain

$$\| r_h(u) \|_{0,2}^2 \leq \left[\frac{2Bh^2}{c}\right]^2 \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} |u''(x)|^2 dx,$$

or, equivalently,

$$\| r_h(u)\|_{0,2} \leq C_r h^2 \| u''\|_{0,2},$$

with $C_r = 2B/c$.

Summarising, the estimates in equation (2.36) of the global truncation error suggest that the FEM is a consistent method with order two in terms of the L^2-norm of the error, and of order one in terms of the same norm of the error first-order derivative. However, in a variational framework, these conclusions are not as clear as in the local sense applying to the FDM. That is why we prefer to clarify the concept of order of an FEM in the following section, which deals precisely with error estimates for discretisation methods, besides their convergence.

Incidentally, we note that an elementary variant of equation (2.36) promptly yields estimates for the interpolation error in the norms of $L^2(0, L)$ and the semi-norm of $H^1(0, L)$, namely,

Error estimates for the $\mathcal{P}_1$ FE interpolate $\tilde{v}_h$ of $v \in \mathbf{H}^2(0, L)$

$$\boxed{\begin{array}{l}\| v - \tilde{v}_h\|_{0,2} \leq \tilde{C}_0 h^2 \| v''\|_{0,2} \\ \| [v - \tilde{v}_h]'\|_{0,2} \leq \tilde{C}_1 h \| v''\|_{0,2}.\end{array}} \quad (2.37)$$

where $\tilde{C}_0$ and $\tilde{C}_1$ are mesh-independent constants.

Incidentally, we point out that finer estimates for C_s and C_r independent of c can be obtained. These estimates are left to the reader as Exercise 2.6. Notice that resulting values of $\tilde{C}_1$ will not account for whether or not a quasi-uniform family of meshes is in use. We strongly suggest the reader to determine fine expressions for both constants in Exercise 2.7.

To conclude this section, we note that the order of consistency of the Vertex-centred FVM can be determined in the same way as in the case of the FDM. Indeed, the corresponding scheme can be

viewed as an FD scheme with a non-uniform grid according to equation (1.26). However, the order of consistency of this (FD or FV) scheme is one instead of two, since for a non-uniform grid it is not possible to get rid of the terms involving u''' in the Taylor expansions.

On the other hand, the Cell-centred FVM works differently, for the order of its local truncation errors is zero. Actually, this scheme cannot be labelled as a consistent scheme by means of Taylor expansions, or more vulgarly in the 'FD sense'. However, the fluxes it is based upon are consistent in another sense, and from this fact it is possible to guarantee that the Cell-centred FVM is completely reliable. We refer to Example 1.1 for an illustration of this assertion. We omit a detailed study of this property here for the sake of brevity, preferring to postpone it to Chapters 5 and 7, where it will be carried out in a similar but wider context. However, we encourage the reader to practice the techniques employed in this section, by determining the local truncation errors for both FVMs. Incidentally, this is the purpose of Exercise 2.8.

2.4 Convergence of the Discretisation Methods

A numerical method for solving a certain type of differential equation cannot be viewed as a reliable one, unless it possesses a property called **convergence**. In simple terms, this means that the numerical values provided by the method will tend in some sense to the corresponding values of the exact solution, as the method's discretisation parameters approach zero.

Assume that we can measure in a certain norm $\|\cdot\|$ both the relevant values of the approximate solution generically denoted by $U_{\mathcal{H}}$ produced by a discretisation method, and the corresponding values of the exact solution denoted by U, where $\mathcal{H}$ represents the set of discretisation parameters. The method is said to converge if $\|U_{\mathcal{H}} - U\|$ tends to zero as all the parameters in $\mathcal{H}$ go to zero. Moreover, we say that the **order of convergence** of the discretisation method is $\mu_l > 0$ in terms of the lth discretisation parameter, say H_l, if the above value goes to zero as fast as $H_l^{\mu_l}$ does, making the utopian assumption that all the other parameters have already tended to zero. Of course, in case there is only one discretisation parameter, or in case all the discretisation parameters are linked to each other by certain relations, the concept of order of convergence can be mastered more realistically, since they are supposed to tend to zero all together in a prescribed manner. An expression bounding $\|U - U_{\mathcal{H}}\|$, possibly implying the method's convergence, is called an **error estimate**.

Equivalently, we say that the approximate solution $U_{\mathcal{H}}$ converges to U as all the parameters in $\mathcal{H}$ go to zero if the above condition is fulfilled. For example, in the FDM, $\mathcal{H}$ is reduced to the grid size h, and $U_{\mathcal{H}}$ is the vector of grid values $\vec{u}_h$. Then, naturally enough, U is the vector $\vec{\tilde{u}}_h = [u(x_1), u(x_2), \cdots, u(x_n)]^T$, and we may check the method's convergence using any norm of $\Re^n$, in particular the maximum norm. Actually, this is the way we will study the convergence of the FDM hereafter.

Now pushing further the maxim quoted in the introduction of this chapter [123] regarding convergence of a numerical method, in this section we give examples that significantly illustrate it. More precisely, we will establish convergence in certain norms by combining the method's stability in the sense of these norms with a strictly positive order of consistency.

2.4.1 Convergence of the Three-point FDM

Here, we endeavour to show that the FDM with an equally spaced grid to solve the modification of (P_1) accommodating the boundary condition $u(0) = a$ converges in the maximum norm of $\Re^n$ with order two, as long as the solution u has some suitable regularity properties. More specifically, we want to establish that $\| \vec{\tilde{u}}_h - \vec{u}_h \|_\infty$ goes to zero like a term of the form Ch^2 as h goes to zero, for a suitable constant C depending on u. This means in particular that the approximate values at the grid points will be closer and closer to the corresponding values of the exact solution as h goes to zero, or as n goes to infinity. This result applies to the modification (2.29) of equation (1.12), for the order of consistency is only one in the case of the latter scheme. Let us define the vector $\vec{\bar{u}}_h := \vec{u}_h - \vec{\tilde{u}}_h$. Noticing that $u_0 = \tilde{u}_0 = a$, we have $\bar{u}_0 = 0$. Moreover, since $B_h \vec{u}_h - \vec{f}_h = \vec{0}$, and from the consistency analysis in the previous section $B_h \vec{\tilde{u}}_h - \vec{f}_h = \vec{r}_h(u)$, we have $B_h \vec{\bar{u}}_h = \vec{r}_h(u)$. This means that the vector $\vec{\bar{u}}_h$ is the solution provided by FD scheme (1.12) with $a = 0$, and the right side is the vector $\vec{r}_h(u)$ instead of $\vec{f}_h$. But, according to our stability result in equation (2.3), we must have

$$\| \vec{\bar{u}}_h \|_\infty \leq C \| \vec{r}_h(u) \|_\infty.$$

Thus recalling equation (2.31), provided the fourth derivative of u is bounded in $[0, L]$, we immediately establish the

Error estimate for the modified Three-point FD scheme (equation 2.29)

$$\| \vec{\tilde{u}}_h - \vec{\underline{u}}_h \|_\infty \leq Cph^2 \| u^{(iv)} \|_{0,\infty}/12. \tag{2.38}$$

This means that the modified FD scheme (2.29) is **second-order convergent** in the maximum norm of the grid point values. Logically enough, by the same arguments we would conclude that scheme (equation (1.12)) is first-order convergent in the same norm. However, while on the one hand **the order of consistency may be sufficient for the same order of convergence to hold, on the other hand it is by no means necessary**. Actually (1.12) turns out to be **second-order convergent**, but the proof of this result is more subtle (cf. Exercise 7.1). Similarly, the reader can figure out quite easily that in the case of non-uniform grids, both schemes are first-order consistent, and hence to the least first order convergent in this norm. The wonder is that numerical experiments show second order convergence even in this case, provided *u* is sufficiently smooth.

For a more concrete interpretation of equation (2.38), let us assume that we are using nested grids in such a way that h is divided by two from a given discretisation level to the next. Confining ourselves to values of n greater than two, equation (2.38) implies that $u_{n/2}$ will tend to $u(L/2)$ like an $O(1/n^2)$ term, or an $O(h^2)$ term, as n increases or h decreases indefinitely. Of course, if we consider the process to be infinite, this conclusion applies to every grid point at any discretisation level. However, convergence in the above sense also has favourable consequences on the quality of the numerical solution as related to any point in the equation's definition domain, as we will see in Chapter 3 (cf. Remark 3.2).

2.4.2 Convergence of the $\mathcal{P}_1$ FEM

In order to establish convergence results for the $\mathcal{P}_1$ FEM applied to (P_3), we will follow the very same recipe already exploited for the FDM. However, in order to do so, it is convenient to consider the following variant of (P_3):

$$\begin{cases} \text{Given } f \text{ and } g \in L^2(0,L) \text{ find } w \in V \text{ such that} \\ \int_0^L pw'v'dx + \int_0^L qwv\ dx = \int_0^L fv\,dx + \int_0^L gv'dx,\ \forall v \in V \end{cases} \quad (P'_3)$$

The idea is to add to the right side a term that mimics the one appearing on the left side of (P_3), but not on its right side. Problem (P'_3) has a unique solution, and we may consider its linear FE approximation denoted by $w_h \in V_h$, that is, the solution of

$$\begin{cases} \text{Find } w_h \in V_h \text{ such that} \\ \int_0^L pw_h'v'dx + \int_0^L qw_hv\,dx = \int_0^L fv\,dx + \int_0^L gv'dx,\ \forall v \in V_h \end{cases} \quad (2.39)$$

Since we are considering the case where $a = b = 0$, the following stability inequalities hold for w_h:

$$\begin{cases} \| w_h' \|_{0,2} \le C_1[\| f \|_{0,2} + \| g \|_{0,2}], \\ \| w_h \|_{0,2} \le C_1'[\| f \|_{0,2} + \| g \|_{0,2}]. \end{cases} \quad (2.40)$$

A simple adaptation of the stability analysis for equation (2.10) carried out in Section 2.1, leading to equations (2.19) and (2.20), allows us to establish stability inequality (equation (2.40)) without any significant difficulty (cf. Exercise 2.9).

Now we recall that $\tilde{u}_h$ is the interpolate of u in V_h^+ defined in Subsection 2.3, together with the consistency results (equation (2.36)). Notice that in the particular case we are considering, $V_h^+ = V_h$. If we replace w_h on the left side of equation (2.39) by $\tilde{u}_h - u_h$, this corresponds to taking $f = r_h(u)$ and $g = s_h(u)$. Therefore applying the stability result (equation (2.40)) we readily obtain

$$\begin{cases} \| [\tilde{u}_h - u_h]' \|_{0,2} \le C_1(\| r_h(u) \|_{0,2} + \| s_h(u) \|_{0,2}), \\ \| [\tilde{u}_h - u_h] \|_{0,2} \le C_1'(\| r_h(u) \|_{0,2} + \| s_h(u) \|_{0,2}), \end{cases} \quad (2.41)$$

Recalling equation (2.36), this leads to the following estimate:

$$\| \tilde{u}_h - u_h \|_{0,2} + \| [\tilde{u}_h - u_h]' \|_{0,2} \le \max[C_1, C_1'][C_r h^2 + C_s h] \| u'' \|_{0,2}. \quad (2.42)$$

The error estimate in equation (2.42) should be sufficient to persuade the user that the $\mathcal{P}_1$ FEM converges with order one in the L^2-norm of the derivative of u, by requiring only that the second-order derivative of u is square integrable in $(0,L)$. However, in this case, we can go further than with the FDM, since we can exhibit the order with respect to norms of $u - u_h$.

Indeed, since $u - u_h = (u - \tilde{u}_h) + (\tilde{u}_h - u_h)$, using the triangle inequality we derive

$$\| u - u_h \|_{0,2} \leq \| \tilde{u}_h - u \|_{0,2} + \| \tilde{u}_h - u_h \|_{0,2}$$

together with a similar relation for the derivatives. Then, noticing that estimates of the same type as equation (2.42) hold for the interpolation error $u - \tilde{u}_h$ according to equation (2.37), and since $h^2 \leq Lh$, we finally obtain for a suitable mesh-independent constant $\tilde{C}$ the

Error estimate for the $\mathcal{P}_1$ FEM

$$\boxed{\begin{array}{l} \| u - u_h \|_{1,2} \leq \tilde{C} h \, \| u'' \|_{0,2}, \\ \text{where } \| u - u_h \|_{1,2} := [\| u - u_h \|_{0,2}^2 + \| [u - u_h]' \|_{0,2}^2]^{1/2} \end{array}} \quad (2.43)$$

Otherwise stated, we have just established that the $\mathcal{P}_1$ FEM is first-order convergent in the L^2-norms of both the function and its first-order derivative. The reader is encouraged to check the path leading to the value of $\tilde{C}$ in terms of L, α, *A* and *B*, in order to consolidate the understanding of such convergence result (cf. Exercise 2.10).

The part of equation (2.43) stating that the $\mathcal{P}_1$ FEM is first-order convergent in the L^2-norm is not optimal. Actually, provided we can assert that u'' belongs to $L^2(0, L)$ *for all* $f \in L^2(0, L)$, it is possible to prove that the order of convergence of $\| u - u_h \|_{0,2}$ as h goes to zero is two. Such a regularity of u'' holds under some suitable assumptions on both p' and q. Here, in order to simplify the argument, besides the boundedness assumptions on p and q of Section 1.1, we assume that p' is bounded and discontinuous to the most at a countable set of points in $(0, L)$. Moreover, for the sake of brevity, we consider only the case where $a = b = 0$, leaving the extension to the general case as Exercise 2.11. Let us verify the assertion that $u'' \in L^2(0, L)$ under such assumptions.

First, we note that p must be continuous and satisfy the corresponding condition in equation (1.1). Then, from (P_1), we have

$$pu'' = -f + qu - p'u'.$$

Therefore, letting C be an upper bound of $|p'(x)|$ in $(0, L)$, we trivially have

$$|u''(x)| \leq (|f(x)| + B|u(x)| + C|u'(x)|)/\alpha \; \forall x \in (0, L).$$

By a simple application of the Cauchy–Schwarz inequality in $\mathfrak{R}^3$, we easily obtain

$$|u''(x)|^2 \leq 3(|f(x)|^2 + B^2|u(x)|^2 + C^2|u'(x)|^2)/\alpha^2 \ \forall x \in (0, L). \tag{2.44}$$

Now, the stability inequality (equation (2.24)) in energy norm that holds for (P_1) implies that

$$\| u' \|_{0,2} \leq \frac{\sqrt{2}C_e}{\sqrt{\alpha}} \| f \|_{0,2}. \tag{2.45}$$

Moreover, from the Friedrichs–Poincaré inequality, we have

$$\| u \|_{0,2} \leq \frac{2\sqrt{2}C_e L}{\sqrt{\alpha}} \| f \|_{0,2}. \tag{2.46}$$

Then, after straightforward manipulations, it follows from equations (2.44), (2.45) and (2.46) that there exists a constant C'' depending only on p, q and L (the reader should determine this constant as Exercise 2.12), such that

$$\| u'' \|_{0,2} \leq C'' \| f \|_{0,2}, \tag{2.47}$$

and we are done.

Next, we proceed to the finer estimate of the error $\| u - u_h \|_{0,2}$ using a so-called **duality argument**. Provided $u \neq u_h$ we can write:

$$\| u - u_h \|_{0,2} = \frac{(u - u_h | u - u_h)_0}{\| u - u_h \|_{0,2}} \tag{2.48}$$

Now, let v be the solution of the **adjoint problem** [2], namely, equation (P_1) taking $f = u - u_h$. Since $u - u_h \in L^2(0, L)$, we have $v'' \in L^2(0, L)$ and, from equation (2.47), $\frac{\| v'' \|_{0,2} \leq C''}{\| u - u_h \|_{0,2}}$. Then, from the definition of v and equation (2.48), we have

$$\| u - u_h \|_{0,2} \leq C'' \frac{(u - u_h | -[pv']' + qv)_0}{\| v'' \|_{0,2}}$$

or yet, using integration by parts and recalling equation (2.22),

$$\| u - u_h \|_{0,2} \leq 2C'' \frac{(u - u_h | v)_e}{\| v'' \|_{0,2}} \tag{2.49}$$

Let $\tilde{v}_h \in V_h$ be the interpolate of v at the mesh nodes. Still assuming that the integral of fv_h in $(0, L)$ is computed exactly, we obviously have $(u|\tilde{v}_h)_e = (u_h|\tilde{v}_h)_e$. Plugging this relation into equation (2.49), it follows that

$$\| u - u_h \|_{0,2} \leq 2C'' \frac{(u - u_h | v - \tilde{v}_h)_e}{\| v'' \|_{0,2}}$$

From Subsection 2.2.2, we know that, as long as $q > 0$, $(\cdot|\cdot)_e$ (resp. $\| \cdot \|_e$) is an inner product (resp. a norm) on the space $H^1(0, L)$ of those functions in $L^2(0, L)$, whose first-order derivatives also belong to $L^2(0, L)$. Moreover, it can be easily shown that $\| w \|_e \leq \sqrt{\max[A, B]/2} \, \| w \|_{1,2} \; \forall w \in H^1(0, L)$. Then, using this inequality together with the Cauchy–Schwarz inequality applied to $(\cdot|\cdot)_e$, we further obtain

$$\| u - u_h \|_{0,2} \leq C'' \max[A, B]/2 \, \| u - u_h \|_{1,2} \frac{\| v - \tilde{v}_h \|_{1,2}}{\| v'' \|_{0,2}} \tag{2.50}$$

Finally resorting to equation (2.37), we can assert that there exists a constant C^* depending neither on the mesh nor on u, such that $\| v - \tilde{v}_h \|_{1,2} \leq C^* h \, \| v'' \|_{0,2}$. Recalling equation (2.43), this readily yields the

Error estimate for the $\mathcal{P}_1$ FEM in the L^2-norm

$$\boxed{\| u - u_h \|_{0,2} \leq C_0 h^2 \, \| u'' \|_{0,2} \, ,} \tag{2.51}$$

where $C_0 = \tilde{C} C^* C'' \max[A, B]$.

2.4.3 Remarks on the Convergence of the FVM

In this subsection, we highlight the convergence properties of the FVM. Basically, the convergence of the Vertex-centred FVM can be studied in the same manner as the FDM, that is, in the sense of the maximum norm. This is because, similarly to the FD scheme with a non-uniform grid, this FV scheme also satisfies a DMP. However, from the formal point of view, there is a fundamental difference between both methods. Indeed, in the case of the FVM, we are dealing with piecewise constant functions and not only with values at a finite set of grid points. However, such a subtlety plays no important role in the convergence analysis.

More concretely, according to equation (1.26), the Vertex-centred FV scheme is nothing but the Three-point FD scheme with a non-uniform grid. Therefore, even if both methods are conceptually different, in practical terms the corresponding convergence results can only be identical. For example, in the case of constant p and q, and a uniform mesh, provided $u^{(iv)}$ is bounded in $[0, L]$, the Vertex-centred FVM is first-order convergent in the maximum norm. Here,

the expected result is the

Error estimate for the Vertex-centred FVM with a uniform mesh

$$\| u - u_h \|_{0,\infty} \leq C_v [h \| u' \|_{0,\infty} + h^2 \| u^{(iv)} \|_{0,\infty}] \tag{2.52}$$

where C_v is a mesh-independent constant; and u_h is the piecewise constant function whose value in the jth FV is u_j. Estimate (2.52) can be derived using the triangle inequality:

$$\| u - u_h \|_{0,\infty} \leq \| \underline{u}_h - u_h \|_{0,\infty} + \| u - \underline{u}_h \|_{0,\infty},$$

where $\underline{u}_h$ is the piecewise constant function whose value in the jth CV V_j is $u(x_j)$. Indeed, from Subsection 2.4.1, we know that $\| \underline{u}_h - u_h \|_{0,\infty} \leq C_d h^2 \| u^{(iv)} \|_{0,\infty}$ for a suitable mesh-independent constant C_d. Furthermore, from the theory of polynomial interpolation (cf. [158]), for a suitable mesh-independent constant C_I it holds that

$$\max_{x \in V_j} |u(x) - \underline{u}_h(x)| \leq C_I h \max_{x \in [0,L]} |u'(x)| \; \forall j.$$

This readily implies equation (2.52).

If the mesh is not uniform, the same qualitative results can be derived on the basis of the corresponding ones applying to the FDM. Still taking constant p and q for simplicity and using a DMP, the reader may check that, in this case, the method's convergence relies on the following:

Error estimate for the Vertex-centred FVM with a non-uniform mesh

$$\| u - u_h \|_{0,\infty} \leq C'_v h [\| u' \|_{0,\infty} + \| u''' \|_{0,\infty}] \tag{2.53}$$

for a suitable mesh-independent constant C'_v.

As far as the Cell-centred FVM is concerned, except for particular cases, the situation is quite different. We can also derive a DMP for this scheme, and the stability in the maximum norm easily follows. However now, even for a uniform mesh, in contrast to the inner cells, the local truncation error in the cell $(0, h)$ is not even an $O(h)$. In order to check this, let us write two Taylor expansions about the point $x = h/2$, assuming that the third derivative of u is continuous in $(0, L)$. For suitable $\xi^- \in (0, h/2)$ and $\xi^+ \in (h/2, 3h/2)$, we have

$$0 = u(0) = u(h/2) - (h/2)u^{'}(h/2) + (h/2)^2 u^{''}(h/2)/2 - (h/2)^3 u^{'''}(\xi^{-})/6$$

and

$$u(3h/2) = u(h/2) + hu^{'}(h/2) + h^2 u^{''}(h/2)/2 + h^3 u^{'''}(\xi^{+})/6.$$

Hence, taking a constant p, we have

$$p\{u(h/2)/(h/2) + [u(h/2) - u(3h/2)]/h\}/2 = -3phu^{''}(h/2)/4 + O(h^2),$$

where the term $O(h^2)$ stands for $ph^2[u^{'''}(\xi^{-})/24 - u^{'''}(\xi^{+})/6]$.On the other hand, the scheme's residual at $x = h/2$ in the pointwise sense is given by

$$[r_h(u)]_{1/2} := p\{u(h/2)/(h/2) + [u(h/2) - u(3h/2)]/h\}/h + qu(h/2) - f(h/2)$$

Thus, $[r_h(u)]_{1/2}$ is seen to equal $pu^{''}(h/2)/4 + O(h)$. Such an $O(1)$ term is not sufficient to establish the convergence in the maximum norm of $\vec{u}_h$ using the same logic as in Subsection 2.4.1. Actually, it has been observed for a long time that the local truncation errors in the FD sense can be much larger than the true absolute errors at CVs' representative points (see e.g. [62]). Fortunately, convergence results in the maximum norm can be proven to hold for Cell-centred FV schemes using more tricky arguments (cf. [68] and references therein), rather than a mere evaluation of local truncation errors. It is even possible to prove second-order convergence, under suitable regularity assumptions on u, as long as the CV representative points are located at their centres (cf. [74]). However, the reader should be aware of the fact that equivalent results cannot be proven for the higher dimensional counterparts of the Cell-centred FVM.

We refer to reference [68] for further details on all those issues related to the FVM. In this book, we confine ourselves to establishing the following result applying to equation (1.28):

Error estimate for the Cell-centred FVM with a non-uniform mesh

$$\boxed{\| u - u_h \|_{0,2} \leq C_c(u)h,} \qquad (2.54)$$

where $C_c(u)$ is a constant depending only on L, p, q and the derivatives of u, u_h being the piecewise constant function whose value in the jth cell is $u_{j-1/2}$ (cf. Subsection 1.4.2). Incidentally, we prefer postponing the justification of equation (2.54) to Chapter 5, as a sort of by-product of results proven to hold for two-dimensional FV schemes analogous to the one-dimensional Cell-centred FV scheme. After examining this material, the reader will be able to check the validity of this error estimate rather easily, and obtain a precise expression for $C_c(u)$,

in the case of a constant p (cf. Exercise 5.17). Finally, it is worth commenting that, in the literature, the quality of the FVM is frequently evaluated by using (discrete) mean-square norms, instead of the maximum norm. For a more comprehensive study of error estimates for the FVM, the author refers to reference [68].

Remark 2.5

In view of the numerical results of Example 1.1, one could effectively conjecture that error estimate (2.54) is suboptimal. Notice, however, that in this case the errors were evaluated only at the representative points, where ***superconvergence*** *takes place. We refer to test problems 1 and 3 of Example 2.1 hereafter for further explanations on this effect.*

2.4.4 Example 2.1: Sensitivity Study of Three Equivalent Methods

The aim of this example is to illustrate the convergence rate of the FDM, the $\mathcal{P}_1$ FEM and the Vertex-centred FVM for solving problem (P_1) in the interval $(0,1)$ by means of numerical values. In particular, we shall deal with the case where the three methods are equivalent, which means that they give rise to the same SLAE. Besides experiments performed under the conditions estimates (2.38), (2.43), (2.51) and (2.52) are valid, we consider situations not addressed in the theory developed in this chapter. More specifically, we take the cases of a discontinuous p and of an unbounded q, for which the exact solution u is known. Moreover, in order to figure out what can be expected in terms of convergence in case f is not in $L^2(0,1)$ (even though not too 'wild'), we perform a **sensitivity analysis** pretending that the exact solution to one of the test problems is unknown. This means that we observe at which rate the numerical solutions obtained with a sequence of nested grids or meshes tend in different norms to the one computed with a mesh much finer than all of them. Three test problems are solved whose exact solution is known, with both smooth and irregular data. In the fourth test problem, the theoretical solution is unknown and a sensitivity analysis was carried out. In all cases, uniform meshes with n intervals were used for $n = 2^m$, with $m = 3, 4, 5, 6, 7$. For each test problem, we display in the subfigure on the right the errors of the solution computed by means of the three equivalent FD–FV–FE methods. We recall that such an equivalence occurs when the trapezoidal rule is employed to calculate the integrals of $qu_h v$ and fv in each element in the FE solution, and the CV representative points are selected for the application of the one-point quadrature rule to

approximate the integrals of q and f in each cell in the Vertex-centred FV solution. The logarithm of the absolute errors measured in different norms are plotted against the logarithm of $h := 1/n$. These log–log plots are shown in Figures 2.4–2.6 and 2.7 for test problems 1, 2, 3 and 4, respectively. The straight lines whose slopes roughly correspond to the actual convergence rate are the best fitted to the error data in the least-squares sense. More precisely, we computed the L^2-norm of the solution and its first-order derivative errors (i.e. the errors in the L^2- and in the H^1-semi-norm), and also in the discrete L^∞-norm, that is, the error maximum absolute value restricted to the mesh nodes. The errors $u - \underline{u}_h$ for the standard FE solution $\underline{u}_h$ are plotted on the left (part a), while the errors $u - u_h$ for the FD–FV–FE solution u_h are displayed on the right (part b). In the latter case, we also show the evolution of the L^2-norm of $u - \bar{u}_h$ where $\bar{u}_h$ is the piecewise linear function that takes the value of u_h at the mesh nodes. In notational terms, we replaced the discrete first-order derivative $u_h'^h$ by the exact derivative $\bar{u}_h'$, but for uniform meshes the result is the same, as the reader can easily check. Both in the error computations and in the evaluation of integrals inherent to the standard FEM, numerical integration by Simpson's rule in each interval was employed.

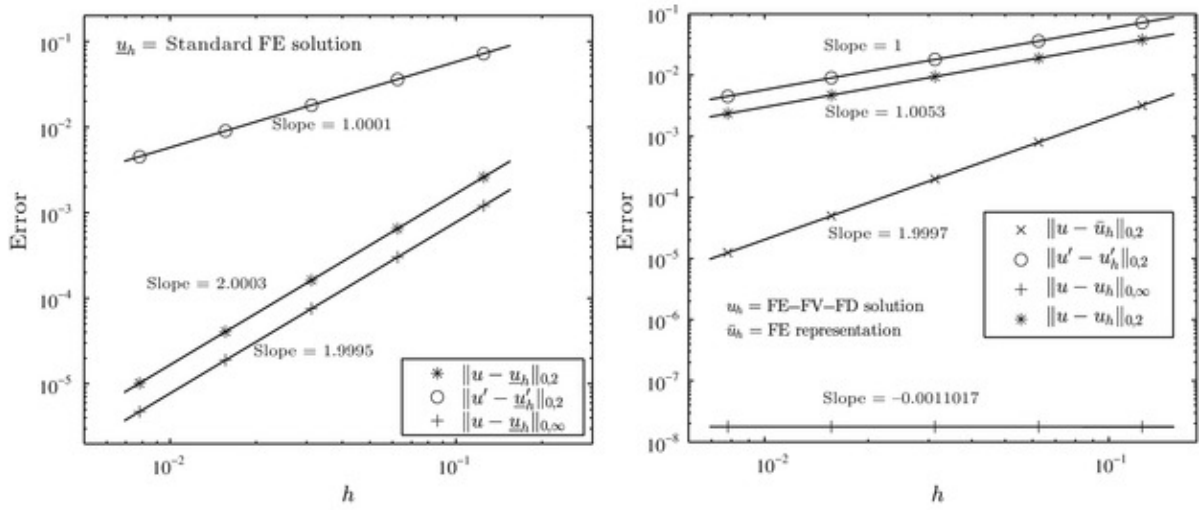


Figure 2.4 Errors for test problem 1.

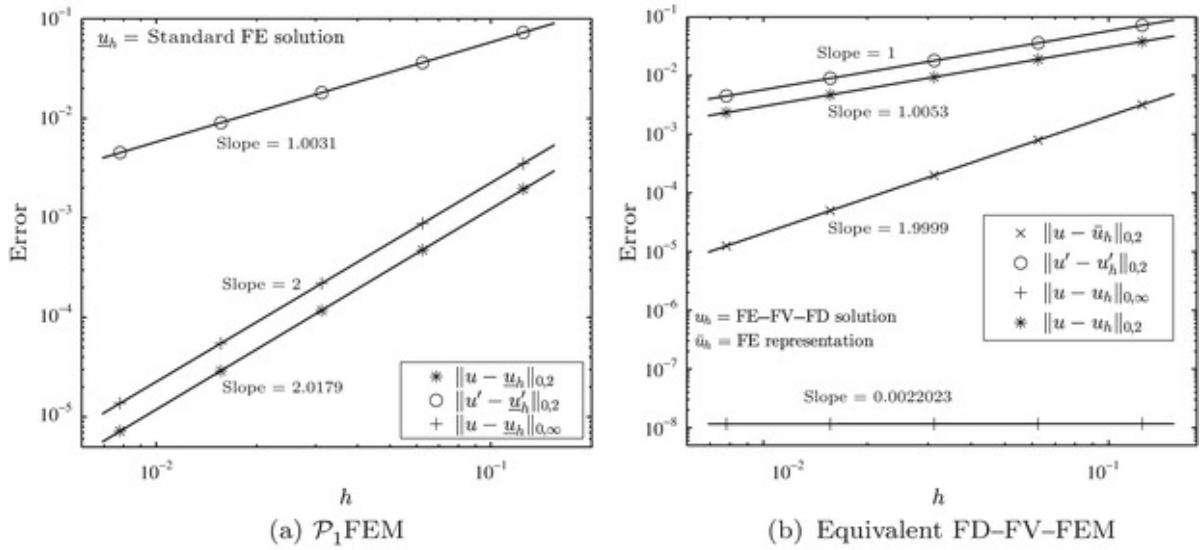


Figure 2.5 Errors for test problem 3.

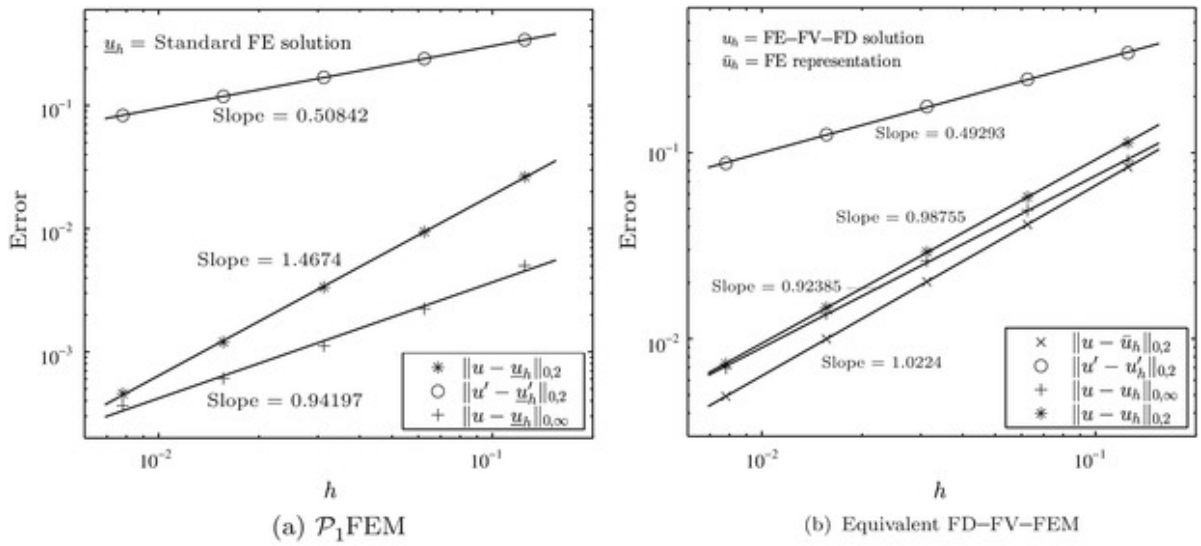
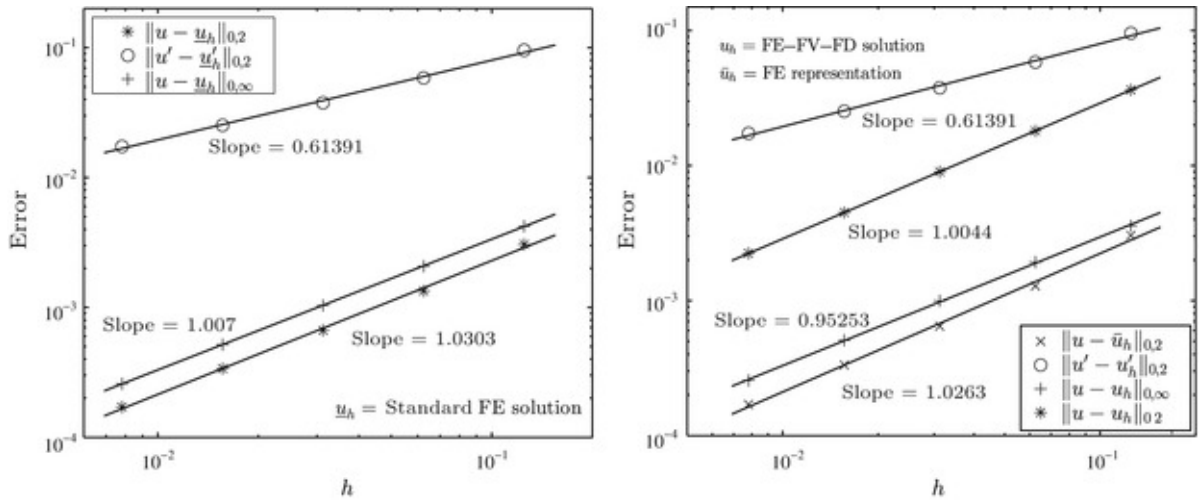


Figure 2.6 Errors for test problem 4.



(a) $\mathcal{P}_1$FEM (b) Equivalent FD–FV–FEM

Figure 2.7 Errors for test problem 2.

Test problem 1: In this test, we took $p = 0.5$, $f(x) = 1 + x$ and a bounded continuous q, namely, $q = x/u(x)$, in such a way that our convergence results apply. The manufactured exact solution is $u(x) = x(2 - x)$. As one can infer from Figure 2.4, the numerical solution at the grid points or mesh nodes are exact practically up to machine precision. This occurs because the exact solution is a polynomial of degree two, and an FD solution reproduces exactly such a function at the grid points, since its third-order derivative vanishes identically (cf. Subsection 2.4.1). The order increase owing to this property of the FDM is carried over to the FE or the FV approximation of a sufficiently smooth function at the mesh nodes. This effect is called **superconvergence** or **supraconvergence**. For more details on this issue, we refer, for instance, to [204], [71] and [21].

Test problem 2: We took the same exact solution u as in test problem 2 of Example 1.1 for a discontinuous p, $q = x/u(x)$ and $f(x) = 1 + x$. We recall that $-[pu']' = 1$ and that the exact solution is 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 u does not belong to $H^2(0, 1)$. Figure 2.5 points to deteriorated convergence rates for both implemented methods. More precisely, the convergence in both the L^2 and the L^∞ senses are roughly of the first order only, while a drop in value apparently from one to the inverse $\Phi - 1$ of the **golden ratio** (or **golden number**) $\Phi = (\sqrt{5} + 1)/2$ can be noticed in the case of the H^1-semi-norm. The one-point decrease in the L^2 convergence rate could be explained by the fact that the convergence rate in H^1 appears twice in the argument leading to the corresponding estimate. But this point requires a more careful analysis, since logically the drop would be rather from two to a value about twice $\Phi - 1$ in this case.

Test problem 3: We were given the same smooth manufactured solution $u(x) = 2x - x^2$ as in test problem 1, but this time we took $p \equiv 0.5$ and $q(x) = 1/xu(x)$. Although $u \in H^2(0,1)$ this case does not really fit into the convergence theory we worked out in this chapter, because neither q is bounded, nor $f \in L^2(0,1)$. Nevertheless, the convergence behaviour of the numerical solution is exactly the optimal one observed in test problem 1, as one can infer from Figure 2.6. Notice that in this test problem, and also in test problem 4, $f = 1 + 1/x$ and q behaves like an $O(1/x^2)$ near the origin. However, both singularities can be removed when computing the left- and right-side terms $\int_0^1 qu_h v \, dx$ and $\int_0^1 fv \, dx$, since both u_h and v are of the form cx for a certain constant c in the leftmost element. In this respect, we recall that in order to let the three methods coincide, we must use a quadrature formula to approximate these integrals. In this way, the singularities of q and f at $x = 0$ are also avoided in practice, if we enforce $[fv](0) = c$ and $[qu_h v](0) = c'$ for suitable constants c and c' in the numerical quadrature formula in case $x = 0$ is one of the quadrature points. Notice that none of these singularities is a problem for the FD or the Vertex-centred FV schemes, since values of q or f at $x = 0$ are nowhere necessary. We refer to Remark 2.6 hereafter for further comments on this test problem.

Test problem 4: We took $q \equiv 1$, but p is the same discontinuous function as in test problem 2 and f is as in test problem 3. It follows that, in this case, the solution u does not belong to $H^2(0,1)$ either. As already pointed out, the convergence rates are observed as if the exact solution was the one obtained by the standard FEM with $n = 2^{12} = 4096$. It is interesting to underline that, in spite of data similarity, the observed convergence rates are quite different from those of test problem 2. Indeed, here the one of the numerical solutions by the standard FEM in $H^1(0,1)$ lowers from one to $1/2$ instead of $\Phi - 1$. Moreover, the convergence rate of the same solution in the L^2-norm is one point greater than in the $H^1(0,1)$-norm, which is less logical than in the case of test problem 2. This could suggest that a kind of superconvergence effect is occurring here, since q is more regular than in the latter case. As for the solution obtained by the three equivalent methods, similar rates are observed, except the aforementioned superconvergence in $L^2(0,1)$. But this is rather coherent with the poorer approximation properties of piecewise constant functions as compared to piecewise linear ones.

Remark 2.6

Classical principles of continuum mechanics require that the total energy $I_e(v)$ *of a bar undergoing longitudinal deformations be finite for every admissible displacement* $v \in V = \{v|\ v \in H^1(0,L), v(0) = 0\}$. *We recall that* $I_e(v) = (v|v)_e - (f|v)_0$, *where* $(u|v)_e = \frac{1}{2}\int_0^L [pu'v' + quv](x)dx$ *(cf. (2.21) and related expressions that follow). In the case of test problem 4, this holds true thanks to the facts that* $v(0) = 0$ *and* q *is bounded. Indeed,* $\int_0^1 f(x)v(x)\,dx = \int_0^1 (1 + 1/x)\left[\int_0^x v'(s)ds\right]dx$, *and by the Cauchy–Schwarz inequality,*

$$\int_0^1 f(x)v(x)dx \leq \int_0^1 \left(\sqrt{x} + \frac{1}{\sqrt{x}}\right)\sqrt{\int_0^x |v'(s)|^2 ds}\,dx$$

This implies that

$$\int_0^1 f(x)v(x)dx \leq \| v' \|_{0,2} \int_0^1 \left(\sqrt{x} + \frac{1}{\sqrt{x}}\right)dx = \frac{8}{3}\ \| v' \|_{0,2}.$$

In practical terms, this result means that the work of the load density $f(x) = 1 + 1/x$, *increasing without a limit as one approaches the end where the bar is kept fixed, is always bounded. Notice that this result would be false if the same end had undergone a given nonzero displacement, say, equal to* a. *Indeed, in this case, any admissible displacement* v *would have to satisfy* $v(0) = a$, *and the work* $(f|v)_0$ *would be unbounded (please check!). In order to draw similar conclusions for test problem 3, more elaborated arguments are necessary, since in this case* q *is also unbounded. But this problem is rather academic, as we are dealing with a smooth manufactured solution. For this reason, we refrain from further commenting on it.*

The observed convergence rates for test problems 2, 3 and 4 require further investigation, but they can certainly be explained using the mathematical tools supplied in this chapter.

2.5 Exercises

2.1 Check that, whatever the inner product $(\cdot|\cdot)$ defined on a linear vector space V, the expression $(v|v)^{1/2}$ defines a norm of all $v \in V$.

2.2 Show that the DMP holds for the Three-point FD scheme with a non uniform grid applied to problem (1.34), taking $b = 0$, $q \equiv 0$ and a constant p.

2.3 Give fine upper bounds for the constants C in equation (2.19) and C' in equation (2.20) in terms of α, B and L.

2.4 Give a fine upper bound for the constant C_e in equation (2.23), by combining equations (2.19) and (2.20).

2.5 Check that a DMP, analogous to the one that holds for the Three-point FDM in case $q(x) > 0$ for every $x \in [0, L]$, applies to the modification of the FV schemes (1.26) and (1.28) in order to incorporate the inhomogeneous boundary condition $u(0) = a$. Conclude the pointwise stability of both schemes in the same sense as for the FDM.

2.6 Give new expressions for the constants C_s and C_r bounding the L^2-norms of the residual functions $s_h(u)$ and $r_h(u)$ proportional to A and B, but not involving the uniformity constant c of a quasi-uniform family of meshes eventually in use. Conclude that the assumption that this family be quasi-uniform is needless for estimating the error $\| u_h - u\|_{1,2}$ by a term of the form $Ch \| u'' \|_{0,2}$ (hint: use the fact that both $|x - x_i|$ and $|x - x_{i-1}|$ are bounded above by h_i $\forall x \in T_i$).

2.7 Give expressions for the constants $\tilde{C}_0$ and $\tilde{C}_1$ in (2.37), independent of the uniformity constant c of a quasi-uniform family of meshes.

2.8 Determine the local truncation errors for the FVMs (1.26) and (1.28) for sufficiently differentiable p and u. What can be concluded about the consistency of each one of these methods? Check in particular the case of $x_{1/2}$. What conclusions can be drawn for scheme (1.28)?

2.9 Prove the validity of equation (2.40) by exhibiting the constants C_1 and C_1'.

2.10 Check the path leading to an expression for the constant $\tilde{C}$ in equation (2.42), in terms of L, α, A and B.

2.11 u being a solution to equation (1.34), verify the validity of the regularity result $u'' \in L^2(0, L)$ under the boundedness assumptions on p and q of Section 1.1, further

assuming that $|p^{'}|$ is bounded by a constant $C_{p}^{'}$ and discontinuous to the most at a finite set of points in $(0, L)$. Determine a constant C_2 in terms of L, B, α and $C_{p}^{'}$ such that $\| u^{''} \|_{0,2} \leq C_2[|a| + |b| + \| f \|_{0,2}]$. Do not assume that q is bounded below in $[0, L]$ by $\beta > 0$.

2.12 Determine an expression for the constant $C^{''}$ in equation (2.47) in terms of L, B, α and $C_{p}^{'}$.

2.13 The purpose of this exercise is to set up a second-order Three-point FD scheme in the maximum norm to solve equation (1.34). The scheme should be a modification of the FD scheme proposed in Exercise 1.7, analogous to equation (2.29). Develop all the steps leading to a second-order convergence result.

2.14 Derive a first-order error estimate for the $\mathcal{P}_1$ FEM (2.10) to solve equation (2.9) in the norm $\| \cdot \|_{1,2}$, assuming that $u \in H^2(0, L)$.

2.15 Based on the result of Exercise 2.14, derive a second-order error estimate in the L^2-norm for equation (2.10), using a duality argument and assuming a suitable regularity of p, q and $p^{'}$.

Notes

1 A trivial consequence of the inequality: $(s\sqrt{\delta} + r/\sqrt{\delta})^2 \geq 0$ for every pair of real numbers $(s; r)$ and $\forall \delta > 0$.

2 Here, the adjoint problem has the same form as the differential equation (P_1) it is related to, but it is not always so. By definition, the adjoint problem of a boundary value problem $\mathcal{L}u = f \in L^2(0, L)$, $\mathcal{L}$ being a differential operator, is $\mathcal{L}^* v = g$ for some given function $g \in L^2(0, L)$, where $\mathcal{L}^*$ is the differential operator such that $(u|\mathcal{L}^* v)_0 = (\mathcal{L}u|v)_0$ for all $u, v \in H^1(0, L)$ satisfying suitable additional integrability and boundary conditions, not necessarily identical.