COMPUTER METHODS IN APPLIED MECHANICS AND ENGINEERING 45 (1984) 57–96
NORTH-HOLLAND

FINITE ELEMENT METHODS: PRINCIPLES FOR THEIR SELECTION*

D.N. ARNOLD, I. BABUŠKA and J. OSBORN
Department of Mathematics, University of Maryland, College Park, MD 20742, U.S.A.

Received 1 March 1983

Principles for the selection of a finite element method for a particular problem are discussed. These principles are stated in terms of the notion of approximability, optimality, and stability. Several examples are discussed in detail as illustrations. Conclusions regarding the selection of finite element methods are summarized in the final section of the paper.

1. Introduction

Larger and larger classes of finite element methods are becoming available for the approximate solution of engineering problems and the selection of a method for a particular problem is an increasingly important question. It is the purpose of this paper to discuss some principles for the selection of finite element methods.

A finite element method or, more generally, a variational method is a discretization of a variational (weak) formulation of the problem under consideration. More specifically, it consists of several items.

(1) *Selection of a variational (weak) formulation of the original problem.* There are, in fact, many such formulations and their choice can significantly affect the resulting finite element method. The choice of a variational formulation leads to the choice of a bilinear form.

(2) *Selection of a trial space.* The trial space consists of those elements (shape functions) with which the solution will be approximated. It is thus chosen so as to provide good approximation properties. The choice of trial space depends, of course, on the set of possible exact solutions under consideration.

(3) *Selection of a test space.* This space is chosen so that the approximate solution is easily computed and so that the error is comparable with the error in the best possible approximation achievable by elements in the trial space.

(4) *Selection of the norm.* The selection of the norm relates to the measure of acceptability of the approximate solution and thus depends on the goals of the computation.

(5) *Selection of the extension procedure.* This procedure describes the manner in which the

* The work of the first and third authors was partially supported by the National Science Foundation under Grant MCS-78-02851, that of the second author by the Office of Naval Research under Contract N00014-77-C-0623.

trial and test spaces (and possibly the variational formulation) are changed when the desired accuracy is not achieved and the approximate solution has to be improved.

The selection of a method for a specific problem depends on the goals of the computation, implementational questions and other practical circumstances. Selection and comparison of methods is not a simple task and in order to find 'optimal' methods it is important to clarify as much as possible the basic notion of a variational method and criteria by which different methods can be compared. It is necessary to emphasize that selection of methods depends on many factors and will always have a relative character. The influence of computer technology on the selection process could be especially important.

Let us turn now to a brief outline of the paper. The paper is partially expository in nature, with the mathematical results having primarily an illustrative as opposed to a practical importance.

Section 2 deals with the principle ideas of a variational method. Section 2.1 introduces the notions of simple variational, variational, directed variational and computational variational methods, concepts we view as important for the discussion of the multiplicity of methods considered today in theory and practice. Section 2.2 discusses approximability and optimality. Approximability refers to the quality of best approximation achievable by the trial space (see (2) above) and optimality refers to the comparison of the approximation yielded by the finite element solution and the best possible approximation achievable by elements in the trial space (see (3) above). In Section 2.3 optimality is elaborated on and related to stability. We introduce the stability constant which is often relatively easy to estimate and in terms of which one can estimate the optimality constant. The ideas introduced in Section 2 are illustrated by a series of examples.

In Section 3 we consider further examples of finite element methods which illustrate several of the ideas introduced in Section 2.

In Section 4 we summarize those conclusions regarding the selection of finite element methods that can be drawn from the discussion in Sections 2 and 3.

Throughout the paper we will use certain function spaces. For an interval $I = (\alpha, \beta)$,

$$L_p(I) = L_p = \begin{cases} \left\{u: \int_I |u|^p \mathrm{d}x < \infty\right\}, & 1 \leq p < \infty, \\ \{u: u \text{ bounded}\}, & p = \infty. \end{cases}$$

On $L_p(I)$ we use the norm

$$|u|_{L_p} = \begin{cases} \left(\int_I |u|^p \mathrm{d}x\right)^{1/p}, & 1 \leq p < \infty, \\ \operatorname{ess\,sup}_{x \in I} |u(x)|, & p = \infty. \end{cases}$$

$H^k(I) = H^k$ is the usual Sobolev space of functions whose first k derivatives are in $L_2(I)$. On this space we use the norm

$$|u|_{H^k(I)} = \left(\sum_{j=0}^{k} |u^{[j]}|^2_{L_2(I)}\right)^{1/2}, \quad u^{[j]} = \frac{\mathrm{d}^j u}{\mathrm{d}x^j}.$$

By $C^l(\bar{I})$ we denote the space of functions whose first l derivatives are continuous on $\bar{I} = [\alpha, \beta]$ with the norm

$$|u|_{C^l(\bar{I})} = \sum_{j=0}^{l} \max_{x \in \bar{I}} |u^{(j)}(x)| .$$

$(L_p)^2$ will denote the spaces of pairs of functions in L_p.

2. Variational methods

2.1. Fundamental ideas

Throughout this paper we suppose we are interested in approximating the unique solution u_0 of some problem by a variational method of discretization. (It is not necessary to give a precise statement of the problem here.) In this section we shall formulate some important ideas which will allow us to discuss variational methods of discretization.

A *simple variational method* is specified by a linear vector space $\mathcal{H}$, a finite dimensional subspace $S \subset \mathcal{H}$ called the *trial space*, a second finite dimensional space V of the same dimension as S called the *test space*, and a *bilinear form* B defined on $\mathcal{H} \times V$. We assume that $B(s, v)$ is *regular* on $S \times V$, i.e., we assume that for every $0 \neq s \in S$ there is a $v \in V$ such that $B(s, v) \neq 0$. This condition is referred to as regularity since, if $\{\varphi_i\}_{i=1}^N$ and $\{\psi_i\}_{i=1}^N$ are bases for S and V, respectively, then $B(s, v)$ is regular if and only if the matrix $B(\varphi_j, \psi_i)$ is regular (or invertible). We will denote the simple variational method by the four-tuple $M = (\mathcal{H}, S, V, B)$.

M is used to determine an approximate solution $u_0(M) \in S$, called the *M-approximate solution*, to the exact solution u_0, which is assumed to lie in $\mathcal{H}$, by requiring that

$$B(u_0(M), v) = B(u_0, v) \tag{2.1.1}$$

holds for all $v \in V$. Using the bases $\{\varphi_i\}$ and $\{\psi_i\}$ and writing $u_0(M) = \sum_{j=1}^N c_j \varphi_j$, we see that (2.1.1) is equivalent to the system of equations

$$\sum_{j=1}^{N} B(\varphi_j, \psi_i) c_j = B(u_0, \psi_i), \quad i = 1, \ldots, N.$$

Since the matrix of this system is regular, the coefficients c_j and hence the approximate solution $u_0(M)$ is uniquely determined. In this way we associate with each $u \in \mathcal{H}$ the M-approximate solution $u(M)$ which is also denoted by Pu or $P(M)u$. Note that $Ps = s$ for any $s \in S$ and therefore that P is a projection. We thus see that with any simple variational method we uniquely associate the projection $P = P(M)$ of $\mathcal{H}$ onto S. Often we will write $M(S, V)$ instead of M to underline the dependence on S and V, especially if the $\mathcal{H}$ and B under consideration are clear from the context. Likewise, instead of $P(M)$ and $u(M)$ we write $P(S, V)$ and $u(S, V)$, respectively.

The space $\mathcal{H}$ must be known, a priori, to contain the exact solution u_0. Furthermore, the bilinear form B must be such that $B(u_0, v)$ is computable from the data that determine the

exact solution u_0, without knowing u_0 explicitly. Let us note that in this section we are not discussing the question of how well the M-approximate solution $u_0(M)$ approximates the exact solution u_0. This is addressed in the next section. Since the exact solution could, a priori, be any element in $\mathscr{H}$ we will often denote it by u.

We now formulate some typical examples; these will be elaborated in the remaining sections.

EXAMPLE 2.1. Let

$$A = \begin{pmatrix} a_{11}(x) & a_{12}(x) \\ a_{21}(x) & a_{22}(x) \end{pmatrix} \tag{2.1.2}$$

be a matrix defined in $I = (-\pi, \pi)$, where $a_{ij} \in L_\infty(I)$, and A is regular (invertible) with

$$A^{-1} = \begin{pmatrix} c_{11}(x) & c_{12}(x) \\ c_{21}(x) & c_{22}(x) \end{pmatrix},$$

where $c_{ij} \in L_\infty(I)$. We then consider the problem of finding $\boldsymbol{u}(x) = (u^{[1]}(x), u^{[2]}(x))^t$ such that

$$A(x)\boldsymbol{u}(x) = \boldsymbol{f}(x) \tag{2.1.3}$$

for a given $\boldsymbol{f}(x) = (f^{[1]}(x), f^{[2]}(x))^t$, where $\boldsymbol{f} \in (L_2)^2$.

A simple variational method $M = (\mathscr{H}, S, V, B)$ is determined by the choices

$$\mathscr{H} = (L_2(I))^2,$$

$$S = V = \Big\{ (s^{[1]}, s^{[2]}) : s^{[k]} = \sum_{j=1}^{n} c_j^{[k]} \sin jx + \sum_{j=0}^{n} d_j^{[k]} \cos jx,$$

$$c_j^{[k]} \text{ and } d_j^{[k]} \text{ real}, \ k = 1, 2 \Big\},$$

$$B(\boldsymbol{u}, \boldsymbol{v}) = \int_{-\pi}^{\pi} \boldsymbol{v}^t A \boldsymbol{u} \, dx. \tag{2.1.4}$$

S and V have the same dimension, namely $N = 4n + 2$. We will later show that B is regular on $S \times V$ under certain additional assumptions.

For any $\boldsymbol{f} \in (L_2(I))^2$, the exact solution $\boldsymbol{u}$ of (2.1.3) lies in $(L_2(I))^2$ and $B(\boldsymbol{u}, \boldsymbol{v})$ is computable in terms of $\boldsymbol{f}$ without knowing $\boldsymbol{u}$ since we have

$$B(\boldsymbol{u}, \boldsymbol{v}) = \int_{-\pi}^{\pi} \boldsymbol{v}^t \boldsymbol{f} \, dx.$$

(We note that we could also take $\mathscr{H} = (L_1(I))^2$ and allow $\boldsymbol{f} \in (L_1(I))^2$, which by our definition would mean a different simple variational method.)

EXAMPLE 2.2. Consider the problem of determining $u(x)$ so that

$$-u''(x) = f(x), \quad x \in I = (-1, 1), \qquad u(-1) = u(1) = 0 \tag{2.1.5}$$

where $f \in L_2(I)$.

(a) Let

$$\mathcal{H} = \mathcal{H}_1 = \mathring{H}^1 \equiv \{u(x): u \in H^1(I),\ u(-1) = u(1) = 0\},$$

$$S = V = \{s(x): s \text{ is a polynomial of degree } n,\ s(-1) = s(1) = 0\}.$$

For $f \in L_2(I)$, the exact solution u of (2.1.5) lies in $\mathcal{H}_1$. The dimension of S and V is $n-1$. Finally we take

$$B(u, v) = -\int_{-1}^{1} uv''\,dx.$$

It is easy to see that B is bilinear on $\mathcal{H}_1 \times V$, is regular on $S \times V$, and that

$$B(u, v) = \int_{-1}^{1} fv\,dx$$

for any $v \in V$, where u is the exact solution of (2.1.5).

(b) We can also take $\mathcal{H} = \mathcal{H}_2 = L_2(I)$. $B(u, v)$ is still defined bilinear in $\mathcal{H}_2 \times V$. This choice for $\mathcal{H}$ is important if f is, for example, a dipole, i.e., the derivative of the Dirac function. In this situation u will be in $\mathcal{H}_2$ but not in $\mathcal{H}_1$. Thus the choice $\mathcal{H} = \mathcal{H}_2$ allows us to treat (2.1.5) with f a dipole. Note that computationally the methods are identical, but according to our definition we are dealing with two different simple variational methods, namely $(\mathcal{H}_1, S, V, B)$ and $(\mathcal{H}_2, S, V, B)$.

EXAMPLE 2.3. Consider the problem of finding $u(x)$ such that

$$u'(x) = f(x), \quad x \in I = (0, 1), \qquad u(0) = 0. \tag{2.1.6}$$

Let $\Delta = \{0 = x_0 < x_1 < \cdots < x_n = 1\}$, where $x_j = j/n$, be a uniform mesh on I and set $I_j = (x_{j-1}, x_j)$, $h_j = x_j - x_{j-1} = 1/n$ and $h = 1/n$. Then let

$$\mathcal{H} = {}^{\circ}H^1 = \{u: u \in H^1(I), u(0) = 0\},$$

$$S = S_n = \{s(x): s \text{ is continuous on } I,\ s \text{ is linear in each } I_j,\ s(0) = 0\}.$$

For the test spaces we will consider two choices:

(a) $V = V_1 = V_{1,n} = \{v(X)\colon v(x)$ is continuous on I, v is linear on each I_j, $v(1) = 0\}$.

(b) $V = V_2 = V_{2,n} = \{v(x)\colon v(x)$ is constant on each $I_j\}$.

In both cases we consider the bilinear form

$$B(u, v) = \int_I u'v \, dx .$$

It is easy to see that $B(u, v)$ is regular on $S \times V$ in both cases.

If $f \in L_2(I)$, then the solution u of (2.1.6) belongs to $\mathcal{H}$ and for any $v \in V$,

$$B(u, v) = \int_I fv \, dx .$$

It often occurs that the approximate solution produced by a simple variational method M_1 is judged to be insufficiently accurate and then another method M_2 is chosen which will give better accuracy. If necessary, a third method M_3 is chosen, and so forth. Although only finitely many computations can be performed in practice, if a procedure is given for determining the new simple method from the previous ones, one is led to an infinite sequence or family of simple variational methods, chosen to produce an approximate solution with acceptable accuracy by considering sufficiently many of the simple methods in the family. The methods M_n are often chosen so as to share the same bilinear form B.

We thus suppose we are given a pair of vector spaces $\mathcal{H}$ and $\mathcal{V}$, a bilinear form B on $\mathcal{H} \times \mathcal{V}$, and consider a family $\mathcal{F}$ of pairs (S, V) of subspaces $S \subset \mathcal{H}$ and $V \subset \mathcal{V}$ so that $\dim S = \dim V < \infty$ and B is regular on $S \times V$. The family of simple methods $M(S, V) = (\mathcal{H}, S, V, B)$ for $(S, V) \in \mathcal{F}$ will be denoted by $\mathcal{M} = (\mathcal{H}, \mathcal{V}, \mathcal{F}, B)$ and will be called *a variational method.* A simple method $M \in (\mathcal{H}, \mathcal{V}, \mathcal{F}, B)$ is completely characterized by the pair $(S, V) \in \mathcal{F}$.

In connection with such families of simple methods it is useful to speak of error (absolute and relative), convergence, and other asymptotic concepts, because the final aim is to obtain an M-approximate solution which approximates the exact solution sufficiently well. Suppose $X \supset \mathcal{H}$ is a Banach space with norm $|\cdot|_X$; we define the absolute error in the approximate solution to be $|u - u(S, V)|_X$. Let α be a function associating to every pair $(S, V) \in \mathcal{F}$ a real positive number $\alpha(S, V)$; $\alpha(S, V)$ will be called the *discretization parameter* associated with the pair (S, V). We say $u(S, V)$ converges to u in X as $\alpha(S, V) \to 0$, written

$$\lim_{\alpha(S, V) \to 0} u(S, V) = u ,$$

if for each ε there is a $\delta > 0$ such that

$$|u(S, V) - u|_X < \varepsilon$$

for any $(S, V) \in \mathcal{F}$ satisfying $\alpha(S, V) < \delta$. We will often consider $\mathcal{H}$ to be equipped with the

norm $|\cdot|_X$. The family $(\mathcal{H}, X, \mathcal{V}, \mathcal{F}, B)$ together with the discretization parameter α will be called a *directed variational method.* For such directed variational methods we will use the notation $\mathcal{M} = (\mathcal{H}, X, \mathcal{V}, \mathcal{F}, B, \alpha)$. The use of $\mathcal{M}$ for both the family of simple methods and the directed method will not cause confusion.

When implementing a variation method we select a sequence $(S_n, V_n) \in \mathcal{F}$, defining the simple methods $M_n = M(S_n, V_n)$, and compute the M_n-approximate solutions $u_n = u(S_n, V_n)$ with increasing n until an acceptable result has been attained. The process of selecting (S_{n+1}, V_{n+1}) from (S_i, V_i), $i = 1, \ldots, n$, and possibly from $u(S_n, V_n)$ and other available information will be called an *extension procedure.* The sequence $\{M_n\}$ will be called a *computational variational method.* Usually our acceptance criterion will be quantified in terms of a norm, as described above.

Let us return now to our examples. We will elaborate the examples introduced earlier.

EXAMPLE 2.1.* Let $\mathcal{F}$ be the family of pairs (S, V), where $S = V$ is the space of trignometric polynomials of arbitrary degree n. Further we select $\mathcal{V} = \mathcal{H} = (L_2(I))^2$. By this selection we have characterized a variational method $\mathcal{M} = (\mathcal{H}, \mathcal{V}, \mathcal{F}, B)$.

Choosing $X = L_2(I)$, $|\cdot| = |\cdot|_{L_2}$ and $\alpha(S, V) = 1/n$, where n is the degree of the polynomials in S (dim $S = 4n + 2$), we define the directed variational method $\mathcal{M} = (\mathcal{H}, X, \mathcal{V}, \mathcal{F}, B, \alpha)$.

Selecting $(S_n, V_n) \in \mathcal{F}$, $S_n = V_n$ being the space of trignometric polynomial of degree n, the extension procedure consists of increasing the polynomial degree by one. The sequence of simple methods $\{M_n\}$, $M_n = (\mathcal{H}, S_n, V_n, B)$, is a computational variational method. A different extension procedure and computational method will be achieved if, for example, the degree of the polynomials is increased in a different way, say by 2.

EXAMPLE 2.2.* Quite analogously to the first example, we select for $\mathcal{F}$ the family of pairs (S_n, V_n), $S_n = V_n$ being the space of the algebraic polynomials of degree n with zero values at $x = 0$, $x = 1$. In case (a) we choose $\mathcal{V} = \overset{\circ}{H}{}^1$, $X = H^1$ and in case (b) $\mathcal{V} = H^2$, $X = L_2(I)$. We further choose $\alpha(S_n, V_n) = 1/n$, where $n - 1 = \dim S_n$. The extension procedure would consist now in some specific manner of increasing the degree of the polynomials. The notions of a variational, directed variational, and computational variational method are now obvious.

EXAMPLE 2.3.* In case (a) we obviously select for $\mathcal{V}$ the space of all continuous, piecewise linear functions subordinate to any uniform mesh which vanish at 1. In case (b) we let $\mathcal{V}$ be the space of all piecewise constant functions subordinate to any uniform mesh. We choose the discretization parameter $\alpha(S, V) = h = 1/n$ and in both cases we consider $X = H^1$. Note that in Examples 2.1* and 2.2* we have $S \subset \tilde{S}$, $V \subset \tilde{V}$ whenever $\alpha(S, V) \geq \alpha(\tilde{S}, \tilde{V})$ but that this is not the case in Example 2.3*.

In Example 2.3*, instead of $X = H^1$ we could select $X = L_2$, etc. We could also choose $\mathcal{H} = H^k \cap {}^\circ H^1$, $k \geq 2$, instead of $\mathcal{H} = {}^\circ H^1$.

2.2. *Approximability and optimality*

Consider a simple variational method $M = (\mathcal{H}, S, V, B) \in \mathcal{M}$, where $\mathcal{M} = (\mathcal{H}, X, \mathcal{V}, \mathcal{F}, B, \alpha)$ is a directed variational method. The purpose of a variational method is to obtain an

M-approximate solution $u(M)$ such that the error (say relative) is smaller than a given tolerance τ. This means that we wish to select M so that

$$|u(M)-u|_X \leq \tau |u|_X \tag{2.2.1}$$

where u is the exact solution of our problem. Thus for a given problem the fundamental goal is as follows: given X and τ, we wish to choose $M=(\mathcal{H}, S, V, B)\in \mathcal{M}$ so that (2.2.1) is satisfied in the most effective way. The word 'effective' must of course be understood in a manner appropriate to the particular situation.

To achieve (2.2.1) it is certainly necessary that

$$Z(u, S, X)=\inf_{s\in S} |u-s|_X/|u|_X \leq \tau . \tag{2.2.2}$$

The quantity $Z(u, S, X)$ measures the relative error in the best possible approximation of u by elements of S with respect to the chosen norm $|\cdot|_X$, i.e., it measures the approximability of u by S with respect to X and is called the *approximability constant of S on U with respect to X* or, more briefly, *the approximability constant.*

That the trial functions are able to approximate the solution well, i.e., that $Z(u, S, X)$ is small, does not alone insure that the M-approximate solution $u(M)$ is close to the exact solution u. We therefore introduce the ratio of the relative error in $u(M)$ to the relative error in the best approximation. For any $u\in \mathcal{H}$ with $|u|_X>0$, define $C(u, M, X)\geq 1$ by

$$|u(M)-u|_X/|u|_X = C(u, M, X)Z(u, S, X) . \tag{2.2.3}$$

If $Z(u, S, X)=0$, we set $C(u, M, X)=1$. The quantity $C(u, M, X)$ measures the optimality of the approximate solution chosen by M and is called the *optimality constant of M on u with respect to X* or, more briefly, *the optimality constant.* When $C(u, M, X)$ is near 1 the approximate solution $u(M)$ is nearly as good as the best possible approximation using the trial space S. We emphasize that, while $Z(u, S, X)$ is independent of the form B and the test space V, the optimality constant $C(u, M, X)$ depends on S, V, B, X and u. The acceptance criterion (2.2.1) is thus simply that the product of $C(u, M, X)$ and $Z(u, S, X)$ does not exceed τ.

Although the exact solution u is unknown in any practical problem, we often know some properties of u, namely that it belongs to $\mathcal{H}$ or to H, where H is a subset of $\mathcal{H}$. We define the *approximability constant of the space S on H with respect to X* as the number

$$Z(H, S, X)=\sup_{u\in H} Z(u, S, X)$$

and the *optimality constant of M on H with respect to X* as

$$C(H, M, X)=\sup_{u\in H} C(u, M, X) .$$

We extend the notion of optimality constant to the directed variational method $\mathcal{M}=$

$(\mathscr{H}, X, \mathscr{V}, \mathscr{F}, B, \alpha)$ by defining

$$C(H, \mathscr{M}, X) = \sup_{M \in \mathscr{M}} C(H, M, X) = \sup_{(S,V)\in\mathscr{F}} C(H, M(S, V), X).$$

The directed variational method $\mathscr{M}$ is called *quasi-optimal on H with respect to X* if $C(H, \mathscr{M}, X)$ is finite.

The set $H \subset \mathscr{H}$ can be arbitrary. E.g., in Example 2.2 (2.2*) in case (a) we have chosen $H = \mathring{H}^1$, but we could select for example $H = H^k \cap \mathring{H}^1$, $k > 1$. We also can select

$$H = \{u \in H^2(I) \cap \mathring{H}^1 : u''(x) > \alpha_1 |u|_{H^2} \text{ on } (\omega, \eta), \text{ where } |\omega - \eta| = \alpha_2\}, \quad \alpha_1, \alpha_2 > 0,$$

etc. Note that Z and C are homogeneous with respect to u, i.e., $Z(cu, S, X) = Z(u, S, X)$ and $C(cu, M, X) = C(u, M, X)$, and so we can restrict our interest to functions u lying on the unit sphere.

The main use of the optimality constant $C(u, M, X)$ and $C(H, M, X)$ is to provide an estimate for the absolute error $|u - u(M)|_X$ in terms of the error of best approximation $\inf_{s\in S}|u - s|_X$. From (2.2.3) we have

$$|u - u(M)|_X \leq C(u, M, X)Z(u, S, X)|u|_X = C(u, M, X) \inf_{s\in S} |u - s|_X,$$

which holds for any $u \in \mathscr{H}$. For any $u \in H$ we have

$$|u - u(M)|_X \leq C(H, M, X) \inf_{s\in S} |u - s|_X.$$

As we have seen above, the effectiveness of a simple variational method is influenced by two factors: approximability, which depends on the selection of the trial space S, and optimality, which depends on the selection of S, V and B. Given X, we thus want to choose M so that $C(u, M, X)$ and $Z(u, S, X)$ are as small as possible. Let us now turn to a more detailed discussion of $C(u, M, X)$. As indicated above we want $C(u, M, X)$ to be not much larger than 1. Now it is known that $C(u, M, X)$ can depend strongly on u (this will be illustrated later by examples). In order to make precise and quantify this notion we say that the solution u is *K-perfect* or *K-perfect with respect to the method M* if $C(u, M, X) \leq K$. We will say u is perfect if it is K-perfect with a small K, otherwise we will call it imperfect. A computational variational method $\{M_i\}$ will be said to be *$\{K_i\}$-perfect* if $C(u, M_i, X) \leq K_i$. In practice we usually want $\{K_i\}$ to be bounded by a fairly small number or to be increasing slowly. We note that even if $C(u, M_i, X) \to \infty$, $u(M_i)$ may still converge to u, for this particular u, since the product $C(u, M_i, X)Z(u, S_i, X)$ may still approach zero.

Since the method M is defined for all $u \in \mathscr{H}$, for a given K it is natural to attempt to characterize the set H of all $u \in \mathscr{H}$ for which $C(u, M, X) \leq K$, i.e., to characterize the largest set H satisfying $C(H, M, X) \leq K$. Usually it is not easy to describe the set in such a way that one could in practice decide whether the exact solution belongs to it, except when $H = \mathscr{H} = X$.

The observation that $C(u, M, X)$ may depend strongly on u is very important. It is common practice to attempt to draw conclusions about the performance of a method M from

experimental computations. These conclusions could be misleading since the computations may have been made for a solution u which is K-perfect with a small K, and for which $Z(u, S, X)$ is small. These conclusions could then be false when some other solutions are considered.

This situation is related to the notion of *robustness*. A simple method M or a directed method $\mathcal{M}$ is said to be *robust* if it performs well in relatively general circumstances. It may be that a less robust method performs better in certain situations. There are, in fact, methods which perform well in certain situations but which fail to converge in others; this is the extreme in nonrobustness. The importance of robustness has to be seen in connection with the observation that computational cost is becoming a smaller and smaller part of the total cost of engineering analysis.

In summary, we would like to have, as much as possible, a directed method $\mathcal{M}$ which is quasi-optimal on $H = \mathcal{H}$ (usually $\mathcal{H} = X$) with the constant $C(H, M, X)$ not too large.

Next we make some simple observations connecting approximability, optimality, convergence and rate of convergence of a directed variational method. A directed variational method is called *convergent on* H if

$$\lim_{\alpha(S, V)\to 0} |u - u(S, V)|_X = 0 \tag{2.2.4}$$

for all $u \in H$. If r is a function defined on $\mathcal{F}$ satisfying $\lim_{\alpha(S, V)\to 0} r(X, V) = 0$, then the method is *convergent with rate r on* H if

$$\sup_{u\in H} \sup_{(S, V)\in F} |u - u(S, V)|_X / r(S, V) < \infty . \tag{2.2.5}$$

We can also define the rate of convergence of a computational variational method $\{M_i\}$. $\{M_i\}$ is said to converge with rate r if

$$\sup_{u\in H} \sup_{i} |u - u(S_i, V_i)|_X / r(S_i, V_i) < \infty . \tag{2.2.6}$$

We note that if a directed variational method is quasi-optimal, then (2.2.4), (2.2.5), (2.2.6), respectively, hold if and only if

$$\lim_{\alpha(S, V)\to 0} Z(u, S, X) = 0 ,$$

$$\sup_{u\in H} \sup_{(S, V)\in \mathcal{F}} Z(u, S, X)|u|_X / r(S, V) < \infty ,$$

$$\sup_{u\in H} \sup_{i} Z(u, S_i, X)|u|_X / r(S_i, V_i) < \infty ,$$

respectively, hold.

We now return to our examples.

EXAMPLE 2.4. We consider the problem introduced in Examples 2.1 and 2.1* with the

matrix $A(x)$ chosen as

$$A(x) = \begin{pmatrix} \lambda - \cos x & \sin x \\ -\sin x & \lambda - \cos x \end{pmatrix},$$

with λ real, $\lambda \neq \pm 1, 0$. In the discussion of this problem it will be convenient to use complex notation rather than vector notation. Thus regarding $\boldsymbol{u}$ and $\boldsymbol{f}$ as complex valued functions we can write $Au = f$ as

$$(\lambda - e^{ix})u = f, \tag{2.2.7}$$

where we are using complex multiplication. For test and trial space we now take

$$S = S_n = V = V_n = \left\{ v\colon v = \sum_{j=-n}^{n} c_j e^{ijx},\ c_j \text{ complex} \right\}.$$

The bilinear form will now be

$$B(u, v) = \int_{-\pi}^{\pi} u(\lambda - e^{ix})v \, dx.$$

It is easy to see that the bilinear form is regular on $S \times V$ under the assumptions on λ imposed above. We let $\mathscr{H} = \mathscr{V} = L_2(I)$ and $X = L_2(I)$, where here $L_2(I)$ is the complex L_2 space, i.e., the space of square integrable complex valued functions with the norm

$$|u|_X^2 = |u|_{L_2(I)}^2 = \int_{-\pi}^{\pi} |u|^2 \, dx,$$

with $|u|$ being the absolute value of u.

If $u(S_n, V_n) = u_n = \sum_{j=-n}^{n} c_j(n)e^{ijx}$ is the approximate solution, then $c_j(n)$ are determined by

$$\int_{-\pi}^{+\pi} (\lambda - e^{ix})u_n e^{ijx} dx = \int_{-\pi}^{+\pi} (\lambda - e^{ix})u e^{ijx} dx, \quad |j| \leq n. \tag{2.2.8}$$

Writing the exact solution in the form

$$u = \sum_{j=-\infty}^{\infty} c_j e^{ijx}, \tag{2.2.9}$$

we get

$$|u|_X^2 = 2\pi \sum_{j=-\infty}^{\infty} |c_j|^2. \tag{2.2.10}$$

It follows immediately from (2.2.8) that the coefficients $c_j(n)$ must satisfy

$$\lambda c_{-n}(n) = \lambda c_{-n} - c_{-n-1},$$
$$\lambda c_j(n) - c_{j-1}(n) = \lambda c_j - c_{j-1}, \quad j = -n+1, \ldots, n.$$

Letting $z_j(n) = c_j(n) - c_j$, these equations can be written as

$$\lambda z_{-n} = -c_{-n-1},$$
$$\lambda z_j = z_{j-1}, \quad j = -n+1, \ldots, n.$$

If $\lambda \neq 0$, this system is uniquely solvable and we obtain

$$z_{-n+j} = -\lambda^{-1-j} c_{-n-1}, \quad j = 0, 1, \ldots, 2n. \tag{2.2.11}$$

It follows from (2.2.9) and (2.2.10) that

$$Z(u, S_n, X)^2 = 2\pi \Big(\sum_{|j|>n} |c_j|^2 \Big) |u|_X^{-2}$$

and from (2.2.9)–(2.2.11) that

$$|u - u_n|_X^2 = |c_{-n-1}|^2 \lambda^{-2} 2\pi \sum_{j=0}^{2n} \lambda^{-2j} + Z(u, S_n, X)^2 |u|_X^2$$
$$= |c_{-n-1}|^2 \lambda^{-2} 2\pi \frac{1 - \lambda^{-4n-2}}{1 - \lambda^{-2}} + Z(u, S_n, X)^2 |u|_X^2$$

and therefore that

$$C^2(u, M_n, X) = 1 + \frac{|c_{-n-1}|^2 \lambda^{-2} ((1 - \lambda^{-4n-2})/(1 - \lambda^{-2}))}{\sum_{|j|>n} |c_j|^2}. \tag{2.2.12}$$

We will now analyze separately the cases $|\lambda| > 1$ and $|\lambda| < 1$.

(a) First we assume $|\lambda| > 1$. Then from (2.2.12) we see that

$$C(u, M_n, X) \leq K = (1 + 1/(\lambda^2 - 1))^{1/2}$$

for all $u \in L_2(I)$ and for all n, and thus

$$C(\mathcal{H}, \mathcal{M}, X) \leq K.$$

The method is therefore quasi-optimal on $\mathcal{H}$ with respect to X and all solutions are perfect.

(b) Now assume $|\lambda| < 1$. Then we see that

$$C^2(\mathcal{H}, M_n, X) = 1 + \lambda^{-2}(1 - \lambda^{-4n-2})/(1 - \lambda^{-2})$$

and the optimality constant deteriorates very quickly with n and we have

$$C(\mathcal{H}, \mathcal{M}, X) = +\infty.$$

Let now $H_i \subset \mathscr{H}$, $i = 1, 2$, be defined by

$$H_1 = \left\{ u \in L_2(I): |c_{-n-1}|^2 \geqslant \alpha^2 \sum_{|j|>n} |c_j|^2 \text{ for all } n \right\}, \quad 0 < \alpha < 1,$$

and

$$H_2 = \{u \in L_2(I): c_{-k-1} = 0,\ k \geqslant 0\},$$

where c_j are coefficients of the exact solution in (2.2.9). Then obviously

$$C(u, M(S_n, V_n), X) \geqslant |\lambda|^{-2n-1}\alpha$$

for all $u \in H_1$ and we see that all solutions belonging to H_1 are very imperfect for large n. On the other hand it is easy to check that all solutions in H_2 are perfect and $\mathscr{M}$ is quasi-optimal on H_2 ($C(H_2, \mathscr{M}, X) = 1$). Assume now that the extension procedure is such that only (S_{2k}, V_{2k}), $k \geqslant 1$, are used. If

$$H_3 = \{u \in L_2(I): c_{-l} = 0,\ l \geqslant 1 \text{ odd}\},$$

then the method is once more quasi-optimal on H_3. (If another extension procedure would be used then these solutions could be very imperfect.)

Consider now the subset H_4 of H_1 defined by

$$H_4 = \left\{ u \in H_1: 2\pi \sum |c_j|^2 |\lambda|^{-2\kappa|j|} = \|u\|^2 \leqslant a < \infty \right\}, \quad \kappa > 2,$$

then for $u \in H_4$ we have

$$|u|_X^2\, Z(u, S_n, X)^2 \leqslant 2\pi |\lambda|^{2(n+1)\kappa} \sum_{|j|>n} |c_j|^2 |\lambda|^{-2\kappa|j|} \leqslant |\lambda|^{2(n+1)\kappa} \|u\|^2$$

and so

$$|u - u_n|_X \leqslant C\|u\|\, |\lambda|^{-2n-2} |\lambda|^{(n+1)\kappa} = C\|u\|\, |\lambda|^{(n+1)(-2+\kappa)}.$$

Hence we see that the method converges for $u \in H_4$, since $\kappa > 2$, in spite of the fact that the solution is very imperfect. (Nevertheless, we can expect computational difficulties caused by round off errors.) Using discretization parameter $\alpha(S_n, V_n) = n^{-1}$, we see that we have, in the above mentioned case, the rate of convergence $r = |\lambda|^{(n+1)(-2+\kappa)}$ (i.e., exponential rate of convergence) which of course is lower that the error achievable by S_n and characterized by $Z(u, S_n, X)$.

(c) Let us now consider the trial space S_n as before but change V_n to $\hat{V}_n$ defined by

$$V = \hat{V}_n = \left\{ v: v = \sum_{j=-n+1}^{n+1} c_j e^{ijx},\ c_j \text{ complex} \right\}.$$

Now the $c_j(n)$ are determined by (2.2.8) with $j = -n+1, \ldots, n+1$. Repeating now the analysis we easily see that for $|\lambda| < 1$ the method is quasi-optimal on $L_2(I)$, but it faces

difficulties for $|\lambda| > 1$ which are analogous to those faced in the case $|\lambda| < 1$ when (S_n, V_n) was used.

Let us now summarize some main points we have seen in this example.

(1) If we select S_n, V_n for the test and trial space, then in the case $|\lambda| > 1$ all solutions $u \in L_2(I)$ are perfect and the method is quasi-optimal. In the case $|\lambda| < 1$ there are solutions which are very imperfect but there exists a large class of solutions which are perfect.

(2) Although a solution can be very imperfect, the method still can converge.

(3) The optimality constant $C(u, M_n, X)$ could deteriorate exponentially with n and numerical experiments will most likely indicate this very quickly, because it is unlikely that only perfect solutions will be used in the experiments.

(4) Changing the test space from V_n to $\hat{V}_n$ significantly changed the performance of the method. The case $|\lambda| > 1$ has a symmetric 'major' part while the case $|\lambda| < 1$ has a nonsymmetric 'main' part. If the symmetric part is the 'major' one, then usually (as in this case) it is desirable to select the same trial and test spaces, while if the nonsymmetric part in the main one ($|\lambda| < 1$), then different trial and test space are usually recommended.

EXAMPLE 2.5. Consider the problem and the method introduced in Examples 2.2 and 2.2*.

(a) Consider first the choice $\mathcal{H} = \mathcal{H}_1 = \mathring{H}^1$, $X = X_1 = H^1$. For any $u \in \mathcal{H}_1$ we easily see that the equations defining $u_n = u(M_n)$, namely,

$$u_n \in S_n,$$
$$B(u_n, v) = B(u, v) \quad \text{for } v \in V_n,$$

are equivalent to

$$u_n \in S_n,$$
$$\int_{-1}^{1} u_n' v' \, dx = \int_{-1}^{1} u' v' \, dx \quad \text{for } v \in V_n,$$

i.e., u_n is the orthogonal projection of u into S_n with respect to the inner product $(u, v) = \int_{-1}^{1} u'v' \, dx$ on $\mathcal{H}_1$. It follows immediately from this that

$$C(\mathcal{H}_1, S_n, X_1) \leq \sqrt{2}.$$

Thus all $u \in \mathcal{H}_1$ are K-perfect with $K = \sqrt{2}$.

(b) Now we turn to the second choice of space considered in Examples 2.2 and 2.2*, namely $\mathcal{H} = \mathcal{H}_2 = L_2(I)$ and $X = X_2 = L_2(I)$.

For any $u \in X_2$ we can write

$$u = \sum_{i=0}^{\infty} b_i \rho_i \tag{2.2.13}$$

where the ρ_i are the Legendre polynomials and

$$b_i = \tfrac{1}{2}(2i+1)\int_{-1}^{1} u\rho_i \,dx\,. \tag{2.2.14}$$

We note that

$$\|u\|_{X_2}^2 = \sum_{i=0}^{\infty} 2b_i^2/(2i+1)\,.$$

Formally u can also be written as

$$u = \sum_{i=1}^{\infty} d_i\varphi_i \tag{2.2.15}$$

where $\varphi_i = \rho_{i+1} - \rho_{i-1}$ and

$$\begin{aligned} &d_1 = -b_0, \qquad d_2 = -b_1, \\ &d_{i-1} - d_{i+1} = b_i, \quad i = 2, 3, \ldots\,. \end{aligned} \tag{2.2.16}$$

We will derive the formula for $C^2(u, S_n, X_2)$ in the special case when $u(x)$ is even with respect to 0. In this case $b_i = 0$ for i odd and $d_i = 0$ for i even. At first we assume u is a polynomial which satisfies the boundary conditions $u(\pm 1) = 0$. Then both series expansions (2.2.13) and (2.2.15) terminate after a finite number of terms.

Clearly,

$$Z(u, S_n, X_2) = |u - \hat{s}|_{X_2}/|u|_{X_2}$$

where $\hat{s}$ is the projection of u into S_n in the space X_2, i.e., where $\hat{s}$ is characterized by

$$\begin{aligned} &\hat{s} \in S_n\,, \\ &\int_{-1}^{1} \hat{s}v \,dx = \int_{-1}^{1} uv \,dx \quad \text{for } v \in S_n\,. \end{aligned}$$

From the basic properties of Legendre polynomials we see that the degree of φ_i is $i+1$ and $\varphi_i(\pm 1) = 0$. Thus $\varphi_1, \ldots, \varphi_{n-1} \in S_n$ and, moreover, it is easy to see that $\varphi_1, \ldots, \varphi_{n-1}$ form a basis for S_n. We write $\hat{s}$ in the form

$$\hat{s} = \sum_{i=1}^{n-1} (d_i + z_i)\varphi_i$$

and attempt to find the z_i. Since u is even we have $d_i = z_i = 0$ for i even. For $n = 2k$ even, $z_1, z_2, \ldots, z_{n-1}$ are easily seen to satisfy the equations

$$\begin{aligned} &(2 + \tfrac{2}{5})z_1 - \tfrac{2}{5}z_3 = 0\,, \\ &-\tfrac{2}{5}z_1 + (\tfrac{2}{5} + \tfrac{2}{9})z_3 - \tfrac{2}{9}z_5 = 0\,, \\ &\qquad \vdots \\ &-\frac{2}{4k-3}\,z_{2k-3} + \left(\frac{2}{4k-3} + \frac{2}{4k+1}\right)z_{2k-1} = -\frac{2}{4k+1}\,d_{2k+1}\,, \end{aligned} \tag{2.2.17}$$

which can be explicitly solved to obtain

$$z_{2j-1} = C(2j-1)j, \quad j = 1, 2, \ldots, k, \tag{2.2.18}$$

where

$$C = -d_{2k+1}/(2k^2+3k+1). \tag{2.2.19}$$

Now we write

$$u - \hat{s} = -\sum_{i=1}^{n-1} z_i(\rho_{i+1} - \rho_{i-1}) + \sum_{i=n}^{\infty} d_i(\rho_{i+1} - \rho_{i-1}) = \rho + \sigma.$$

With this notation we have

$$|u - \hat{s}|_{X_2}^2 = |\rho + \sigma|_{X_2}^2 = |\rho|_{X_2}^2 + |\sigma|_{X_2}^2 + 2\int_{-1}^{1} \rho\sigma \, dx. \tag{2.2.20}$$

Now

$$\begin{aligned} |\rho|_{X_2}^2 &= \int_{-1}^{1} \left| \sum_{i=1}^{n-1} z_i(\rho_{i+1} - \rho_{i-1}) \right|^2 dx \\ &= \sum_{i=1}^{n-1} z_i \sum_{j=1}^{n-1} z_j \int_{-1}^{1} (\rho_{j+1} - \rho_{j-1})(\rho_{i+1} - \rho_{i-1}) \, dx. \end{aligned}$$

Using the defining equations for $z_1, \ldots, z_{2k-1}$ we thus have

$$|\rho|_{X_2}^2 = -\frac{2}{4k+1} d_{2k+1} z_{2k-1} = \frac{2d_{2k+1}^2 k(2k-1)}{(4k+1)(2k^2+3k+1)}. \tag{2.2.21}$$

Writing σ in the form

$$\sigma = \sum_{j=2k}^{\infty} d_j(\rho_{j+1} - \rho_{j-1}) = -d_{2k+1}\rho_{2k} + \sum_{j=1}^{\infty} b_{2k+2j}\rho_{2k+2j},$$

we see that

$$|\sigma|_{X_2}^2 = \frac{2}{4k+1} d_{2k+1}^2 + 2\sum_{j=1}^{\infty} \frac{b_{2k+2j}^2}{(4k+4j+1)}. \tag{2.2.22}$$

Finally we see that

$$\int_{-1}^{1} \rho\sigma \, dx = \frac{2}{4k+1} z_{2k-1} d_{2k+1} = -|\rho|_{X_2}^2. \tag{2.2.23}$$

Thus, combining (2.2.20)–(2.2.23), we get

$$\begin{aligned} Z^2(u, S_n, X_2)|u|_{X_2}^2 &= |u - \hat{s}|_X^2 = -|\rho|_{X_2}^2 + |\sigma|_{X_2}^2 \\ &= -\frac{2d_{2k+1}^2 k(2k-1)}{(4k+1)(2k^2+3k+1)} + \frac{2}{4k+1} d_{2k+1}^2 + 2\sum_{j=1}^{\infty} \frac{b_{2k+2j}^2}{4k+4j+1} \\ &= \frac{2d_{2k+1}^2}{2k^2+3k+1} + 2\sum_{j=1}^{\infty} \frac{b_{2k+2j}^2}{4k+4j+1}. \end{aligned} \tag{2.2.24}$$

Next we calculate $|u - u_n|_{X_2}$. In terms of the d_i and φ_i we can give a simple formula for u_n, namely,

$$u_n = \sum_{i=1}^{n-1} d_i \varphi_i \,. \tag{2.2.25}$$

To see this observe that

$$u' = \sum_{i=1}^{\infty} d_i \varphi_i' = \sum_{i=1}^{\infty} d_i (2i+1)\rho_i \,,$$
$$u_n' = \sum_{i=1}^{n-1} d_i (2i+1)\rho_i$$

and hence

$$\int_{-1}^{1} (u' - u_n')\varphi_j' \, dx = -\int_{-1}^{1} (u - u_n)\varphi_j'' \, dx = 0, \quad j = 1, \ldots, n-1 \,.$$

From (2.2.15), (2.2.22) and (2.2.25) we get

$$|u - u_n|_{X_2}^2 = \left| \sum_{i=n}^{\infty} d_i \varphi_i \right|_{X_2}^2 = |\sigma|_{X_2}^2$$
$$= \frac{2}{4k+1} d_{2k+1}^2 + 2 \sum_{j=1}^{\infty} \frac{b_{2k+2j}^2}{4k+4j+1} \,. \tag{2.2.26}$$

Finally, combining (2.2.24) and (2.2.26), we have

$$C^2(u, M_n, X_2) = |u - u_n|_{X_2}^2 / |u - \hat{s}|_{X_2}^2$$
$$= \frac{2d_{2k+1}^2/(4k+1) + 2\sum_{j=1}^{\infty} b_{2k+2j}^2/(4k+4j+1)}{2d_{2k+1}^2/(2k^2+3k+1) + 2\sum_{j=1}^{\infty} b_{2k+2j}^2/(4k+4j+1)}$$
$$= \frac{d_{2k+1}^2/(4k+1) + A_{2K+1}}{d_{2k+1}^2/(2k^2+3k+1) + A_{2k+1}} \tag{2.2.27}$$

where

$$A_{2k+1} = \sum_{j=1}^{\infty} b_{2k+2j}^2/(4k+4j+1)$$

and $n = 2k$. We have established (2.2.27) for any even polynomial satisfying the boundary conditions. A simple limiting argument establishes it for any even $u \in L_2(I)$.

It is immediate from (2.2.27) that

$$C_1\sqrt{n} \leq C(X_2, M_n, X_2) \leq C_2\sqrt{n} \tag{2.2.28}$$

and so

$$C(X_2, \mathcal{M}, X_2) = +\infty .$$

From (2.2.27) we also see that if u is such that

$$d_{2k+1}^2/(4k+1) < KA_{2k+1}, \quad \text{for all } k ,$$

then u is $\sqrt{K+1}$-perfect. For the particular choice $u = \bar{u} = 1$, we get $b_j = 0$, $j \geq 1$, and we see that

$$C(\bar{u}, M_n, X_2) = \left(\frac{2k^2+3k+1}{4k+1}\right)^{1/2} \simeq C\sqrt{n} . \tag{2.2.29}$$

By comparing (2.2.28) and (2.2.29) we see that $\bar{u}$ belongs to the set of most imperfect solutions. It is easy to see that

$$Z(\bar{u}, S_n, X_2)|\bar{u}|_{X_2} \leq C/n$$

and hence

$$\|\bar{u} - \bar{u}_n\|_{X_2} \leq Cn^{-1/2} .$$

Thus the method converges for $\bar{u}$ although $C(\bar{u}, S_n, X_2) \uparrow \infty$.

If we define u by $u = \sum_{i=0}^{\infty} b_i \rho_i$ where $b_i = i^{-\varepsilon}$ for i even, with $0 < \varepsilon < \frac{1}{2}$, then we see that

$$\begin{aligned} |u - u_n|_{X_2}^2 \geq \frac{2d_{n+1}^2}{2n+1} &= \frac{2(b_0 + b_2 + \cdots + b_n)^2}{2n+1} \\ &= 2\frac{(1 + 2^{-\varepsilon} + 4^{-\varepsilon} + \cdots + n^{-\varepsilon})^2}{2n+1} \\ &\geq 2\frac{(\frac{1}{2}n\, n^{-\varepsilon})^2}{2n+1} = \frac{n^{2-2\varepsilon}}{2(2n+1)} \uparrow \infty . \end{aligned}$$

This illustrates the fact that the method could diverge even though $Z(u, S_n, X) \to 0$. Although the method is not robust, we see that the optimality constant grows relatively slowly with n ($C \sim \sqrt{n}$).

Let us now summarize some main points we have seen in this example.

(1) The method $\mathcal{M}_1 = (\mathring{H}^1, H^1, \mathcal{F}, B, \alpha)$ is quasi-optimal but the method $\mathcal{M}_2 = (\mathring{H}^1, L_2, \mathcal{F}, B, \alpha)$ is not quasi-optimal, although computationally the methods are identical.

(2) The set of perfect solutions for the method $\mathcal{M}_2$ is relatively large and very likely the method would work well. The performance, in fact, would be good unless the solution under consideration were one of the relatively rare imperfect solutions. The performance can be good also if the solution is imperfect since for reasonably smooth solutions u satisfying the boundary conditions, $Z(u, S_n, L_2) \to 0$ more rapidly than $\sqrt{n} \to \infty$. (Recall that $\sqrt{n}$ gives the growth of the optimality constant.)

2.3. Stability and the stability condition

We have seen that the optimality constants $C(u, M, X)$, $C(H, M, X)$ and $C(H, \mathcal{M}, X)$ play essential roles in understanding variational methods. The notion of K-perfect solutions is, in fact, defined in terms of $C(u, M, X)$. We have also seen in Examples 2.4 and 2.5 that the calculation or estimation of optimality constants can be rather subtle. It is important to be able to estimate these constants for wide classes of problems. Toward this end we introduce the notion of the stability constant which is often easier to estimate and in terms of which we can estimate the optimality constants.

Let $M = (\mathcal{H}, S, V, B)$ be a simple variational method and let $\mathcal{H} \subset X$. For $u \in \mathcal{H}$ define

$$D(u, M, X) = \sup_{s \in S} |P(u+s)|_X / |u+s|_X \tag{2.3.1}$$

and

$$D(H, M, X) = \sup_{u \in H} D(u, M, X)\,. \tag{2.3.2}$$

$D(u, M, X)$ is called the *stability constant of M on u with respect to X* and $D(H, M, X)$ is called the *stability constant of M on H with respect to X*. For a directed variational method $\mathcal{M}$ we define $D(H, \mathcal{M}, X) = \sup_{m \in \mathcal{M}} D(H, M, X)$. $D(H, \mathcal{M}, X)$ is called the *stability constant of $\mathcal{M}$ on H with respect to X*. We say $\mathcal{M}$ is *stable on H with respect to X* if $D(H, \mathcal{M}, X)$ is finite. We now show the relation between the stability and optimality constants.

THEOREM 2.6. The stability and optimality constant satisfy

$$D(u, M, X) - 1 \leq C(u, M, X) \leq D(u, M, X) + 1\,, \tag{2.3.3}$$

and

$$D(H, M, X) - 1 \leq C(H, M, X) \leq D(H, M, X) + 1\,. \tag{2.3.4}$$

PROOF. We easily see that

$$\begin{aligned} |u - Pu|_X = |(u-s) - P(u-s)|_X &\leq |u-s|_X + |P(u-s)|_X \\ &\leq [1 + D(u, M, X)]|u-s|_X \end{aligned}$$

for any $s \in S$, which proves the second inequality in (2.3.3). For any $s \in S$ we have

$$\begin{aligned} \frac{|P(u+s)|_X}{|u+s|_X} &\leq \frac{|(u+s) - P(u+s)|_X + |u+s|_X}{|u+s|_X} \\ &\leq 1 + |u - Pu|_X / |u+s|_X \\ &\leq 1 + |u - Pu|_X / \inf_{s \in S} |u-s| = 1 + C(u, M, X)\,, \end{aligned}$$

which proves the first inequality in (2.3.3). Now (2.3.4) follows immediately from (2.3.3).

We see that a directed variational method is quasi-optimal if and only if it is stable.

Let us now turn to a further discussion of stability. Suppose $\mathcal{H}$ is closed in X. We easily see that $D(\mathcal{H}, M, X) = \|P\|$, the operator norm of P. In Section 2.1 we introduced the idea of a computational variational method as a sequence of simple variational methods $M_n = (\mathcal{H}, S_n, V_n, B)$ selected from a directed variational method $\mathcal{M} = (\mathcal{H}, X, \mathcal{V}, \mathcal{F}, B, \alpha)$. Assume we are given such a sequence and suppose

$$D(\mathcal{H}, M_n, X) = \|P(M_n)\| \to \infty \quad \text{as } n \to \infty . \tag{2.3.5}$$

We will now show that this implies that there is a $z \in \mathcal{H}$ such that the method does not converge to z.

THEOREM 2.7. Suppose (2.3.5) *holds and suppose* $\mathcal{H}$ *is closed in* X. *Then there is a* $z \in \mathcal{H}$ *such that* $P(M_n)z \not\to z$.

PROOF. Suppose the method converges for all $u \in \mathcal{H}$, i.e., $|P(M_n)u - u|_X \to 0$ as $n \to \infty$ for all $u \in \mathcal{H}$. This implies $P(M_n)u$ is bounded for each u. The uniform boundedness principle [6, p. 190] then implies that $\|P(M_n)\|$ is bounded in n. This would contradict the hypothesis (2.3.5).

Theorems 2.6 and 2.7 show that methods which are not quasi-optimal are not robust in the sense that there exists solutions for which the methods do not converge. It should be noted, however, that the solution z in Theorem 2.7 could be very 'wild' or 'irregular' and the computational variational method might still converge for a large class of u in $\mathcal{H}$.

Let us note a further significance of D. Suppose $u \in \mathcal{H}$ and $s \in S$ satisfies $Z(u, S, X) = |u - s|_X$. Set $\xi = Pu - s$. We easily see that

$$|\xi|_X = |Pu - s|_X = |P(u - s)|_X \leq D(u, M, X)|u - s|_X .$$

Thus D is a magnification factor relating $|\xi|_X$ and $|u - s|_X$. Usually we replace $D(u, M, X)$ by $D(\mathcal{H}, M, X)$. It should be noted that this replacement could lead to a rather pessimistic estimate.

We noted earlier that the stability constants are often easier to estimate than the optimality constants. We will now show how the stability constants can be estimated in terms of the bilinear form B.

THEOREM 2.8. Suppose the space V *is furnished with a norm* $|v|_V$. *Assume*

(1) *for any* $u \in \mathcal{H}$ *there is a constant* $C(u)$ *such that*

$$|B(u + s, v)| \leq C(u)|u + s|_X|v|_V \quad \textit{for all } s \in S \textit{ and } v \in V \tag{2.3.6}$$

and

$$\text{(2)} \qquad \inf_{\substack{s \in S \\ |s|_X = 1}} \sup_{\substack{v \in V \\ |v|_V = 1}} |B(s, v)| \equiv \gamma(S, V) > 0 . \tag{2.3.7}$$

Then

$$D(u, M, X) \leq \frac{|C(u)|}{\gamma(S, V)} \tag{2.3.8}$$

for all $u \in \mathcal{H}$.

PROOF. From (2.3.6) and (2.3.7) we have

$$\begin{aligned}\gamma(S, V)|P(u+s)|_X &\leq \sup_{\substack{v \in V \\ |v|_V = 1}} |B(P(u+s), v)| = \sup_{\substack{v \in V \\ |v|_V = 1}} |B(u+s, v)| \\ &\leq C(u)|u+s|_X\end{aligned}$$

for all $s \in S$, from which (2.3.8) follows.

REMARK 2.9. If (2.3.6) holds for all $u \in \mathcal{H}$ with $C(u)$ replaced by $C(\mathcal{H})$, then (2.3.8) holds for all $u \in \mathcal{H}$ with $C(u)$ replaced by $C(\mathcal{H})$, i.e.,

$$D(\mathcal{H}, M, X) \leq C(\mathcal{H})/\gamma(S, V).$$

We can also estimate $D(\mathcal{H}, M, X)$ below by $1/\gamma(S, V)$, provided certain additional assumptions are made. Suppose $\mathcal{H}$ and $\mathcal{V}$, furnished with the norms $|\cdot|_X$ and $|\cdot|_{\mathcal{V}}$, respectively, are a pair of Hilbert spaces or reflexive Banach spaces, B is a bilinear form on $\mathcal{H} \times \mathcal{V}$, and $S \subset \mathcal{H}$, $V \subset \mathcal{V}$ are a pair of finite dimensional subspaces. For $v \in V$ we set $|v|_V = |v|_{\mathcal{V}}$. Now assume

$$|B(u, v)| \leq C|u|_X|v|_{\mathcal{V}} \quad \text{for } u \in \mathcal{H},\ v \in \mathcal{V}, \tag{2.3.9}$$

$$\inf_{\substack{u \in \mathcal{H} \\ |u|_X = 1}} \sup_{\substack{v \in \mathcal{V} \\ |v|_{\mathcal{V}} = 1}} |B(u, v)| \equiv \omega > 0 \tag{2.3.10}$$

and

$$\sup_{u \in \mathcal{H}} |B(u, v)| > 0 \quad \text{for } 0 \neq v \in \mathcal{V}. \tag{2.3.11}$$

We note that in the presence of (2.3.9), (2.3.10) and (2.3.11) are necessary and sufficient for the variational problem

$$\begin{aligned}&u \in \mathcal{H}, \\ &B(u, v) = F(v) \quad \text{for } v \in \mathcal{V}\end{aligned}$$

to be uniquely solvable for each bounded linear functional F on $\mathcal{V}$. In this situation we also have

$$|u|_X \leq \omega^{-1} \sup_{v \in \mathcal{V}} \frac{|F(v)|}{|v|_{\mathcal{V}}} = \omega^{-1}\|F\|.$$

This result is proved in [1, p. 112].

THEOREM 2.10. Suppose (2.3.9)–(2.3.11) *hold. Let* $\bar{u} \in S$ *and set*

$$\frac{1}{|\bar{u}|_X} \sup_{\substack{v \in V \\ |v|_Y = 1}} |B(\bar{u}, v)| \equiv Q(\bar{u}) . \tag{2.3.12}$$

Then there exists a $u \in \mathcal{H}$ *such that* $P(M)u = \bar{u}$ *and*

$$D(u, M, X) \geq \omega / Q(\bar{u}) . \tag{2.3.13}$$

PROOF. Obviously $F(v) \equiv B(\bar{u}, v)$, $v \in V$, defines a bounded linear functional on V. It is possible to show $F(v)$ can be extended to a bounded linear functional on $\mathcal{Y}$ with

$$\sup_{\substack{v \in \mathcal{Y} \\ |v|_Y = 1}} |F(v)| = \sup_{\substack{v \in V \\ |v|_Y = 1}} |F(v)| = Q(\bar{u})|\bar{u}|_X .$$

This is the content of the Hahn–Banach Theorem [6, p. 134].

As noted above, from (2.3.9)–(2.3.11) it follows there is a $u \in \mathcal{H}$ satisfying

$$B(u, v) = F(v) \quad \text{for all } v \in \mathcal{Y},$$

and

$$|u|_X \leq Q(\bar{u})|\bar{u}|_X / \omega . \tag{2.3.14}$$

It follows immediately from the definition of F that $P(M)u = \bar{u}$. Using (2.3.14) and the definition of $D(u, M, X)$ we thus have

$$D(u, M, X) \geq \frac{|Pu|_X}{|u|_X} = \frac{|\bar{u}|_X}{|u|_X} \geq \frac{\omega}{Q(\bar{u})} .$$

This completes the proof.

THEOREM 2.11. Suppose (2.3.7), (2.3.9)–(2.3.11) *hold. Then*

$$D(\mathcal{H}, M, X) \geq \omega / \gamma(S, V) . \tag{2.3.15}$$

PROOF. This result follows immediately from Theorem 2.10 and the definitions of $D(\mathcal{H}, M, X)$ and $\gamma(S, V)$.

Let us now elaborate Theorem 2.10. It shows that functions u with $D(u, M, X)$ large should be sought among functions whose projection $\bar{u} = P(M)u$ have small $Q(\bar{u})$. Of course not every $u \in \mathcal{H}$ satisfying $P(M)u = \bar{u}$ leads to large $D(u, M, X)$, as shown by the fact that for $\bar{u} \in S$, $P\bar{u} = \bar{u}$ and $D(\bar{u}, M, X) = 1$. If we define

$$R(\bar{u}) = \{w \in \mathcal{H} : P(M)w = \bar{u}\} , \tag{2.3.16}$$

then for $u \in R(\bar{u})$, $D(u, M, X)$ will be largest for that u with smallest $|u|_X$. Thus if we define

$$W(\bar{u}) = \inf_{w \in R(\bar{u})} |w|_X ,$$

we see immediately that there is a $w \in \mathcal{H}$ such that

$$D(w, M, X) \geq |\bar{u}|_X / W(\bar{u}) .$$

Obviously every choice of $\bar{u} \in S$ and $u \in R(\bar{u})$ leads to a lower bound for $D(\mathcal{H}, M, V)$:

$$D(\mathcal{H}, M, V) \geq |\bar{u}|_X / |u|_X .$$

This gives a practical tool for estimating $D(\mathcal{H}, M, X)$ from below.

Condition (2.3.7) is easily seen to imply the regularity of the bilinear form B on $S \times V$.

A directed variational method $\mathcal{M} = (\mathcal{H}, X, \mathcal{V}, \mathcal{F}, B, \alpha)$ is said to satisfy the stability condition if

$$\gamma(S, V) \geq \gamma > 0 \tag{2.3.17}$$

for all $(S, V) \in \mathcal{F}$. This condition is sometimes also called the inf–sup condition, or the LBB or BB condition.[1] Theorems 2.6 and 2.8 show that when (2.3.17) holds, then $D(\mathcal{H}, M, X)$ is finite and the method is stable.

An especially important situation occurs when $\mathcal{H} = \mathcal{V}$, $\mathcal{F}$ is the family of pairs (S, V) with $S = V$ and

$$B(u, u) \geq \gamma |u|_X^2, \quad \gamma > 0 \tag{2.3.18}$$

for any $u \in S$, $(S, S) \in \mathcal{F}$. Then $\gamma(S, S) \geq \gamma$. In this case we say that the form is coercive.

Suppose $M_n = M(S_n, V_n)$, $(S_n, V_n) \in \mathcal{F}$, is a computational variational method for which $\gamma(S_n, V_n) \to 0$ as $n \to \infty$. In addition suppose M_n satisfies the hypotheses of Theorem 2.10 for each n. Then Theorem 2.10 shows that there is a sequence $u_n \in \mathcal{H}$ such that

$$D(u_n, M_n, X) \to \infty .$$

Theorem 2.10 does not show the existence of a $u \in \mathcal{H}$ such that $u(M_n) \not\to u$. This, however, follows from Theorem 2.7.

Combining this observation with Theorem 2.8 we see that the stability condition (2.3.17) is necessary and sufficient for quasi-optimality of the directed variational method $\mathcal{M} = (\mathcal{H}, X, \mathcal{V}, \mathcal{F}, B, \alpha)$. The condition is necessary in the sense that if it fails then there will exist at least one $u \in \mathcal{H}$ for which the method fails. As we have noted above, however, the method may still perform well for all solutions of major practical interest. Various statements in the

[1] This terminology relates to [1, 5] where this condition was first introduced in connection with the analysis of the finite element methods.

literature that certain mixed methods work well inspite of the fact that the LBB (BB) condition is violated have to be understood in light of this observation.

Let us return now to the problem and methods introduced in our previous examples and apply the results discussed in this section.

EXAMPLE 2.12. (Cf. Examples 2.1, 2.1* and 2.4). Suppose $A(x)$ in (2.1.3) is of the form

$$A(x) = A_0 + B(x)$$

where A_0 is a constant invertible matrix satisfying

$$\|A_0^{-1}B(x)\| \leq q < 1$$

where $\|\cdot\|$ denotes the matrix norm associated with the Euclidean vector norm. Take $X = \mathcal{H} = (L_2(I))^2$ and let $\mathcal{V} = (L_2(I))^2$ as before.

It is immediate that

$$B(\boldsymbol{u}, \boldsymbol{v}) \leq C|\boldsymbol{u}|_X|\boldsymbol{v}|_{\mathcal{V}} \quad \text{for } \boldsymbol{u} \in \mathcal{H}, \boldsymbol{v} \in \mathcal{V},$$

with $C = 2\max_{i,j}\|a_{ij}(x)\|_{L_\infty(I)}$. Given $\boldsymbol{u} \in \mathcal{H}$, let $\boldsymbol{v} = (A_0^{-1})\boldsymbol{u}$. We easily see that

$$B(\boldsymbol{u}, \boldsymbol{v}) = \int_{-\pi}^{\pi} \boldsymbol{u}A_0^{-1}(A_0 + B)\boldsymbol{u}\, dx = \int_{-\pi}^{\pi} \boldsymbol{u}(I + A_0^{-1}B)\boldsymbol{u}\, dx$$
$$\geq |\boldsymbol{u}|_X^2(1 - q)$$

and

$$|\boldsymbol{v}|_{\mathcal{V}} \leq Q|\boldsymbol{u}|_X$$

where $Q = \|(A_0^{-1})^t\|$, from which we obtain (2.3.10) with

$$\omega = (1 - q)/Q > 0 .$$

Thus conditions (2.3.9) and (2.3.10) hold for this example. Condition (2.3.11) is also easily seen to hold. (Note that the assumption that A_0 is a constant matrix was not used.)

Our main goal is to show that (2.3.7) holds. In Section 2.2 we have proved the infinite dimensional analogue. The proof of (2.3.7) is similar but we use now the fact that A_0 is a constant matrix. For $\boldsymbol{u} \in S$, let $\boldsymbol{v} = (A_0^{-1})^t\boldsymbol{u}$. Then $\boldsymbol{v} \in V$ since A_0 is a constant matrix. The above estimate thus shows that

$$\gamma(S, V) \geq (1 - q)/Q > 0 .$$

Therefore we see that the assumptions in Theorem 2.8 are satisfied and, moreover, the stability condition (2.3.17) is satisfied with $\gamma(S, V) \geq (1 - q)/Q$. Although the quasi-optimality of the method under our assumptions could be proven by the approach shown in Section 2.2, obviously the approach used here is much simpler and can be used in many situations.

EXAMPLE 2.13. (Cf. Examples 2.2, 2.2* and 2.5) Consider the problem introduced in Example 2.2. Let $\mathcal{H} = \mathcal{H}_1 = \mathring{H}^1$, $X = X_1 = H^1$ and $\mathcal{V} = \mathring{H}^1$ with the H^1-norm as in case (a). We obviously have

$$B(u, u) \geq 2^{-1/2}|u|_X^2 \quad \text{for } u \in \mathcal{H}_1,$$

i.e., B is coercive, and we thus see that the assumptions of Theorem 2.8 are satisfied and hence the quasi-optimality of this method has again been established. As we have seen in Example 2.5 a similar analysis would fail in case (b).

EXAMPLE 2.14. (Cf. Examples 2.3 and 2.3*.) Let $\mathcal{H} = \mathring{H}^1$, $H = H^1$, and $\mathcal{V} = L_2$.

(a) First let us consider the space S_n and $V_{2,n}$ introduced in case (b) in Example 2.3. Then (2.3.6) and (2.3.7) are satisfied and as before we conclude that the method is quasi-optimal and stable.

(b) Now consider the case of test spaces $V_{1,n}$ introduced in case (a). Our goal is to show that

$$\mathcal{C}_1 n \leq D(\mathcal{H}, M_n, X) \leq \mathcal{C}_2 n \tag{2.3.19}$$

where $\mathcal{C}_1$, $\mathcal{C}_2 > 0$, independent of n, and with $M_n = M(S_n, V_{1,n})$.

We will prove the left-hand side of (2.3.19) by applying Theorem 2.10. To this end we construct a particular $\bar{u}_n$ and estimate $Q(\bar{u}_n)$, as defined in (2.3.12). Let $\bar{u}_n \in S_n$ be defined by

$$\bar{u}_n'(x) = (-1)^j(j - \tfrac{1}{2})/n, \quad x_{j-1} < x < x_j,\ j = 1, 2, \ldots, n\ ; \tag{2.3.20}$$

the graph of $\bar{u}_n'$ is shown in Fig. 1.

Note that

$$|\bar{u}_n|_X \geq C > 0\,. \tag{2.3.21}$$

Now let $\psi_0, \ldots, \psi_{n-1} \in V_{1,n}$ denote the basis functions defined by

$$\psi_i(x) = 0 \quad \text{for } |x - x_i| > 1/n, \quad i = 0, \ldots, n-1\,,$$

$$\psi_0(x) = 1 - nx, \quad 0 < x < x_1\,,$$

$$\psi_i(x) = \begin{cases} nx - i + 1, & x_{i-1} < x < x_i\,, \\ -nx + i + 1\,, & x_i < x < x_{i+1}, \end{cases} \quad i = 1, \ldots, n-1\,.$$

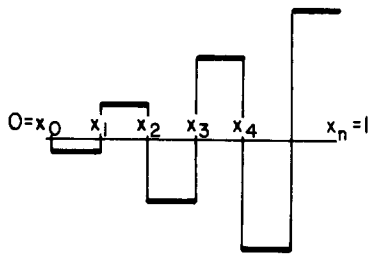

Fig. 1. Graph of the function $\bar{u}'$.

Then we have

$$B(\bar{u}_n, \psi_i) = \begin{cases} \frac{1}{2}(-1)^{i-1} n^{-2}, & i = 1, \ldots, n-1, \\ -\frac{1}{4} n^{-2}, & i = 0. \end{cases} \tag{2.3.22}$$

Any $\psi \in V_{1,n}$ can be written in the form

$$\psi = \sum_{i=0}^{n-1} c_i \psi_i$$

and we easily see that

$$\frac{1}{3n}\left[\tfrac{1}{2}c_0^2 + c_1^2 + \cdots + c_{n-1}^2\right] \leq |\psi|_{\frac{2}{3}}^2 \leq \frac{1}{n}\left[\tfrac{1}{2}c_0^2 + c_1^2 + \cdots + c_{n-1}^2\right].$$

Therefore,

$$|B(\bar{u}_n, \psi)| = \left| \sum_{i=0}^{n-1} c_i B(\bar{u}_n, \psi_i) \right| = \left| \left[\sum_{i=1}^{n-1} \tfrac{1}{2} c_i (-1)^{i-1} - \tfrac{1}{4} c_0 \right] n^{-2} \right|$$
$$\leq n^{-1} |\psi|_1 .$$

Thus

$$Q(\bar{u}_n) \leq C n^{-1}. \tag{2.3.23}$$

Since (2.3.10) holds with $\omega = 1/\sqrt{2}$, we see from Theorem 2.10 that $D \geq Cn$. This proves the left-hand side of (2.3.19).

Before turning to the proof of the right-hand side we will construct a particular $w \in R(\bar{u}_n)$ which will play the role of u in Theorem 2.10. We begin by constructing $\xi_0(x), \ldots, \xi_{n-1}(x)$ so that

$$\xi_i(x) = 0 \quad \text{for } |x - x_i| > 1/n,$$

$$\int_{x_{i-1}}^{x_{i+1}} \xi_i \psi_j \, dx = \frac{(-1)^{i-1}}{n} \delta_{ij}, \quad i, j = 0, \ldots, n-1, \text{ not both } 0, \tag{2.3.24}$$

$$\int_0^{x_1} \xi_0 \psi_0 \, dx = \frac{-1}{2n},$$

where

$$\delta_{ij} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}.$$

We will take the ξ_i to be linear on each subinterval, so (2.3.24) determines them uniquely. Specifically we choose $\xi_i(-1)^{i-1}$ as shown in Fig. 2.

Then we define

$$\rho(x) = \frac{1}{2n} \sum_{i=0}^{n-1} \xi_i(x) \tag{2.3.25}$$

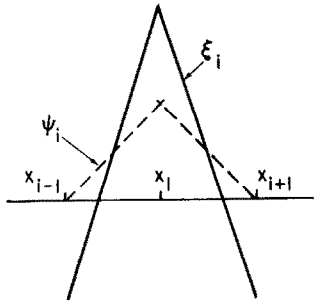

Fig. 2. The graph of the function ξ_i.

and

$$w(x) = \int_0^x \rho(t)\mathrm{d}t. \tag{2.3.26}$$

It is easy to see that

$$|w|_X \leq C/n. \tag{2.3.27}$$

From (2.3.22), (2.3.24) and the definition of w we have

$$B(\bar{u}_n, \psi_i) = \int_0^1 \bar{u}_n' \psi_i \mathrm{d}x = \int_0^1 w' \psi_i \mathrm{d}x = B(w, \psi_i), \quad i = 0, \ldots, n-1,$$

which shows that $P_n w = \bar{u}_n$. Using (2.3.20) and (2.3.26) we thus have

$$\frac{|P_n(w)|_X}{|w|_X} = \frac{|\bar{u}_n|_X}{|w|_X} \geq Cn,$$

which was our goal.

Let us remark that $\bar{u}_n$ is not unique. It is essential, however, that it is what we referred to earlier as a wild or irregular function. It is also of interest to note that the function $\tilde{u}_n \in S_n$ defined by

$$\tilde{u}_n'(x) = (-1)^j, \quad x_{j-1} < x < x_j,$$

will not lead to the desired result since we could only prove $Q(\tilde{u}_n) \geq C\sqrt{n}$.

We will now prove the right-hand side of (2.3.19) by applying Theorem 2.8. Let $u \in S_n$ and set

$$w_i = u'(x_i - 1/2n), \quad i = 1, 2, \ldots, n.$$

Then select $v \in V_{1,n}$ so that

$$V(x_i) = v_i,$$

where

$$v_i = w_i + w_{i+1}, \quad i = 1, \ldots, n-1,$$

$$v_0 = w_1, \qquad v_n = 0.$$

We easily see that

$$B(u, v) = \frac{1}{2n}(w_1, \ldots, w_n)\begin{bmatrix} 2 & 1 & & & \\ 1 & 2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 2 & 1 \\ & & & 1 & 1 \end{bmatrix}\begin{bmatrix} w_1 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{bmatrix}.$$

Consider the auxiliary $(2n-1)\times(2n-1)$ matrix

$$\hat{A} = \begin{bmatrix} 2 & 1 & & & \\ 1 & 2 & 1 & & \\ & \ddots & \ddots & \ddots & 1 \\ & & & 2 & 1 \\ & & & 1 & 2 \end{bmatrix}.$$

and the associated quadratic form

$$\Phi(z) = (z_1, \ldots, z_{2n-1})\hat{A}(z_1, \ldots, z_{2n-1})^t.$$

The eigenfunctions are $\sin \pi jk/(2n)$, $k = 1, \ldots, 2n-1$, and thus it is easily seen that the least eigenvalue $\lambda_1(n)$ of $\hat{A}$ satisfies

$$\lambda_1(n) \geq Cn^{-2}.$$

Thus

$$\Phi(z) \geq Cn^{-2} \sum_{i=1}^{2n-1} z_i^2. \tag{2.3.28}$$

Setting

$$\hat{z} = (w_1, \ldots, w_{n-1}, w_n, w_{n-1}, \ldots, w_1)^t,$$

we see that

$$B(u, v) = \frac{1}{4n}\Phi(\hat{z})$$

and therefore from (2.3.28) we obtain

$$\begin{aligned} B(u, v) &\geq Cn^{-3}(2w_1^2 + \cdots + 2w_{n-1}^2 + w_n^2) \\ &\geq Cn^{-3} \sum_{i=1}^{n} w_i^2 \geq Cn^{-2}|u|_X^2. \end{aligned} \tag{2.3.29}$$

We also see that

$$\begin{aligned} |v|_{\mathcal{V}}^2 &\leq Cn^{-1} \sum_{i=0}^{n} v_i^2 \\ &= Cn^{-1}[w_1^2 + (w_1 + w_2)^2 + \cdots + (w_{n-1} + w_n)^2] = CB(u, v)\,. \end{aligned} \tag{2.3.30}$$

Combining (2.3.29) and (2.3.30) we obtain

$$\frac{|B(u, v)|}{|v|_{\mathcal{V}}} \geq C|B(u, v)|^{1/2} \geq Cn^{-1}|u|_X\,,$$

from which we get

$$\gamma(S_n, V_{1,n}) = \inf_{\substack{u \in S_n \\ |u|_X = 1}} \sup_{\substack{v \in V_{1,n} \\ |v|_{\mathcal{V}} = 1}} |B(u, v)| \geq Cn^{-1}\,.$$

Since

$$|B(u, v)| \leq |u|_X |v|_{\mathcal{V}} \quad \text{for } u \in \mathcal{H},\ v \in \mathcal{V}\,,$$

thus from Theorem 2.8 we have

$$D(u, M_n, X) \leq Cn, \quad \text{for } u \in \mathcal{H}\,.$$

This proves the right-hand side of (2.3.19).

Next we show that, although $D(\mathcal{H}, M_n, X)$ is large, the method still works well for smooth solutions. Specifically, we will prove that

$$|u - u_n|_X \leq Ch|u|_{C^3(\bar{I})}\,, \tag{2.3.31}$$

for $u \in C^3(\bar{I})$, where $u_n = P(S_n, V_{1,n})u$ and $h = 1/n$. Let $u_i = u_n(x_i)$. It is easy to see that the finite element equations reduce to

$$\begin{aligned} &u_0 = 0, \qquad \tfrac{1}{2}(u_1 - u_0) = \int_{x_0}^{x_1} f\psi_0 \,\mathrm{d}x\,, \\ &\tfrac{1}{2}(u_{i+1} - u_{i-1}) = \int_{x_{i-1}}^{x_{i+1}} f\psi_i \,\mathrm{d}x, \quad i = 1, \ldots, n-1 \end{aligned} \tag{2.3.32}$$

where $\psi_0, \ldots, \psi_{n-1}$ are the basis functions for $V_{1,n}$ introduced earlier.

Since $u \in C^3(\bar{I})$, $f \in C^2(\bar{I})$ and we can write

$$\begin{aligned} &2\int_{x_0}^{x_1} f\psi_0 \mathrm{d}x - \int_{x_0}^{x_1} f \,\mathrm{d}x = \alpha_0 h^2\,, \\ &2\int_{x_{i-1}}^{x_{i+1}} f\psi_i \mathrm{d}x - \int_{x_{i-1}}^{x_{i+1}} f \,\mathrm{d}x = \alpha_i h^3, \quad i = 1, \ldots, n-1\,, \end{aligned} \tag{2.3.33}$$

where

$$|\alpha_i| \leq C|f|_{C^2(\bar{I})}. \tag{2.3.34}$$

Now, using (2.3.32),

$$[u_{i+1} - u(x_{i+1})] - [u_{i-1} - u(x_{i-1})] = 2\int_{x_{i-1}}^{x_{i+1}} f\psi_i \mathrm{d}x - \int_{x_{i-1}}^{x_{i+1}} f\,\mathrm{d}x, \quad i = 1, \ldots, n-1$$

and so, setting $\xi_i = u_i - u(x_i)$, (2.3.32) implies

$$\begin{aligned} &\xi_{i+1} - \xi_{i-1} = \alpha_i h^3, \quad i = 1, \ldots, n-1 \\ &\xi_0 = 0. \end{aligned} \tag{2.3.35}$$

From (2.3.34) and (2.3.35) we obtain

$$|\xi_j| \leq C|f|_{C^2(\bar{I})} h^2, \quad j \text{ even}$$

and

$$|\xi_j| \leq C|f|_{C^2(\bar{I})} h^2 + |\xi_1|, \quad j \text{ odd}.$$

Using (2.3.32) and (2.3.33) again we have

$$\xi_1 = u_1 - u(x_1) = 2\int_0^{x_1} f\psi_0 \mathrm{d}x - \int_0^{x_1} f\,\mathrm{d}x = \alpha_0 h^2.$$

Thus,

$$|\xi_j| \leq C|f|_{C^2(\bar{I})} h^2, \quad j = 0, 1, \ldots, n,$$

with C independent of n.

Letting $\mathcal{I}_n u$ denote the S_n-interpolant of u, we have

$$\begin{aligned} |u - u_n|_X &\leq |u - \mathcal{I}_n u|_X + |\mathcal{I}_n u - u_n|_X \\ &\leq Ch|u|_{C^3(\bar{I})} + |\mathcal{I}_n u - u_n|_X. \end{aligned} \tag{2.3.36}$$

We further see that

$$|\mathcal{I}_n u(x_i) - u_n(x_i)| = |\xi_i| \leq Ch^2|u|_{C^3(\bar{I})}.$$

From this we get

$$|\mathcal{I}_n u - u_n|_{L_2} \leq Ch^2|u|_{C^3(\bar{I})},$$

and thus, using the fact that $|s|_X \leq Cn|s|_{L_2}$ for $s \in S_n$ (the so-called inverse assumptions for S_n), we have

$$|\mathcal{I}_n u - u_n|_X \leq Ch|u|_{C^3(\bar{I})}. \tag{2.3.37}$$

Combining (2.3.36) and (2.3.37) we have

$$|u - u_n|_X \leq Ch|u|_{C^3(\bar{I})}.$$

This is the desired result.

Define now

$$H = \{u: |u|_{C^3(\bar{I})} < \infty,\ u(0) = 0,\ u''(x) \geq \alpha_1 |u|_{C^3(\bar{I})} \text{ or } u''(x) \leq -\alpha_1 |u|_{C^3(\bar{I})} \text{ on } (\omega, \eta) \text{ with } |\omega - \eta| = \alpha_2\},$$
$$\alpha_1, \alpha_2 > 0.$$

Consider now the method using $(S_n, V_{1,n})$. Then we easily see that

$$C(H, M_n, H^1) \leq K \tag{2.3.38}$$

where K depends on α_1, α_2 but is independent of n. In fact it is easy to see that for $u \in H$,

$$Z(u, S_n, H^1) \geq Ch\alpha_1\alpha_2 |u|_{C^3(\bar{I})} / |u|_{H^1}$$

and (2.3.38) is obtained by combining this estimate with (2.3.31).

Let us summarize some main points shown in case (b).

(1) The method is not quasi-optimal on $\mathscr{H} = {}^{\circ}H^1$; nevertheless, the instability is not too strong. The analysis has been based on results relating to Theorem 2.8.

(2) The method is quasi-optimal on the set H, i.e., all functions belonging to H are K-perfect. We note that H is not closed in ${}^{\circ}H^1$.

(3) The result that H is a set of perfect solutions cannot be directly proven by Theorems 2.7, 2.8, 2.10 and 2.11. A special analysis is necessary.

3. Further examples of finite element methods

3.1. Introduction

In this section we examine three different finite element methods for the approximate solution of a simple model problem, namely that of a longitudinally loaded bar on an elastic support. The classical displacement formulation of this problem is

$$\begin{aligned} &Lu \equiv -(E(x)F(x)u'(x))' + b(x)u(x) = p(x), \quad 0 < x < 1, \\ &u(0) = u(1) = 0. \end{aligned} \tag{3.1.1}$$

Here $u(x)$, $0 < x < 1$, denotes the longitudinal displacement and $E(x)$ denotes the modulus of elasticity, $F(x)$ the cross-sectional area, $b(x)$ the spring constant of the elastic support and $p(x)$ the longitudinal load. We will let $a(x) = E(x)F(x)$ and assume

$$0 < \beta_1 \leq a(x) \leq \beta_2, \qquad 0 \leq b(x) \leq \beta_2, \tag{3.1.2}$$

but otherwise allow a and b to be rather general functions. They could be, for example, constant functions (corresponding to a bar with uniform cross-section and elastic properties and a uniform elastic support) or step functions with many steps (arising in the study of composite materials, for example).

We now cast this problem in variational form. Let

$$B(u, v) = \int_0^1 (au'v' + buv)\,dx. \tag{3.1.3}$$

Then we easily see that (3.1.1) can be formulated as

$$\begin{aligned} &u \in \mathring{H}^1 \equiv \{u\colon u \in H^1,\ u(0) = u(1) = 0\}, \\ &B(u, v) = \int_0^1 pv\,dx \quad \text{for all } v \in \mathring{H}^1. \end{aligned} \tag{3.1.4}$$

B is defined in $\mathscr{H} \times \mathscr{V}$, where $\mathscr{H} = \mathscr{V} = \mathring{H}^1$.

To complete the specification of a directed variational method, in addition to the bilinear form we must select a family $\mathscr{F}$ of trial and test spaces S and V, the space X and the discretization parameter α. We will present three choices for S and V in Sections 3.2 and 3.3.

3.2. *The standard finite element method*

Let $\Delta = \{0 = x_0 < x_1 < \cdots < x_{n(\Delta)} = 1\}$, $n(\Delta) \geq 2$, be an arbitrary mesh on $[0, 1]$ and set $I_j = I_j(\Delta) = (x_{j-1}, x_j)$, $h_j = h_j(\Delta) = x_j - x_{j-1}$ and $h = h(\Delta) = \max_{0 \leq j \leq n(\Delta)} h_j(\Delta)$. Then set

$$\begin{aligned} S = S_\Delta = V = V_\Delta \\ \equiv \{s(x)\colon s(x) \text{ is continuous on } [0, 1],\ s(x) \text{ is linear on each } I_j \text{ and } s(0) = s(1) = 0\}. \end{aligned} \tag{3.2.1}$$

Let $\mathscr{H} = \mathscr{V} = \mathring{H}^1$, $\mathscr{F} = \{(S_\Delta, V_\Delta)\colon \Delta \text{ any mesh}\}$, $X = H^1$ and $\alpha(S_\Delta, V_\Delta) = h(\Delta)$. We can now discuss the directed variational method

$$\mathscr{M} = (\mathscr{H}, H^1, \mathscr{V}, \mathscr{F}, B, \alpha).$$

The optimality of this method is given in the following theorem.

THEOREM 3.1. The method $\mathscr{M}$ is quasi-optimal on $\mathring{H}^1$ with respect to H^1, i.e.,

$$C(\mathring{H}^1, \mathscr{M}, H^1) \leq C, \tag{3.2.2}$$

with C depending only on β_1 and β_2 but independent of Δ. Thus,

$$|u - P(S_\Delta, V_\Delta)u|_{H^1} \leq C \inf_{v \in S_\Delta} |u - v|_{H^1}. \tag{3.2.3}$$

PROOF. This standard result is proved by showing (2.3.18) and using Theorems 2.8 and 2.6. In fact, for all $u \in \mathring{H}^1$ we have

$$B(u, u) = \int_0^1 ((au')^2 + bu^2)\mathrm{d}x$$

$$\geqslant \beta_1 \int_0^1 (u')^2 \mathrm{d}x \geqslant \tfrac{1}{2}\beta_1 |u|_{H^1}^2 .$$

As stated in Theorem 3.1, the optimality constant $C(\mathring{H}^1, \mathcal{M}, H^1)$ is finite and is bounded uniformly with respect to the class of coefficients a and b satisfying (3.1.2). Regarding approximability let us state the standard result for the elements under consideration:

$$\inf_{s \in S_\Delta} |u - s|_{H^1} \leqslant Ch|u|_{H^2}, \tag{3.2.4}$$

where C is independent of Δ. Thus for any $u \in \mathring{H}^1 \cap H^2$,

$$Z(u, S_\Delta, H^1) \leqslant Ch|u|_{H^2}/|u|_{H^1} . \tag{3.2.5}$$

If our solution u lies in H^2, then from (3.2.3) and (3.2.4) we have the error estimate

$$|u - P(S_\Delta, V_\Delta)u|_{H^1} \leqslant Ch|u|_{H^2} .$$

We note that if $a'(x)$ and $b(x)$ are bounded, then we can prove that

$$|u|_{H^2} \leqslant C|p|_{L_2} .$$

From this and (3.2.5) we obtain

$$|u - P(S_\Delta, V_\Delta)u|_{H^1} \leqslant Ch|p|_{L_2} . \tag{3.2.6}$$

C here depends on the maximum of a' and b; (3.2.6) is not, however, valid for problems with rough coefficients (coefficients which are step functions, for example).

Next, let us consider the same family of simple methods but choose $X = L_2$. We present a result showing the resulting method is not quasi-optimal.

THEOREM 3.2. Suppose $a(x) = 1$ *and* $b(x) = 0$. *Then for any constant* $C > 0$ *and any* Δ *there is a* $u \in \mathring{H}^1$ *such that*

$$|u - P(S_\Delta, V_\Delta)|_{L_2} > C \inf_{s \in S_\Delta} |u - s|_{L_2} .$$

Thus,

$$C(\mathring{H}^1, M(S_\Delta, V_\Delta), L_2) = +\infty \quad \textit{for any } \Delta .$$

PROOF. Suppose $\Delta = \{0, \frac{1}{2}, 1\}$ and let u be as shown in Fig. 3 with $u(\frac{1}{2}) = 1$. It is easily shown that, since $P(S_\Delta, V_\Delta)u$ is the S_Δ-interpolant of u, as shown in Fig. 3 by the dashed curve, $|u - P(S_\Delta, V_\Delta)u|_{L_2}$ is nearly as large as $|P(S_\Delta, V_\Delta)u|_{L_2} = 1/\sqrt{3}$ and $\inf_{s \in S_\Delta} |u - s|_{L_2}$, which is $\leq |u|_{L_2}$, is nearly as small as zero. This completes the proof for this simple mesh. The proof for a general mesh is similar.

It is interesting to compare Theorems 3.1 and 3.2 with Example 2.5. In that example we treated the same boundary value problems but based our approximation on polynomials instead of piecewise linear functions. In Example 2.5 we obtained $C(\mathring{H}^1, M(S_n, V_n), L_2) \leq C\sqrt{n}$ in contrast to the result $C(\mathring{H}^1, M(S_\Delta, V_\Delta), L_2) = +\infty$ obtained above.

THEOREM 3.3. Suppose $a'(x)$ and $b(x)$ are bounded. Then for any $u \in H^2$ we have

$$|u - P(S_\Delta, V_\Delta)u|_{L_2} \leq Ch^2 |u|_{H^2} \leq Ch^2 |P|_{L_2},$$

where C is independent of u and Δ but depends on a and b. If

$$H = \{u : u \in H^2 \cap \mathring{H}^1,\ u''(x) \geq \alpha_1 |u|_{H^2} \text{ or } u''(x) \leq -\alpha_1 |u|_{H^2} \text{ on } (\omega, \eta) \text{ where } |\omega - \eta| = \alpha_2\}$$

with $\alpha_1, \alpha_2 > 0$, then any $u \in H$ is K-perfect with K depending on α_1, α_2, a and b. Thus,

$$C(H, \mathcal{M}, L_2) \leq C.$$

PROOF. The first part of the result is standard (see e.g. [2]) and the second part follows from the fact that

$$Z(u, S_\Delta, L_2) \geq \frac{Ch^2 \alpha_1 \alpha_2 |u|_{H^2}}{|u|_{L_2}}.$$

We consider now one more choice for X, namely $X = L_\infty$. Then the following result is proved in [3].

THEOREM 3.4. Suppose that a, b are sufficiently smooth. Then there is a constant C such that for any $u \in L_\infty$,

$$|u - P(S_\Delta, V_\Delta)u|_{L_\infty} \leq C \inf_{s \in S_\Delta} |u - s|_{L_\infty},$$

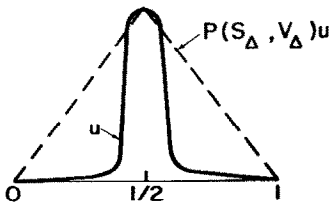

Fig. 3. The graph of function u.

with C independent of u and Δ but depending on β_1, β_2 and the maximum of the first derivatives of a and b. Thus,

$$C(\mathring{H}^1, \mathcal{M}, L_\infty) \leq C.$$

It is interesting to compare Theorems 3.1, 3.2 and 3.4. The computationally identical method is quasi-optimal on $\mathring{H}^1$ with respect to H^1 and L_∞ but is not with respect to L_2.

3.3. *A second method for solving* (3.1.4)

In this subsection we suppose $b(x) = 0$. Here we choose

$$\begin{aligned}\hat{S} = \hat{S}_\Delta = \hat{V} = \hat{V}_\Delta \\ = \{s(x)&: s \text{ is continuous on } [0,1],\ s(x) \text{ is a solution of } (as')' = 0 \text{ on each } I_j, \\ &s(0) = s(1) = 0\}.\end{aligned} \tag{3.3.1}$$

We again let $\mathcal{H} = \mathcal{V} = \mathring{H}^1$, $X = H^1$, $\hat{\mathcal{F}} = \{(\hat{S}_\Delta, \hat{V}_\Delta)$: Δ any mesh$\}$ and $\alpha(S_\Delta, V_\Delta) = h(\Delta)$. We will discuss the directed variational method

$$\mathcal{M} = (\mathring{H}^1, H^1, \mathring{H}^1, \hat{\mathcal{F}}, B, \alpha).$$

Exactly as in Theorem 3.1 we see that this method is quasi-optimal on $\mathcal{H}$ with respect to H^1. The methods differ, however, in regard to the approximation properties of the trial spaces employed. The approximation properties of the trial space introduced in (3.3.1) is given in the following theorem, proved in [4].

THEOREM 3.5. There is a constant C depending only on β_1 and β_2 such that

$$\inf_{s \in S_\Delta} |u - s|_{H^1} \leq Ch|Lu|_{L_2} = Ch|p|_{L_2}. \tag{3.3.2}$$

PROOF. See [4].

Combining this result with the above mentioned quasi-optimality we obtain

$$|u - P(\hat{S}_\Delta, \hat{V}_\Delta)u|_{H^1} \leq Ch|p|_{L_2}. \tag{3.3.3}$$

This estimate should be compared with (3.2.6). They differ in that while (3.3.3) holds for all $a(x)$ and $b(x)$ satisfying (3.1.2), (3.2.6) holds only for smooth a and b. Thus we see that the method $\mathcal{M}$ considered in this section is very accurate. Because of the unusual test and trial space it is, however, not easily implemented. The method introduced in Section 3.4 will be as accurate as the method discussed here while being as easily implemented as the standard finite element method discussed in Section 3.2.

3.4. A third method for solving (3.1.4)

As underlined in Section 2.2, the trial and test spaces play very different roles. The trial space is chosen for its approximation properties and the test space is chosen so that the method is optimal or nearly optimal and so that the method is easily implemented. The standard method, which uses piecewise linear functions for both trial and test spaces is optimal and is easily implemented. The trial space has poor approximation properties for problems with rough coefficients, however. The method discussed in Section 3.3 is optimal and has good approximation properties while not being easily implemented because of the way the load enters in the computation. The method discussed in this section will use the trial space $\hat{S}_\Delta$ used in the second method and use the test space V_Δ used in the standard method. This choice will simplify the implementation, making it virtually of the same complexity as the standard method, while preserving the advantages of the second method.

Suppose as above that $b(x) = 0$. Let $\hat{S}_\Delta$ be as defined in (3.3.1) and let V_Δ be as defined in (3.2.1). We let $\mathscr{H} = \overset{\circ}{H}{}^1$, $X = H^1$, $\mathscr{V} = \overset{\circ}{H}{}^1$, $\mathscr{F} = \{(\hat{S}_\Delta, V_\Delta)\}$ and $\alpha(\hat{S}_\Delta, V_\Delta) = h(\Delta)$, and consider the directed variational method $\mathscr{M} = (\mathscr{H}, H^1, \mathscr{V}, \mathscr{F}, B, \alpha)$.

THEOREM 3.6. There is a positive constant γ depending only on β_1 and β_2 such that

$$\inf_{\substack{s\in\hat{S}_\Delta \\ |S|_{H^1}=1}} \sup_{\substack{v\in V_\Delta \\ |V|_{H^1}=1}} |B(S, V)| = \gamma(S_\Delta, V_\Delta) \geqslant \gamma \quad \text{for all } \Delta . \tag{3.4.1}$$

PROOF. For $s \in \hat{S}_\Delta$ let $\bar{s} \in V_\Delta$ be defined by

$$\bar{s}(x_j) = s(x_j), \quad j = 0, 1, \ldots, n .$$

Set

$$\hat{a}_j = \left(\int_{I_j} a^{-1}\mathrm{d}x\right)^{-1} h_j . \tag{3.4.2}$$

Then we easily see that

$$as'|_{I_j} = \hat{a}_j \bar{s}'|_{I_j} .$$

Using this relation we have

$$\begin{aligned} B(s, \bar{s}) &= \int_0^1 as'\bar{s}'\,\mathrm{d}x = \sum_{j=1}^{n} \int_{I_j} \frac{(as')^2}{\hat{a}_j}\,\mathrm{d}x \\ &\geqslant \frac{\beta_1^2}{\beta_2}\int_0^1 (s')^2\mathrm{d}x \geqslant \tfrac{1}{2}\frac{\beta_1^2}{\beta_2}|s|_{H^1}^2 \end{aligned}$$

and

$$\begin{aligned} |\bar{s}|_{H^1}^2 &\leqslant 2\int_0^1 (\bar{s}')^2\mathrm{d}x = 2\sum_j \int_{I_j} (as'/a_j)^2\mathrm{d}x \\ &\leqslant (2\beta_2^2/\beta_1^2)|s|_{H^1}^2 . \end{aligned}$$

Hence

$$\gamma(S_\Delta, V_\Delta) \geq \beta_1^3/2\sqrt{2}\beta_2^2 \equiv \gamma .$$

Combining Theorem 3.6 with Theorems 2.6 and 2.8, we see that the optimality constant $C(\mathring{H}^1, \mathcal{M}, H^1)$ is bounded uniformly over an entire class of coefficients satisfying (3.1.2). In addition, from Theorem 3.5 we see that we have good approximation properties. Thus,

$$|u - P(\hat{S}_\Delta, V_\Delta)u|_{H^1} \leq \bar{C}h|Lu|_{L_2} \leq \bar{C}h|p|_{L_2} , \tag{3.4.3}$$

with $\bar{C}$ depending only on β_1 and β_2. Although estimates (3.2.6) and (3.4.3) are similar, the values of the constants C in (3.2.6) and C in (3.4.3) could be very different if $a(x)$ is not smooth ($\bar{C} \ll C$).

One further point needs to be considered. We are using the bilinear form $B(s, v) = \int_0^1 as'v'\,dx$, as in Section 3.2, but we are now using different trial and test spaces and the trial space consists of less simple elements. However, it is easily seen that the stiffness matrix is symmetric and is as easily computed as the stiffness matrix for the standard method.

THEOREM 3.7. Let $\bar{\varphi}_1, \ldots, \bar{\varphi}_{n-1}$ be the standard basis for S_Δ and V_Δ and $\varphi_1, \ldots, \varphi_{n-1}$ be the standard bases for $\hat{S}[\varphi_i(x_j) = \delta_{ij}]$. Then

$$B(\varphi_j, \bar{\varphi}_i) = \int_0^1 a\varphi_j'\bar{\varphi}_i'\,dx = \int \hat{a}_j\bar{\varphi}_j'\bar{\varphi}_i'\,dx . \tag{3.4.4}$$

PROOF. Again we use the relation

$$as'|_{I_j} = \hat{a}_j\bar{s}'|_{I_j}$$

for any $s \in S_\Delta$, (3.4.4) follows immediately from this.

It should be noted that the stiffness matrix for the standard finite element method discussed in Section 3.2 is

$$\int_0^1 a_j\bar{\varphi}_j'\varphi_i'\,dx \tag{3.4.5}$$

where

$$a_j = \left(\int_{I_j} a\,dx\right)\Big/ h_j . \tag{3.4.6}$$

Since the load vector (the right-hand side in the discretized problem) is the same in the standard method as in the method discussed in this section, we see that the methods differ only in the occurrence of $\hat{a}_j$ or a_j, defined in (3.4.2) and (3.4.6), respectively, in the stiffness matrices, defined in (3.4.4) and (3.4.5), respectively. Obviously $\hat{a}_j$ and a_j are very close when $a(x)$ is nearly constant but they can be very different when $a(x)$ is rapidly changing. This shows that the results obtained by the standard method and the third method could be very

different, the third method performing equally well for problems with smooth coefficients and strikingly better for problems with rough coefficients. The analysis here has been for the H^1 norm. We remark that similar results hold for the L_2-norm. Let us also remark that the third method can be extended to equations with $b \not\equiv 0$ so as to preserve its good properties. For further results see [4].

4. Conclusions regarding the selection of finite element methods

4.1. Introduction

As we have stated in Section 1 and have shown in the previous sections, there are many finite element methods which can be considered for any specific problem. In this section we discuss the application of the ideas elaborated on above to the rational selection or design of effective methods.

4.2. Definitions of the various types of variational methods

It is essential to clarify as much as possible the aim of an engineering computation, the set of possible solutions, the environment in which the computations are to be made, and the various types of computational procedures actually in use in computational engineering. Toward this end we have introduced the notions of simple variational, variational, directed variational and computational variational methods. Nearly all of the computational procedures used in practice fall within this framework. We mention here the h-version and the p-version of the finite element method, displacement and mixed methods, various adaptive approaches, etc. The examples discussed in Sections 2 and 3 show that the same computational procedure can be viewed as different directed variational or computational variational methods, the difference being related to the use of different norms with which to measure the error. We have seen that the different methods have significantly different behavior. It appears that without precise definitions of the various types of variational methods, a careful discussion of finite element methods leading to a rational choice of a method for a specific problem would not be possible.

4.3. Approximability and optimality

The notions of approximability and optimality are two central ideas to be considered in the comparison of finite element methods. Clearly the trial space should be tailored as well as possible to the class of possible solutions. The approximability constant $Z(u, S, X)$ is a measure of how well this has been done. The trial spaces introduced in Sections 3.3 and 3.4 are obviously preferable to the standard polynomial spaces.

The test space should be selected so as to lead to a small optimality constant $C(u, M, X)$ and so as to lead to computational simplicity. The optimality constant is a measure of how well the approximate solution performs in comparison with the best possible approximation. In Example 2.5 with $X = X_2 = L_2$ we have seen that for the solution $\bar{u}$, $C(\bar{u}, M_n, X_2) \sim \sqrt{n}$ and $Z(\bar{u}, S_n, X_2) \sim 1/n$. Thus the error satisfies $|\bar{u} - u_n|_{X_2} \sim 1/\sqrt{n}$, i.e., the good approximability properties of S_n are partially eroded because of the large optimality constant. We have also seen that $\bar{u}$ leads to the largest optimality constant. The method does, however, converge.

The second and third methods treated in Section 3 both have small optimality constants, but the third method is much easier to implement than the second.

The selection of the test spaces heavily influences the set of solutions which are perfect, i.e., the set of solutions for which the method works well, assuming good approximability. Sometimes all 'reasonable' solutions are perfect, while the imperfect solutions are 'wild'. Example 2.14 is an illustration of this. In other problems reasonable solutions could be very imperfect. Example 2.4 is an illustration of this. That example also clearly shows the effect of the choice of the test space on the class of perfect solutions. In any case, the method which is perfect for the largest class of solutions is preferable.

If a large class of solutions are imperfect and the optimality constant $\to +\infty$ as $n \to +\infty$ (n being the number of degrees of freedom, say), then a method with a smaller C is preferable to one with a large C. The standard finite element method (cf. the discussion in Section 3.2 in the constant coefficient case) and the method discussed in Example 2.5, when considered in connection with $X = L_2$, can be compared in this way. In the first case we get $C(\mathring{H}^1, M_\Delta, L_2) = \infty$ while in the second case $C(L_2, M_n, L_2) \sim \sqrt{n}$.

In many situations, rigorous estimates of the optimality constant are not available and judgments concerning choice of methods must be based on computational experience. In these situations, one must attempt to gain insight on the class of perfect solutions from the computational experience. Nevertheless, we have to be aware that there is a possibility that only perfect solutions will be tested and the conclusions could be misleading.

Approximability and optimality together influence the performance of the method. A method could have a deteriorating optimality constant and still give reasonably good accuracy if the approximability is sufficiently good to offset the lack of optimality. Thus we are interested in the set of solutions for which the method converges as well as the set of perfect solutions.

Examples 2.4 and 2.14 clearly illustrate the influence of the choice of trial and test space on optimality. If the 'major part' of the bilinear form B is symmetric (cf. Example 2.4 for $|\lambda| > 1$), then using the same space for the trial and test space is often advisable. If the 'major part' of B is nonsymmetric (cf. Example 2.4 for $|\lambda| < 1$ and Example 2.14), then it is often useful to consider different trial and test spaces.

4.4. *Stability and the stability condition*

The stability constant should be viewed as a tool to be used in the analysis of optimality. $C(u, M, X)$ and $D(u, M, X)$, and also $C(H, M, X)$ and $D(H, M, X)$, are closely related, as shown in Theorem 2.6. The constant $\gamma(S, V)$ is in turn used to estimate the stability constant, as shown in Theorem 2.8. We note that $\gamma(S, V)$ does not depend on the exact solution u and thus that an analysis of stability or optimality based on an estimate for $\gamma(S, V)$ must concentrate on 'worst' possible cases. The stability condition (2.3.17) is necessary in the sense that if it is violated, then the method must diverge for at least one solution, as shown by Theorems 2.7 and 2.11. Nevertheless, a method can violate this condition and still perform well for a large class of solutions.

Recently there have been statements in the literature to the effect that certain methods perform well computationally even though the stability condition (also sometimes called the inf–sup, LBB, or BB conditions) is violated. This occurrence can be explained by noting that in the computations only perfect solutions were considered. These particular solutions may

have been considered partly on the basis of physical insight, but partly by accident. Thus we see that a detailed analysis of the structure of the set of perfect solutions is highly desirable. Certainly formal insistance that the stability condition is satisfied is inappropriate.

We see from Theorem 2.6, that the method discussed in Example 2.5, when considered in connection with the L_2-norm (case (b)), does not satisfy the stability condition, since the optimality constant $\to \infty$ as $n \to \infty$. Still, on the basis of experimental evidence, it is likely that the method would perform well because only 'reasonable' solutions would be considered. A similar situation occurs when we consider the performance of the standard finite element method in the L_2-norm or when considering the second choice of spaces in Example 2.14. In contrast, it is very likely that experimentally one would generally see that the method discussed in the Example 2.4 has a serious deficiency when $|\lambda| < 1$.

References

[1] I. Babuška, Error bounds for finite element method, Numer. Math. 16 (1971) 322–332.
[2] I. Babuška and A.K. Aziz, Survey lectures in the mathematical foundations of the finite element method, in: A.K. Aziz, ed., The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (Academic Press, New York, 1972).
[3] I. Babuška and J.E. Osborn, Analysis of finite element methods for second order boundary value problems using mesh dependent norms, Numer. Math. 34 (1980) 41–62.
[4] I. Babuška and J.E. Osborn, Generalized finite element methods: their performance and their relation to mixed methods, SIAM J. Numer. Anal. 20 (1983) 510–536.
[5] F. Brezzi, On the existence uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers, RAIRO 22 (8) (1974) 129–151.
[6] A.E. Taylor and D.C. Lay, Introduction to Functional Analysis (Wiley, New York, 2nd ed., 1980).