FINITE ELEMENT METHODS FOR NON-SELF-ADJOINT PROBLEMS

K. W. MORTON

NUMERICAL ANALYSIS REPORT 3/81

Lectures presented at the SRC Numerical Analysis Summer School and Workshop, University of Lancaster, 20th July - 20th August, 1981. Proceedings to be published by Springer-Verlag.

FINITE ELEMENT METHODS FOR NON-SELF-ADJOINT PROBLEMS

K. W. Morton

1. INTRODUCTION

Finite element methods now dominate the solution of those elliptic problems that are derivable from quadratic extremal principles. Their practical development has been carried out by engineers under strong guidance from physical principles and subsequently their mathematical structure and error analysis studied in detail by mathematicians - as general references see Zienkiewicz (1977), Babuška & Aziz (1972), Strang & Fix (1973), Oden & Reddy (1976), Ciarlet (1978). From this work it became clear that the methods could be very widely applied, using a variational formulation referred to by the two groups respectively as 'the method of weighted residuals' or 'the weak form of the equations'.

However the success of these methods has sprung very largely from their optimal approximation properties. These follow naturally from extremal principles but are much harder to achieve for general elliptic problems that are not derived in this way and which are therefore no longer self-adjoint. It is this question that we shall consider in this short course of lectures. It is one of the most important and active areas of current research in the development of finite element methods.

1.1 A self-adjoint example

Consider the classical Dirichlet problem for Poisson's equation on an open polygonal domain Ω of $\mathbb{R}^2$ with boundary Γ :

$$-\nabla^2 u = f \quad \text{on } \Omega \tag{1.1a}$$

$$u = 0 \quad \text{on } \Gamma. \tag{1.1b}$$

Then u is also the solution of the extremal problem:

$$\underset{v \in H_0^1(\Omega)}{\text{minimise}} \int_\Omega [\tfrac{1}{2}|\nabla v|^2 - fv]\,d\Omega \ , \tag{1.2}$$

where we denote by $H^m(\Omega)$ the usual Sobolev space of functions with square integrable m^{th} derivatives over Ω and by $H_0^m(\Omega)$ the closure in this space of the set of functions whose support is confined to the interior of Ω, i.e. which are zero on the boundary. (More generally we shall use the latter notation to denote functions which are zero on that part of the boundary where Dirichlet data is given).

Suppose now we triangulate Ω and denote by S_0^h the set of functions V which are continuous on Ω, linear in each triangle and zero on the boundary (in general that part with Dirichlet data): that is, we can write

$$V(\underline{x}) = \sum_{(j)} V_j \phi_j(\underline{x}), \tag{1.3}$$

where the summation is over all the interior vertices of the triangulation and $\{\phi_j\}$ are the basis functions which are pyramid-shaped: i.e., ϕ_j is piecewise linear, unity at node j and zero at all other nodes. Then the Ritz-Galerkin approximation U to u is given by

$$\underset{V\epsilon S_0^h}{\text{minimise}} \int_\Omega [\tfrac{1}{2}|\underline{\nabla}V|^2 - fV]d\Omega \quad : \tag{1.4}$$

that is, using the notation $(\cdot,\cdot)$ for the L_2 inner product over Ω of either vectors or scalars, we have the Galerkin equations

$$(\underline{\nabla}U, \underline{\nabla}\phi_i) = (f, \phi_i) \qquad \forall\ \phi_i \in S_0^h. \tag{1.5}$$

Similarly, from (1.2) u satisfies the weak form of (1.1) :

$$(\underline{\nabla}u, \underline{\nabla}w) = (f,w) \qquad \forall\ w \in H_0^1(\Omega). \tag{1.6}$$

Since $S_0^h \subset H_0^1(\Omega)$, i.e. we are using a conforming finite element approximation, we can take ϕ_i for v in (1.6) and subtracting (1.5) obtain

$$(\underline{\nabla}(u-U), \underline{\nabla}\phi_i) = 0 \qquad \forall\ \phi_i \in S_0^h. \tag{1.7}$$

It follows, after a little manipulation, that

$$||\underline{\nabla}(u-U)||^2 = \inf_{V\epsilon S_0^h} ||\underline{\nabla}(u-V)||^2 , \tag{1.8}$$

where $||\cdot||$ denotes the L_2 norm. This is the fundamental optimal approximation property of U.

In practical applications it is often the vector field $\underline{\nabla}u$ which is of most interest; and $\underline{\nabla}U$ is the least squares best fit to it from those piecewise constant approximations obtained by taking gradients of functions in S_0^h. By comparing with u^I, the piecewise linear interpolant of u, one readily finds that $\underline{\nabla}U$ is generally accurate to $O(h)$, where h is the maximal diameter of the triangles in the triangulation (and it is assumed that the latter satisfies a regularity condition such as all the angles are bounded from zero, or the weaker condition that they are bounded from π - see Strang & Fix, 1973, and Ciarlet, 1978.) On the other hand, from a well-known argument due to Aubin and Nitsche, for the error in U we have $||u-U|| = O(h^2)$. More generally, if S_0^h had contained all functions which were piecewise polynominal up to degree k in each triangle we would have $||u-U|| = O(h^{k+1})$ but $||\underline{\nabla}(u-U)|| = O(h^k)$.

However, a most important practical consideration is that superconvergence phenomena enable ∇u to be estimated with an $O(h^2)$ error. For bilinear elements on rectangles ∇U approximates ∇u to $O(h^2)$ at the centroid of each element, a fact that has long been exploited by engineers and recently established rigorously and extended to more general quadrilateral elements by Zlamal (1977) - see also Zlamal (1978), Lesaint & Zlamal (1979). Although for triangular elements this extra order of convergence does not generally occur at the centroids, it has long been believed (and supported by numerical evidence) that second order accuracy can be recovered from gradients of U along each triangle side though this has not yet been proved (see Strang & Fix (1973) p169).

Without such superconvergence, which stems from the optimal approximation property, finite element methods would hardly be competitive with traditional finite difference methods, where similar divided difference results hold - see for instance Thomée & Westergren (1968).

1.2 Diffusion-convection problems

In studying the effects of losing self-adjointness we shall concentrate on the important class of problems called diffusion-convection problems:

$$-\nabla.(a\nabla u - \underline{b}u) + cu = f \qquad \text{in } \Omega \tag{1.9a}$$

$$u = g \quad \text{on } \Gamma_D \ , \ \partial u/\partial n = 0 \quad \text{on } \Gamma_N, \tag{1.9b}$$

where $\Gamma = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$ and $\Gamma_D \neq \emptyset$. Here the positive scalar a can be regarded as an isotropic diffusion coefficient, the vector $\underline{b}$ a convective velocity, f a given source and c a depletion rate. Equation (1.9a) represents a conservation law for the quantity u which might, for instance, be the concentration of a pollutant, the temperature of a coolant or density of a population: when $\underline{b} \neq \underline{0}$ it is not derivable from an extremum principle.

We may define an approximation $U(\underline{x})$ directly from Galerkin equations like (1.5), after first dealing with the inhomogeneous Dirichlet data on Γ_D. To ensure we maintain a strictly conforming approximation, we assume that the triangulation has been carried out so that Γ_D consists of a set of triangle sides with no more than one from any triangle: then define $G(\underline{x})$ as the piecewise blended interpolant which equals $g(\underline{x})$ on Γ_D, varies linearly in each of these triangles on the rays from Γ_D to the vertex not on Γ_D and is zero elsewhere. We define S_0^h, consistently with earlier usage, as the set of piecewise linear functions which are zero on Γ_D, and S_E^h as

$$S_E^h = \{V(\underline{x}) = G(\underline{x}) + W(\underline{x}) \mid W \in S_0^h\}. \tag{1.10}$$

Then the Galerkin approximation U to u is given by

$$(a\underline{\nabla}U, \underline{\nabla}\phi_i) + (\underline{\nabla}\cdot(\underline{b}U) + cU, \phi_i) = (f, \phi_i) \qquad \forall \phi_i \in S_0^h. \tag{1.11}$$

It is readily seen that u also satisfies these equations but, instead of an optimality property like (1.8), the most that we shall be able to deduce about the error is the following: defining the energy norm $||\cdot||_{AC}$ for the symmetric part of the operator by

$$||v||_{AC}^2 = (a\underline{\nabla}v, \underline{\nabla}v) + (cv, v), \tag{1.12}$$

we obtain for some constant K

$$||u-U||_{AC}^2 \le K \inf_{V\in S_E^h} ||u - V||_{AC}^2 . \tag{1.13}$$

The constant K will be bounded independently of h so that $\underline{\nabla}u$ will still have overall $O(h)$ accuracy, and indeed U will be $O(h^2)$ accurate, but the superconvergence phenomena are lost. Moreover, K depends on the local mesh Péclet numbers defined as bh/a, where b is the magnitude of $\underline{b}$, and these may be very large indeed. In practice the approximation may become very poor, exhibiting spurious oscillations which make it worthless.

1.3 A one-dimensional model problem

The origin of these oscillations can be exhibited by a simple model problem:

$$-au'' + bu' = 0 \qquad \text{on } (0,1) \tag{1.14a}$$

$$u(0) = 0, \quad u(1) = 1 \tag{1.14b}$$

where a,b are positive constants. The Galerkin equations (1.11) for a piecewise linear approximation on a uniform mesh of size h with $Jh = 1$, reduce to

$$h^{-1}(U_j-U_{j-1})(a+\tfrac{1}{2}bh) + h^{-1}(U_{j+1}-U_j)(-a+\tfrac{1}{2}bh) = 0 \tag{1.15a}$$

i.e. $$-\delta^2 U_j + (bh/a)\,\Delta_0 U_j = 0 \quad , \quad j = 1,2\ldots, J-1.$$

The central differences here, $\delta^2 U_j = U_{j+1}-2U_j+U_{j-1}$ and $\Delta_0 U_j = \frac{1}{2}(U_{j+1}-U_{j-1})$, are typical of a Galerkin approximation and it has long been recognised that they may give rise to spurious oscillation when bh/a is large. In fact, we can solve this system explicitly to find for $j=0,1,\ldots,J$ and $bh/a \neq 2$

$$U_j = \frac{\mu_0^j - 1}{\mu_0^J - 1} \qquad \text{where } \mu_0 = \frac{2+bh/a}{2-bh/a} . \tag{1.16}$$

When $bh/a = 2$, we have $U_j = 0$ for $j = 0,1,\ldots,J-1$ which is actually the exact solution of the reduced problem obtained by setting $a = 0$ in (1.14a): all the oscillatory solutions obtained from $bh/a > 2$ are entirely spurious and if one attempts to approximate the singular perturbation problem, $a \to 0$ with bh fixed,

one finds that the Galerkin equations become singular when J is even and then $U_j \to \infty$ for j odd. The exact solution of (1.14) is

$$u(x) = \frac{e^{bx/a}-1}{e^{b/a}-1}, \quad 0 \le x \le 1, \tag{1.17}$$

giving an exponential boundary layer at $x = 1$ as $a \to 0$, and (1.16) is a reasonable approximation only so long as μ_0 is a reasonable approximation to $e^{bh/a}$.

It might be argued, see for instance Gresho & Lee (1979), that it is unreasonable to attempt to approximate (1.14) when b/a is large without a local mesh refinement near the boundary layer. But in more complicated problems such layers are difficult to locate and the refinement expensive to implement. Thus most would agree that it is valuable to have available methods which will give good accuracy away from the boundary layer while using coarse meshes. What is certainly true is that, in various norms, the best piecewise linear fit is capable of giving an adequate representation of the solution under these circumstances: and with an appropriate choice of norm it can even give valuable information about the boundary layer, such as its half-width. Unfortunately, the Galerkin method does not give an approximation which is anywhere near optimal in any sense.

It should be noted in the above problem that, if the boundary condition at $x = 1$ were the Neumann condition, the exact solution would be identically zero. The Galerkin equations would have an extra equation $U_J - U_{J-1} = 0$, obtained from ϕ_J in (1.11), and their solution would also be identically zero for all bh/a. Thus the problem of spurious oscillations is a product of both the lack of self-adjointness and the boundary condition. We shall see however, in an example in Section 4, that even for a Neumann boundary condition the Galerkin method gives very poor accuracy compared with other methods.

1.4 Upwind differencing and Petrov-Galerkin methods

One-sided or upwind differencing has long been used in difference methods to avoid the oscillations described above. Completely replacing Δ_0 in (1.15b) by the backward difference operator Δ_- can however give rise to excessive false diffusion: writing $\Delta_- = \Delta_0 - \frac{1}{2}\delta^2$, the equations for U^- become

$$-(1+\tfrac{1}{2}bh/a)\,\delta^2 U_j^- + (bh/a)\,\Delta_0 U_j^- = 0, \quad j = 1,2,\ldots,J-1 \tag{1.18}$$

with the solution of the same form as in (1.16) but with μ_0 replaced by $\mu_- = 1 + bh/a$. This is clearly always monotone but, for instance, for $bh/a = 2$ gives $U_{J-1}^- \approx 1/3$ rather than $u(1-h) \approx e^{-2}$, a typical example of the enhanced diffusion apparent from (1.18).

More sophisticated schemes, using exponential fitting, go back to Allen & Southwell (1955): here the technique gives

$$-(1+\tfrac{1}{2}\xi bh/a)\delta^2 U_j^e + (bh/a)\Delta_0 U_j^e = 0, \quad j=1,2,\ldots,J-1 \tag{1.19}$$

where $$\xi = \coth(\tfrac{1}{2}bh/a) - (\tfrac{1}{2}bh/a)^{-1}, \tag{1.20}$$

which exactly reproduces the nodal values $u(jh)$ of u. Moreoever, recent results have shown that by the use of local values of ξ in a variable coefficient problem one can obtain an approximation which is uniformly accurate at the nodes as $bh/a \to \infty$ - see Doolan et al. (1980) as a general reference for these developments.

Such results relate to one-dimensional problems. In higher dimensions difficulties occur with cross-wind diffusion, that is enhanced diffusion perpendicular to the velocity vector $\underline{b}$. Much less progress has been made here with finite difference methods.

A large part of the development of finite element methods for diffusion-convection problems has been inspired by the earlier work on difference methods. Several techniques for generating upwind schemes have been proposed and used quite successfully on two-dimensional problems. We shall consider these in more detail in Section 3. The earliest of them (Christie et al. 1976) is based on a generalisation of the Galerkin formulation in which a different set T_0^h of test functions ψ_i, is introduced in (1.11) instead of the trial function basis $\{\phi_i\}$: this Petrov-Galerkin approximation is then given by

$$(a\underline{\nabla}U, \underline{\nabla}\psi_i) + (\underline{\nabla}.(\underline{b}U) + cU, \psi_i) = (f, \psi_i) \qquad \forall \psi_i \in T_0^h. \tag{1.21}$$

The problem is how to choose T_0^h for a given choice of S^h. Furthermore what criterion should be used for the assessment of accuracy and how should error bounds be derived? In particular, how closely should one adhere to the finite difference viewpoint, with the attendant emphasis on approximating nodal values and the awkwardness of estimating accuracy through estimating truncation error and bounding the inverse of the discrete operator?

All of the methods described in Section 3 to some extent adopt the finite difference viewpoint. In Section 4 an alternative approach is considered which is based on approximately symmetrising the bilinear form in (1.11). This leads to a near-optimal approximation to u in an integral norm which results naturally from the symmetrisation. We shall also show that the Section 3 methods can be regarded as approximate symmetrization in norm $||\cdot||_{AC}$. Thus the next section is devoted to developing the mathematical framework needed to study variational problems of the diffusion-convection type together with their approximation by generalised Galerkin procedures.

2. VARIATIONAL FORMULATION AND APPROXIMATION

2.1 Abstract problems and their approximation

The theoretical basis for Petrov-Galerkin methods is provided by the following generalisation of the Lax-Milgram lemma given by Babuška & Aziz (1972).

Theorem 2.1 Suppose $B(\cdot,\cdot)$ is a bilinear form on $H_1 \times H_2$, where H_1 and H_2 are real Hilbert spaces, which is continuous and coercive in the sense that there exist positive constants C_1 and C_2 such that

$$\text{(i)} \quad |B(v,w)| \le C_1 ||v||_{H_1} ||u||_{H_2} \qquad \forall v \in H_1, \ \forall w \in H_2; \tag{2.1a}$$

$$\text{(ii)} \quad \inf_{v\in H_1} \sup_{w\in H_2} \frac{|B(v,w)|}{||v||_{H_1} ||w||_{H_2}} \ge C_2 ; \tag{2.1b}$$

$$\text{(iii)} \quad \sup_{v\in H_1} |B(v,w)| > 0 \qquad \forall w \ne 0. \tag{2.1c}$$

Then for $\forall f \in H_2'$, there is a unique $u_0 \in H_1$ such that

$$B(u_0,w) = f(w) \qquad \forall w \in H_2 \tag{2.2a}$$

and

$$||u_0||_{H_1} \le ||f||_{H_2'}/C_2 \ . \tag{2.2b}$$

Proof By (2.1a) and the Riesz representation theorem, for each $v \in H_1$ there is a Riesz representer Rv of $B(v,w)$ in H_2 such that

$$(Rv,w)_{H_2} = B(v,w) \qquad \forall v \in H_1, \ \forall w \in H_2 \tag{2.3a}$$

and also that

$$||R||_{\mathcal{L}(H_1,H_2)} \le C_1 \ . \tag{2.3b}$$

That the mapping $R: H_1 \to H_2$ is closed follows from the closed graph theorem: furthermore, it follows from (2.1b) that

$$||Rv||_{H_2} = \sup_{w\in H_2} \frac{|B(v,w)|}{||w||_{H_2}} \ge C_2 ||v||_{H_1} \ . \tag{2.4}$$

Then by (2.1c) the mapping R must be onto: for otherwise, by the projection theorem, $\exists w^* \ne 0$ such that

$$(Rv,w^*)_{H_2} = 0 \qquad \forall v \in H_1$$

which contradicts (2.1c). From (2.4) we then have

$$||R^{-1}||_{\mathcal{L}(H_2,H_1)} \le 1/C_2 \tag{2.5}$$

and if w_0 is the Riesz representer of f in H_2 we can write $u_0 = R^{-1} w_0$ to obtain (2.2a,b). ∎

Corollorary 1 If H_1 is a subspace of H_2 it is sufficient to replace (2.1b) by taking the supremum over H_1 and requiring that

$$|B(v,v)| \geq C_2^{\prime}||v||^2_{H_1} \qquad \forall v\in H_1. \tag{2.6}$$

Corollorary 2 If $H_1 = H_2$ and (2.6) is satisfied then Theorem 2.1 reduces to the Lax-Milgram lemma.

Theorem 2.2 (A generalisation of Céa's lemma). Suppose $B(\cdot,\cdot)$ on $H_1 \times H_2$, f and u_0 are as in Theorem 2.1 and that M_1, M_2 are subspaces of H_1, H_2 respectively such that, for some positive constant C_2^M :

$$\text{(i)} \quad \inf_{V\in M_1} \sup_{W\in M_2} \frac{|B(V,W)|}{||V||_{H_1}||W||_{H_2}} \geq C_2^M ; \tag{2.7}$$

$$\text{(ii)} \quad \sup_{V\in M_1} |B(V,W)| > 0 \qquad \forall W\neq 0,\ W\in M_2. \tag{2.8}$$

Then there is a unique $U_0\in\overline{M}_1$ given by

$$B(U_0,W) = f(W) \qquad \forall W\in M_2 \tag{2.9}$$

and moreover,

$$||u_0-U_0||_{H_1} \leq \left[1+C_1/C_2^{\overline{M}}\right] \inf_{V\in M_1}||u_0-V||_{H_1} . \tag{2.10}$$

Proof With Rv defined as in Theorem 2.1, let P be the orthogonal projection $H_2\to\overline{M}_2$ and define S, in a similar way to R, as the mapping from $\overline{M}_1$ onto $\overline{M}_2$ such that

$$(SV,W)_{H_2} = B(V,W) \qquad \forall V\in\overline{M}_1, \forall W\in\overline{M}_2. \tag{2.11}$$

Then S is the restriction of PR to $\overline{M}_1$ because for $V\in\overline{M}_1$, $W\in\overline{M}_2$

$$(PRV,W)_{H_2} = (RV,W)_{H_2} = B(V,W).$$

Hence with w_0 the Riesz representer of f in H_2 and Pw_0 the representer in $\overline{M}_2$, we can set $U_0 = S^{-1}Pw_0 = S^{-1}PRu_0$ to obtain (2.9) from (2.11). Moreover, suppose V is any element of M_1, so that $S^{-1}PRV = V$, then we have

$$u_0 - U_0 = (I-S^{-1}PR)u_0 = (I-S^{-1}PR)(u_0-V);$$

$$\therefore \quad ||u_0-U_0||_{H_1} \leq ||I-S^{-1}PR|| \inf_{V\in M_1}||u_0-V||_{H_1}$$

$$\leq [1+C_1/C_2^M] \inf_{V\in M_1}||u_0-V||_{H_1} . \quad \blacksquare$$

Corollorary 3 If H_1 is a subspace of H_2 and (2.6) holds and if $M_1=M_2$, then Theorem 2.2 reduces to Céa's lemma. For, from (2.2a) and (2.9), $B(u_0-U_0,V) = 0$ $\forall V\in M_1$: thus, by (2.1a) and (2.6),

$$\|u_0-U_0\|_{H_1}^2 \le (C_2')^{-1}\, B(u_0-U_0,\, u_0-U_0) = (C_2')^{-1} B(u_0-U_0, u_0-V)$$
$$\le (C_1/C_2')\|u_0-U_0\|_{H_1}\|u_0-V\|_{H_1} \qquad \forall\, V\in M_1.$$

That is, (2.10) for this Galerkin case is replaced by

$$\|u_0-U_0\|_{H_1} \le (C_1/C_2') \inf_{V\in M_1} \|u_0-V\|_{H_1}. \tag{2.12}$$

Self-adjoint case. If $B(\cdot,\cdot)$ is symmetric as well as coercive, it can be used to define an inner product and thence a Hilbert space H so that we set $H_1=H_2=H$. Then (2.1a) is replaced by the Cauchy-Schwarz inequality $|(v,w)_H| \le \|v\|_H\|w\|_H$ with $C_1=1$ and (2.6) holds with $C_2'=1$. Thus, in Theorem 2.1, R becomes the identity mapping and the solution u_0 is just the Riesz representer of f in H. In the Theorem 2.2, S becomes the orthogonal projection of $\bar{M}_1$ onto $\bar{M}_2$ and $SU_0=Pu_0$. Moreoever, we can interpret C_2^M as measuring the extent to which elements of $\bar{M}_1$ can be approximated from $\bar{M}_2$: from (2.7) we have $C_2^M \le 1$ and, $\forall\, V\in M_1$,

$$\|V\|_H^2 = \|V-SV\|_H^2 + \|SV\|_H^2 \ge \|V-SV\|_H^2 + (C_2^M)^2\|V_H\|^2$$

i.e. $$\|V-SV\|_H^2 \le [1-(C_2^M)^2]\|V\|_H^2 \qquad \forall\, V\in M_1. \tag{2.13}$$

If we denote by U_0^* the orthogonal projection of u_0 onto $\bar{M}_1$, i.e. the Galerkin approximation, we have

$$\|u_0-U_0\|_H^2 = \|u_0-U_0^*\|_H^2 + \|U_0^*-U_0\|_H^2. \tag{2.14}$$

Then, rewriting the last term, recalling that $(u_0-U_0,W)_H = 0\ \forall\, W\in M_2$ and using (2.13), we obtain

$$\|U_0^*-U_0\|_H^2 = (u_0-U_0, U_0^*-U_0)_H$$
$$= (u_0-U_0, (I-S)(U_0^*-U_0))_H$$
$$\le \|u_0-U_0\|_H [1-(C_2^M)^2]^{\frac{1}{2}} \|U_0^*-U_0\|_H$$

i.e. $$\|U_0^*-U_0\|_H \le [1-(C_2^M)^2]^{\frac{1}{2}} \|u_0-U_0\|_H. \tag{2.15}$$

Hence we obtain the error bound for the Petrov-Galerkin method in this case,

$$\|u_0-U_0\|_H \le (1/C_2^M)\, \|u_0-U_0^*\|_H. \tag{2.16}$$

This is sharper than the bound obtained by merely putting $C_1 = 1$ in (2.10). In particular, as $C_2^M \to 1$ to give the Galerkin case, this error constant correctly tends to unity.

It was the fact that R became the identity which enabled this argument to go through so simply. In the general case, we could introduce the adjoint operator R^* to R and consider approximating $V\in M_1$ from R^*M_2 : if $W_V\in\bar{M}_2$ gives the best approximation,

$$||V-R^*W_V||^2_{H_1} = ||V||^2_{H_1} - ||R^*W_V||^2_{H_1}$$

and

$$||R^*W_V||_{H_1} = \sup_{W\in M_2} \frac{(R^*W_V,R^*W)_{H_1}}{||R^*W||_{H_1}} = \sup_{W\in M_2} \frac{(V,R^*W)_{H_1}}{||R^*W||_{H_1}}$$

$$\geq \frac{B(V,SV)}{||R^*SV||_{H_1}} \geq \frac{||SV||^2_{H_2}}{C_1||SV||_{H_2}} \geq \frac{C_2^M}{C_1}||V||_{H_1}$$

i.e. $$||V-R^*W_V||^2_{H_1} \leq [1-(C_2^M/C_1)^2]||V||^2_{H_1} \qquad \forall\ V\in M_1. \tag{2.17}$$

Also from $B(u_0-U_0,W) = 0$, $\forall W\in M_2$, we have

$$(u_0-U_0,\ R^*W)_{H_1} = 0 \qquad \forall\ W\in M_2$$

so that if U_0^* is the orthogonal projection of u_0 onto $\bar{M}_1$ (but not now the Galerkin approximation),

$$||U_0^*-U_0||^2_{H_1} = (u_0-U_0, U_0^*-U_0)_{H_1}$$

$$= \inf_{W\in M_2} (u_0-U_0, U_0^*-U_0-R^*W)_{H_1}$$

$$\leq ||u_0-U_0||_{H_1}[1-(C_2^M/C_1)^2]^{\frac{1}{2}}||U_0^*-U_0||_{H_1}.$$

Thus in the same way as (2.16) we obtain the following corrolorary: the derivation also indicates more clearly than (2.7) the appropriate choice of M_2 to ensure that U_0 is a near optional approximation to u_0 from M_1. We will use this later.

Corollorary 4 In Theorem 2.2 the error bound (2.10) can be improved to

$$||u_0-U_0||_{H_1} \leq (C_1/C_2^M) \inf_{V\in M_1}||u_0-V||_{H_1}. \tag{2.18}$$

2.2 Diffusion-convection problems

We recall the general statement of the problem in (1.9a,b) and shall henceforth assume that Neumann boundary conditions are never imposed on inflow boundaries: that is, if $\underline{n}$ is the unit outward normal to Γ, then

$$\underline{n}.\underline{b} \geq 0 \qquad \text{on} \quad \Gamma_N. \tag{2.19}$$

It is also useful to denote by Γ^- the inflow boundary, i.e. that part of Γ on which $\underline{n}.\underline{b} < 0$, and similarly by Γ^+, Γ^0 the outflow and tangential boundaries. So we have assumed that $\Gamma_N \cap \Gamma^- = \emptyset$.

In making such an assumption we have also assumed some smoothness of the boundary. It will be sufficient to assume throughout that Γ is Lipschitz continuous in the sense of Nečas (1967) - see also Ciarlet (1978) and Oden & Reddy (1976) as general references for results needed here: that is, there are a finite number of local co-ordinate systems such that every part of the boundary is defined by a Lipschitz continuous function in at least one of them. This means that $\underline{n}$ is defined almost everywhere on Γ, which may have corners and edges but no cusps. It also means that a trace operator tr is defined on $H^1(\Omega)$, extending the restriction of $v:\bar{\Omega} \to \mathbb{R}$ to Γ as a continuous linear mapping $tr : H^1(\Omega) \to L_2(\Gamma)$. Thus we can write in a conventional way

$$H_0^1(\Omega) = \{v \in H^1(\Omega) \,|\, tr\ v = 0 \ \text{ on } \ \Gamma_D\}. \tag{2.20}$$

By implication, too, Ω is bounded and a Poincaré-Friedrichs inequality holds: there exists a positive constant $C(\Omega)$ such that

$$||v||_{0,\Omega} \le C(\Omega)|v|_{1,\Omega} \qquad v \in H_0^1(\Omega), \tag{2.21}$$

where the notation $||\cdot||_{m,\Omega}$ is used for the usual Sobolev norm for $H^m(\Omega)$ and $|\cdot|_{i,\Omega}$ for $i=1,\ldots,m$ denotes the corresponding semi-norms. Finally, Green's formulae also hold: for example, with $u \in H^1(\Omega)$, $\underline{v} \in [H^1(\Omega)]^d$, $\Omega \subset \mathbb{R}^d$, $d=1,2$ or 3 we have

$$\int_\Omega u \underline{\nabla}\cdot\underline{v}\, d\Omega = -\int_\Omega \underline{v}\cdot\underline{\nabla} u\, d\Omega + \int_\Gamma u\underline{n}\cdot\underline{v}\, d\Gamma. \tag{2.22}$$

To apply (2.22) to (1.9) we assume that

$$0 < a \in C^0(\bar{\Omega}),\ \underline{b} \in [H^1(\Omega)]^d \ \text{ and } \ 0 \le c \in L_2(\Omega). \tag{2.23}$$

Then, denoting by L the diffusion-convection operator on the left of (1.9a), we have for $v \in H^2(\Omega)$, $w \in H^1(\Omega)$:

$$\begin{aligned}(Lv,w) + (a\underline{n}\cdot\underline{\nabla}v, w)_{\Gamma_N} &= (a\underline{\nabla}v - \underline{b}v,\ \underline{\nabla}w) + (cv,w) \\ &\quad + (\underline{n}\cdot(\underline{b}v - a\underline{\nabla}v), w)_{\Gamma_D} + (\underline{n}\cdot\underline{b}v, w)_{\Gamma_N},\end{aligned} \tag{2.24}$$

where $(\cdot,\cdot)_{\Gamma_N}$ denotes the L_2 inner product over Γ_N and similarly for Γ_D. A more convenient basic definition of the bilinear form to which we shall apply Theorems 2.1 and 2.2 than (2.24) is however the following:

$$B(v,w) := (a\underline{\nabla}v, \underline{\nabla}w) + (\underline{\nabla}\cdot(\underline{b}v) + cv, w) - (a\underline{n}\cdot\underline{\nabla}v, w)_{\Gamma_D}. \tag{2.25}$$

This is clearly continuous on $H^1(\Omega) \times H_0^1(\Omega)$ which covers most cases of interest: in Section 4, however, we shall try lifting the condition that $w=0$ on $\Gamma_D \cap \Gamma^-$

and this will entail restricting the class of v's.

The inhomogeneous boundary data of (1.9) is incorporated in the formulation by assuming that g is derived by the trace operator $(tr)_D$ corresponding to Γ_D operating on some function $G:\Omega \to \mathbb{R}$. Then we define the following linear functional corresponding to all the data:

$$F(w) := (f-\underline{\nabla}\cdot(\underline{b}G) - cG,w) - (a\underline{\nabla}G,\underline{\nabla}w) + (a\underline{n}\cdot\underline{\nabla}G,w)_{\Gamma_D} . \tag{2.26}$$

To ensure that this is bounded over $w\in H_0^1(\Omega)$ it is sufficient to assume:

$$f\in L_2(\Omega) , \quad g=(tr)_D G \text{ s.t. } G\in H^1(\Omega). \tag{2.27}$$

Again, widening the class of w will entail restricting that of G. We shall furthermore assume that the convective medium is incompressible in obtaining the following Theorem.

Theorem 2.3 Suppose that for the problem (1.9a,b) the assumptions (2.19), (2.23) and (2.27) are satisfied and that

$$\underline{\nabla}\cdot\underline{b} = 0. \tag{2.28}$$

Then a weak solution exists of the form $u=u_0+G$, where $u_0\in H_0^1(\Omega)$ is uniquely defined by

$$B(u_0,w) = F(w) \qquad \forall w\in H_0^1(\Omega) \tag{2.29}$$

and $B(\cdot,\cdot)$, $F(\cdot)$ are defined by (2.22) and (2.24).

Proof We check the hypotheses of Theorem 2.1, or rather of the Lax-Milgram lemma, with $H_1=H_2=H_0^1(\Omega)$. For the coercivity we have

$$B(v,v) = (a\underline{\nabla}v,\underline{\nabla}v) + (cv,v) + (\underline{\nabla}\cdot(bv),v) ;$$

and from (2.28), (2.19) there follows for $u\in H_0^1(\Omega)$

$$(\underline{b}\cdot\underline{\nabla}v,v) = (\underline{\nabla}\cdot(\underline{b}v),v) = -(\underline{b}\cdot\underline{\nabla}v,v) + (\underline{n}\cdot\underline{b}v,v)_{\Gamma_N}$$

$$\text{i.e.} \qquad (\nabla\cdot(\underline{b}v),v) = \tfrac{1}{2}(\underline{n}\cdot\underline{b}v,v)_{\Gamma_N} \geq 0. \tag{2.30}$$

Since we have $a>0$, $c\geq 0$ and $\Gamma_D \neq \emptyset$, (2.21) ensures that (2.6) is satisfied for some C_2'. Similarly, and with the use of the Cauchy-Schwarz inequality, the continuity condition (2.1a) is satisfied for some C_1. We defer until the next sub-section discussion on the sharpest bounds attainable for the ratio C_1/C_2 except to note that this will depend on bounding $(\underline{b}\cdot\underline{\nabla}v,w)$ in terms of $(a\underline{\nabla}v,\underline{\nabla}v)$ and $(a\underline{\nabla}w,\underline{\nabla}w)$.

As this solution u is clearly uniquely defined, independently of the choice of G, and since a classical solution of (1.9a,b) because of (2.24) also satisfies (2.29), identification of the weak and classical solutions depends only on the regularity of the latter. For general theorems covering this we refer the reader to Agmon, Douglis & Nirenburg (1964).

2.3 Galerkin approximation

The analysis of finite element methods takes its simplest form if we make the following standard assumptions regarding the approximation properties of the space S^h of trial functions:

(i) S^h is conforming, i.e. $S^h \subset H^1(\Omega)$;

(ii) S^h has order $r > 1$ and is regular, i.e. $\forall g \in H^\ell(\Omega), \ell \geq 1$, there exists an element $V \in S^h$ such that for some constant K

$$||g-V||_{s,\Omega} \leq Kh^\mu ||g||_{\ell,\Omega} \qquad s = 0,1 \qquad (2.31)$$

where $\mu = \min (r-s, \ell-s)$.

To ensure these properties we shall assume Γ is polygonal and Ω is subdivided into elements which have maximal diameter h and satisfy a regularity condition, such as all interior angles are uniformly bounded from zero. In $\mathbb{R}^2$, elements will either be triangles or quadrilaterals, with corresponding elements in $\mathbb{R}^3$: parametric transformations will generally be necessary to map quadrilaterals in global variables into rectangles in local variables and such transformations (c.f. isoparametric elements) are often used to approximate curved boundaries, but this is beyond the scope of the present lectures.

We shall mainly consider piecewise linear approximation on triangles or bilinear approximation on rectangles, both of which are examples of $r=2$ in (2.31). Piecewise quadratic, or biquadratic, elements similarly give $r=3$: generally speaking, if all polynomial functions up to degree k on each element are contained in S^h, and the parameters are chosen to ensure continuity between elements, then (2.31) will hold with $r=k+1$. We also omit consideration of Hermitian elements, such as Hermite cubics, so ensuring that we can write the general member $V \in S^h$ in the Lagrangian form

$$V(\underline{x}) = \sum_{(j)} V_j \phi_j(\underline{x}), \qquad (2.32)$$

where ϕ_j is the basis function corresponding to node $(\underline{x}_j)$ such that $\phi_j(\underline{x}^i) = \delta_{ij}$ and hence $V_j = V(\underline{x}_j)$. Thus $S^h = \text{span}\{\phi_j\}$.

We suppose further that Γ_D is composed of an integral number of element sides so that, defining

$$S_0^h = S^h \cap H_0^1(\Omega), \qquad (2.33)$$

we find that S_0^h is spanned by a subset of $\{\phi_j\}$. Then we can introduce the Galerkin approximation U_0 to the solution u_0 of (2.29),

$$U_0 \in S_0^h : \quad B(U_0, \phi_i) = F(\phi_i) \qquad \forall\, \phi_i \in S_0^h . \qquad (2.34)$$

Thus we have

$$B(u_0 - U_0, \phi_i) = 0 \qquad \forall\, \phi_i \in S_0^h . \qquad (2.35)$$

Let us estimate the error u_0-U_0 first in the norm $||\cdot||_{AC}$ introduced in (1.12). From the assumptions (2.23) on a, $\underline{b}$ and c and the boundedness of Ω we can introduce constants P_1, P_2 such that

$$|(\underline{\nabla}\cdot(\underline{b}v),w)| \leq P_1||v||_{AC}||w||_{0,\Omega} \tag{2.36a}$$

and

$$||w||_{0,\Omega} \leq P_2||w||_{AC}. \tag{2.36b}$$

We can regard the product P_1P_2 as a global Péclet number, particularly when c=0, as it has the dimension of bL/a where L is a scale length. We also suppose that $u_0 \in H^r(\Omega)$ and that by (2.31) there therefore exists a member $\tilde{U}_0$ of S_0^h such that for some constants K and P_3

$$||u_0-\tilde{U}_0||_{0,\Omega} \leq K_r h^r ||u_0||_{r,\Omega} \tag{2.37a}$$

and

$$||u_0-\tilde{U}_0||_{AC} \leq K_r P_3 h^{r-1} ||u_0||_{r,\Omega}. \tag{2.37b}$$

Then from (2.30), (2.35) and (2.36a)

$$\begin{aligned} ||u_0-U_0||_{AC}^2 &\leq B(u_0-U_0,u_0-U_0) = B(u_0-U_0,u_0-\tilde{U}_0) \\ &= (u_0-U_0,u_0-\tilde{U}_0)_{AC} + (\underline{\nabla}\cdot[\underline{b}(u_0-U_0)],u_0-\tilde{U}_0) \\ &\leq ||u_0-U_0||_{AC}\left[||u_0-\tilde{U}_0||_{AC}+P_1||u_0-\tilde{U}_0||_{0,\Omega}\right]; \end{aligned} \tag{2.38}$$

that is, by (2.37)

$$||u_0-U_0||_{AC} \leq [1+P_1h/P_3]\, K_r P_3 h^{r-1} ||u_0||_{r,\Omega}. \tag{2.39}$$

Here we can regard P_1h/P_3 as a mesh Péclet number which is seen to completely represent the loss of accuracy attributable to the convection term when the Galerkin method is used. To compare this result with that obtained directly from Céa's lemma, (2.12), we see from (2.36) that C_1 in (2.1a) can be taken as $1+P_1P_2$ and C_2' from (2.27) taken as unity: thus we have been able to replace a global Péclet number with a mesh Péclet number.

To obtain an error estimate in a lower order norm, in particular in the L_2 norm, we use the device due to Aubin and Nitsche. We let $v_0 \in H^2(\Omega)$ be the solution of the adjoint problem

$$B^*(v_0,w) = (u_0-U_0,w) \qquad \forall w \in H_0^1 \tag{2.40}$$

for which, by the ellipticity of the equation, there must be an estimate of the form

$$||v_0||_{2,\Omega} \leq P_4||u_0-U_0||_{0,\Omega}. \tag{2.41}$$

Then taking $w=u_0-U_0$ in (2.40) we obtain

$$||u_0-U_0||^2_{0,\Omega} = B^*(v_0,u_0-U_0) = B(u_0-U_0,v_0)$$
$$= B(u_0-U_0,\ v_0-\tilde{V}_0)$$
$$\le ||u_0-U_0||_{AC}\left[||v_0-\tilde{V}_0||_{AC} + P_1||v_0-\tilde{V}_0||_{0,\Omega}\right]. \qquad (2.42)$$

Here, $\tilde{V}_0$ is any member of S_0^h and we can assume it is chosen so that bounds analogous to (2.37) hold with $r=2$. Then substituting also from (2.39) and (2.41) we obtain

$$||u_0-U_0||_{0,\Omega} \le [1+P_1h/P_3]^2 \ K_rK_2P_3P_4h^r||u_0||_{r,\Omega} . \qquad (2.43)$$

Thus the usual extra power of h is obtained but at the cost of extra constants, in particular a further factor from the mesh Péclet number. Note that as $a\to 0$, although $P_4\sim a^{-1}$ we also have $P_3\sim a^{\frac{1}{2}}$ and so $P_3^2P_4 = O(1)$.

It is worth noting that this same technique can be used to obtain the pair (2.37 a and b). Suppose U_0^* is the optimal approximation to u_0 in $||\cdot||_{AC}$, that is

$$(u_0-U_0^*,\phi_i)_{AC} = 0 \qquad \forall\ \phi_i\in S_0^h. \qquad (2.44)$$

Then we can introduce w_0 by

$$(w_0,w)_{AC} = (u_0-U_0^*,w) \qquad \forall\ w\in H_0^1 \qquad (2.45)$$

for which we shall have a bound $||w_0||_{2,\Omega} \le P_4^*||u_0-U_0^*||_{0,\Omega}$ and an optimal approximation W_0^* with $||w_0-W_0^*||_{AC} \le K_2P_3^*h||w_0||_{2,\Omega}$. Hence we obtain, in the same way as (2.42),

$$||u_0-U_0^*||_{0,\Omega} \le K_2P_3^*P_4^*h||u_0-U_0^*||_{AC} \qquad (2.46)$$

and substituting this in (2.38) obtain

$$||u_0-U_0||_{AC} \le [1+K_2P_1P_3^*P_4^*h]||u_0-U_0^*||_{AC} . \qquad (2.47)$$

We again see that $K_2P_1P_3^*P_4^*h$ can be regarded as a mesh Péclet number and, with a bound on $||u_0-U_0^*||_{AC}$, (2.47) can be used to replace (2.39).

Thus, too (2.47) shows that as $h\to 0$ the Galerkin approximation U_0 eventually becomes "near optimal" and one can expect superconvergence results to hold: the practical difficulty is that this will occur for only extremely small h when the Péclet number is large.

2.4 The one-dimensional model problem

We conclude this section by applying some of the results in the earlier subsections to the model problem (1.14). We reformulate and generalise this to

$$-au_0''+bu_0' = f, \qquad u_0(0) = u_0(1) = 0, \qquad (2.48)$$

with $f\equiv -b$ giving for $u=u_0+x$ the same result as (1.14). Working in $H_0^1(0,1)$ equipped with the norm $||v||_{AC}^2 = a||v'||_0^2$, the mapping R of Thorem 2.1 can be explicitly derived:

$$B(v,w) = \int_0^1 (av'w' + bv'w)dx = \int_0^1 a(Rv)'\, w'dx \qquad \forall\, v,w\in H_0^1$$

i.e. $$a(Rv)' - av' + bv = \text{const.} = b\bar{v}$$

i.e. $$(Rv)(x) = v(x) - (b/a)\int_0^x [v(t)-\bar{v}]dt, \tag{2.49}$$

where $\bar{v} = \int_0^1 v dt$: R^* has the same form with the sign of b changed. Similarly we find

$$(R^{-1}w)(x) = \int_0^x e^{b(x-t)/a}\, w'(t)dt - [1-e^{-b/a}]^{-1}\,[e^{bx/a}-1]\int_0^1 e^{-bt/a} w'(t)dt. \tag{2.50}$$

It is clear directly from (2.6) that $C_2'=1$ and from (2.49) that

$$||Rv||_{AC}^2 = ||v||_{AC}^2 + (b^2/a)\int_0^1 (v-\bar{v})^2 dx \;: \tag{2.51}$$

it is therefore evident that $||R^{-1}||=1$ and from a Fourier analysis one can show that

$$||R|| = (1+ b^2/4\pi^2a^2)^{\frac{1}{2}}. \tag{2.52}$$

This then is the constant which appears in Cea's lemma (2.12).

For the Galerkin approximation using piecewise linear elements on a uniform mesh we can also carry out the analysis leading to (2.47). It is easy to see that $P_1=b/a^{\frac{1}{2}}$ and $P_4^*=1/a$ and elementary approximation theory gives $K_2=1/\pi$ with $P_3^*=a^{\frac{1}{2}}$. Thus (2.47) becomes

$$||u_0-U_0||_{AC} \le [1 + bh/a\pi]\, ||u_0-U_0^*||_{AC}, \tag{2.53}$$

a much sharper result than that given by Cea's lemma.

Moreover, u_0-U_0 is given explicitly by (1.16) and (1.17), and it is easy to see (cf. (3.1) below) that U_0^* actually interpolates u_0. Thus we can readily calculate the ratio of the two norms in (2.53): we find that

$$||u_0-U_0||_{AC}^2 = b(1-\mu_0^{-J})^{-1}\frac{\mu-\mu_0}{1-\mu\mu_0} + O(e^{-b/a}), \tag{2.54a}$$

$$||u_0-U_0^*||_{AC}^2 = b\left[\frac{1}{2} - \frac{a}{bh}\,\frac{\mu-1}{\mu+1}\right] + O(e^{-b/a}), \tag{2.54b}$$

where μ_0 is given by (1.16) and $\mu=e^{bh/a}$, so that the ratio does not take a simple form. However, denoting bh/a by β, the two limiting forms are as follows:

$$||u_0-U_0||_{AC}/||u_0-U_0^*||_{AC} \sim 1+\frac{23}{240}\beta^2 \qquad \text{as} \quad \beta\to 0 \tag{2.55a}$$

$$\sim (\tfrac{1}{2}h\beta)^{\frac{1}{2}} \qquad \text{as} \quad \beta\to\infty, \text{ even } J. \tag{2.55b}$$

Apart from the fact that it diverges, this second limit is not particularly useful since even U_0^* is a very poor approximation to u_0 in this norm: indeed this

limit follows directly from the observations that $||U_0||_{AC}/||u_0||_{AC} \to (\frac{1}{2}h\beta)^{\frac{1}{2}}$ and $||U_0^*||_{AC}/||u_0||_{AC} \to (2/\beta)^{\frac{1}{2}}$ as $\beta \to \infty$.

A more dramatic demonstration of the inadequacy of the Galerkin approximation when β is large is provided by looking at the discrete equations which represent the approximation process $U_0 = S^{-1}Pw_0$. Here w_0, the Riesz representer of $-b$ in H_{AC}, is given by

$$w_0(x) = -\tfrac{1}{2}(b/a)x(1-x)$$

and P corresponds to taking nodal values. Then from equation (2.49) for R, we obtain for $SU_0 \equiv PRU_0 = Pw_0$

$$U_j^0 - \beta\left[\sum_1^{j}{}' U_i^0 - jh\sum_1^{J} U_i^0\right] = -\tfrac{1}{2}\beta hj(J-j),\ j=1,2,..,J-1; \qquad (2.56)$$

here $\{U_j^0,\ j=0,1,\ldots,J\}$ are the nodal values of $U_0(x)$ and the prime on the first sum indicates that only $\frac{1}{2}U_j^0$ is included. It is readily seen that for even J the vector $\{0,1,0,1,\ldots,1,0\}$ is annihilated by the operations in the square brackets. This is the vector for which the norm $||S^{-1}||=1$ is attained and there will always be a component of this of order of magnitude β in the solution U_0: thus it is that for even J the Galerkin solution exhibits unbounded oscillations as $\beta \to \infty$. One can similarly see why for odd J the oscillations are much less violent.

3. PETROV-GALERKIN METHODS USING EXPONENTIAL, UPWINDING AND STREAMLINE-DIFFUSION TECHNIQUES

In surveying these three (overlapping) techniques, we shall generally introduce them for the 1D model problem (1.14) before indicating their developments for more general problems. We shall also work in the space H_{AC} with norm $||\cdot||_{AC}$ defined in (1.12). This is a particularly appropriate norm in this case, in view of the motivation of several of the key ideas by finite difference methods. For, if the trial space S_E^h consists of piecewise linear functions with nodes $\{x_j\}$, $j = 0,1,\ldots,J$, $x_0=0$, $x_J=1$, the best fit $U^* \in S_E^h$ to u in this norm satisfies

$$a\int_0^1 (u' - U^{*\prime})\phi_j'\,dx = 0,\ j = 1,2,\ldots,J-1$$

$$\text{i.e.} \qquad \frac{\Delta_-[u(x_j)-U_j^*]}{\Delta_- x_j} = \frac{\Delta_+[u(x_j)-U_j^*]}{\Delta_+ x_j}\ . \qquad (3.1)$$

Denoting the common value by D, we find by multiplying each ratio by the denominator and summing that $D=0$: hence $u(x_j)-U_j^* =$ constant $= 0$. Thus the best piecewise linear fit in this norm is also the best (i.e. exact) fit at the nodes.

3.1 Use of piecewise exponentials

As with finite differences, several early methods exploited the exponential character of solutions to one dimensional diffusion-convection problems. Three or four differing approaches have been adopted.

(i) Liouville transform (Guymon, 1970). If in the 1D model problem (1.14) we set

$$w(x) = e^{-\frac{1}{2}bx/a}u(x) \tag{3.2a}$$

the problem is symmetrized to

$$-aw'' + (b^2/4a)w = 0, \quad w(0) = 0, \quad w(1) = e^{-\frac{1}{2}b/a} . \tag{3.2b}$$

This may then be solved by a Galerkin method to give a best fit in the mixed norm

$$\int_0^1 [aw'^2 + (b^2/4a)w^2]dx, \tag{3.3a}$$

which tends to the L_2 best fit as $b/a \to \infty$. However, any errors will be amplified by $\exp(\frac{1}{2}bx/a)$ on transforming back to the original variables and this is ill-conditioned in the singular perturbation limit. Equivalently, we can see that after transforming back we have an optimal approximation in the norm

$$\int_0^1 av'^2\, e^{-bx/a}dx \tag{3.3b}$$

which concentrates attention away from the boundary layer near $x=1$ which is of most interest. Guymon et al. (1970) have also extended this technique to two-dimensional flow problems, but the above arguments indicate that it should be used only with very great care.

(ii) Exponential trial space (K.E. Barrett, 1974, 1977). When any inhomogenous term in the equation is such that the solution is predominantly exponential in character, the following basis functions (on a uniform mesh) would seem a natural choice for the trial space: with $\beta=bh/a$ and $\phi_j(x)=\phi(h^{-1}x-j)$ we set

$$(1-e^{-\beta})\phi(t) = \begin{cases} e^{\beta t} - e^{-\beta} & -1 \le t \le 0 \\ 1 - e^{\beta(t-1)} & 0 \le t \le 1. \end{cases} \tag{3.4}$$

The Galerkin method for the model problem then of course gives the exact exponential solution. For the more general problem $-au''+ bu' = f(x)$, $u(0)$ and $u(1)$ given, then by (2.12) and (2.52) the approximation U has an error bound

$$||u-U||_{AC} \le (1 + b^2/4\pi^2a^2)^{\frac{1}{2}} \inf_{V\in S_E^h} ||u-V||_{AC}. \tag{3.5}$$

On this basis when b/a is large the trial space of exponentials has to be capable of very close approximation to the solution if the method is to be used with confidence: although from the previous section we might expect this factor to be replaced by a mesh Péclet number, we shall not pursue these estimates further and

instead we will consider below the accuracy attained at the nodes. One could also use these trial functions together with piecewise linear test functions: such a method is considered by Griffiths & Lorenz (1978), who show that this gives a lower error bound than any of the alternative upwind test functions (3.12) discussed in the next sub-section. When the coefficients a and b depend on x, local values of β can be used in each trial function and a similar error bound to (3.5) obtained.

(iii) Exponential test space (Hemker, 1977). This is motivated by some of the earliest work on superconvergence at the nodes, by de Boor & Swartz (1973) and Douglas & Dupont (1973). Consider a general one-dimensional problem, let $G^*_\xi(x)$ be the Green's function of the adjoint problem and denote the delta function $\delta(x-\xi)$ by $\delta_\xi(x)$. Then the weak form of the equation for G^*_ξ is

$$B(v, G^*_\xi) = (\delta_\xi, v) = v(\xi) \qquad \forall\, v \in H^1_0. \tag{3.6}$$

Now suppose U is a Petrov-Galerkin approximation to u obtained with a test space T^h so that

$$B(u-U, W) = 0 \qquad \forall\, W \in T^h. \tag{3.7}$$

Then (3.6) and (3.7) together give

$$u(\xi)-U(\xi) = B(u-U, G^*_\xi) = B(u-U, G^*_\xi - W) \qquad \forall\, W \in T^h, \tag{3.8}$$

and from (2.1a) we have

$$|u(\xi)-U(\xi)| \le C_1 ||u-U||_{AC} \inf_{W \in T^h} ||G^*_\xi - W||_{AC}. \tag{3.9}$$

As G^*_ξ has a discontinuous gradient at $x=\xi$, the last factor here will be reasonably small only when ξ is a mesh point. Then for any sensible choices of S^h and T^h the order of accuracy at the nodes should be double that in the $||\cdot||_{AC}$ norm: it should be noted, however, that for linear elements this is no improvement over the L_2 error bounds obtained by the Aubin-Nitsche arguments as in (2.43).

In the 1D model problem G^*_ξ consists of piecewise negative exponentials and hence, on a uniform mesh, we should take as test basis functions $\psi_j(x) = \psi(h^{-1}x - j)$, where

$$(1-e^{-\beta})\,\psi(t) = \begin{cases} 1-e^{-\beta(t+1)} & -1 \le t \le 0 \\ e^{-\beta t} - e^{-\beta} & 0 \le t \le 1 \end{cases}; \tag{3.10}$$

these are the reflection of the trial functions given by (3.4) about $t=0$. Then G^*_ξ is approximated exactly and nodal values of u are exactly reproduced even for $-au'' + bu' = f(x)$, for general f, and any reasonable choice of trial space.

For a piecewise linear trial space, an alternative interpretation based on the error bounds of Section 2 is possible. From (2.17) and the subsequent argument,

it is clear that $R^*T_0^h$ should be such as to approximate S_0^h well. An explicit expression for R was given in (2.49) for the operator in the model problem and in the $||\cdot||_{AC}$ norm. From this it is an easy calculation to show that indeed

$$R^*\psi_j = \tfrac{1}{2}(1+e^{-\beta})\phi_j \qquad j=1,2,\ldots,J-1 \tag{3.11}$$

exactly, where the $\{\phi_j\}$ are the piecewise linear basis functions. Thus the resulting approximation is optimal in $||\cdot||_{AC}$ and hence exact at the nodes.

One of the disadvantages of using exponentials as either trial or test functions is the difficulty of evaluating the inner products involving these rapidly varying functions. Hemker (1977) has developed specialised quadrature formulae for this purpose. He also considered using these test functions only where the solution varied rapidly, as has Axelsson (1981) who used the very similar technique of introducing the negative exponential as a weight function in the bilinear form. More fundamental difficulties arise when any of these exponential-based techniques are extended into two dimensions and little progress has so far been reported.

3.2 Upwind methods

Zienkiewicz (1975) seems to have been the first to raise the possibility of choosing the test space in a Petrov-Galerkin scheme in order to obtain the same effects as upwind differencing. Mitchell and his colleagues quickly took up the challenge and a number of promising techniques were developed - see Christie et al. (1976), Heinrich et al. (1977) and the survey article Heinrich & Zienkiewicz (1979).

For the operator in the 1D model problem, it is apparent from the foregoing that either a positive exponential trial space in a Galerkin formulation or a negative exponential test space in a Petrov-Galerkin scheme will reproduce the Allen-Southwell difference operator: but clearly many other test spaces could achieve this. One of the simplest that may be used with a piecewise linear basis on a uniform mesh, $\{\phi_j\}$, is the following: with $\psi_j(x) = \psi(h^{-1}x-j)$ and $\sigma_j(x) = \sigma(h^{-1}x-j)$ we set

$$\psi(t) = \phi(t) + \alpha\sigma(t) \tag{3.12a}$$

with

$$\sigma(t) = \begin{cases} -3t(1-|t|) & |t| \le 1 \\ 0 & |t| > 1 \,. \end{cases} \tag{3.12b}$$

We see that $(\phi_j', \sigma_i') = 0$ for $1 \le i, j \le J-1$ so that the terms in the stiffness matrix arising from the diffusion operator do not depend on the parameter α: but for the convection terms we obtain

$$(\phi_j', \sigma_i) = -(\phi_j, \sigma_i') = \begin{cases} -\frac{1}{2} & j=i\pm1 \\ 1 & j=i \\ 0 & |j-1|>1 \,. \end{cases} \tag{3.13}$$

Thus for the 1D model problem we obtain

$$-a\delta^2 U_i + bh(\Delta_0 U_i - \tfrac{1}{2}\alpha\delta^2 U_i) = 0 \ . \tag{3.14}$$

Setting $\alpha=1$ gives the fully upwinded scheme of (1.18) and any choice such that $2a+\alpha bh>bh$, that is $\alpha>1-2a/bh$, avoids an oscillatory solution to the difference scheme by ensuring that it satisfies a maximum principle. The Allen & Southwell exponentially-fitted scheme is obtained by setting, as in (1.20)

$$\begin{aligned}\alpha = \xi &:= \coth(\tfrac{1}{2}bh/a) - (\tfrac{1}{2}bh/a)^{-1} \\ &= \coth\tfrac{1}{2}\beta - 2/\beta \ .\end{aligned} \tag{3.15}$$

It is easily seen that ξ varies smoothly from -1 to $+1$ as β ranges from $-\infty$ to $+\infty$, with $\xi\sim\beta/6$ as $\beta\to 0$ and $\xi\sim 1-2/\beta$ as $\beta\to\infty$.

It is interesting to see what choice of α is indicated by the error bounds in Theorem 2.2, and its Corollorary 4, when the norm $||\cdot||_{AC}$ is used: a detailed analysis is given by Griffiths & Lorenz (1978). In Theorem 2.2 only C_2^M of (2.7) is affected by the choice of test functions, which should thus be chosen to maximise $||S^{-1}||$. Denoting by A and B the stiffness matrices representing the difference operators $-a\delta^2$ and $bh\Delta_0$, we see from the defining relation (2.11) that an expression for S^{-1} can be obtained from the following (we denote by $\underset{\sim}{V}$ the vector of nodal values of $V\in S_0^h$ and similarly for $W\in T_0^h$):

$$(1+3\alpha^2)A(\underset{\sim}{S}\underset{\sim}{V}) = [(1+\tfrac{1}{2}\alpha\beta)A + B]\underset{\sim}{V} \ . \tag{3.16}$$

The matrix on the left represents the inner product $(\cdot,\cdot)_{AC}$ in the basis (3.12a) of T_0^h, obtained using (3.13) and the fact that $(\sigma_j',\sigma_i') = 3(\phi_j',\phi_i')$. Then using Fourier analysis we find

$$C_2^M = \min \frac{||SV||_{AC}}{||V||_{AC}} = (1+3\alpha^2)^{-\frac{1}{2}}[(1+\tfrac{1}{2}\alpha\beta)^2 + \tfrac{1}{4}\beta^2\tan^2\tfrac{1}{2}\pi h]^{\frac{1}{2}}. \tag{3.17}$$

The maximum value is given quite accurately by neglecting the term $\tan^2\frac{1}{2}\pi h$, leading to the choice $\alpha = \beta/6$: this agrees with (3.15) for small β but at first sight seems quite unreasonable for large β.

Before considering this point further, let us derive the choice of α obtained by optimising the error bound in (2.17) : in particular, we choose α to minimise

$$\min_{\gamma} ||\gamma R^*\psi_j - \phi_j||^2_{AC}. \tag{3.18}$$

Using the expression for R in (2.49) and exploiting the fact that $\sigma' = -6(\phi-\tfrac{1}{2})$, it is a straightforward computation to obtain

$$||\gamma R^*\psi_j - \phi_j||^2_{AC} = (a/h)\int_{-1}^{1} \gamma^2[(1-\gamma^{-1})\phi' + (\beta-6\alpha)\phi + 3\alpha + \alpha\beta\sigma]^2 dt \tag{3.19}$$

and to find that this is minimised by

$$\alpha = \tfrac{5}{9}\beta(\beta^2+3)/(\beta^2+10). \tag{3.20}$$

This gives the correct behaviour, $\alpha\sim\beta/6$, for $\beta\to 0$ and very similar behaviour to that derived from (3.17), namely $\alpha\sim 5\beta/9$ for $\beta\to\infty$. Moreover, the latter is very easily understood in terms of the function fitting needed to minimise (3.19) : since ϕ and the constant 3α are the only even functions, one needs $\alpha=O(\beta)$; then $\gamma\alpha\beta\sigma$ is of very similar form to ϕ', the best fit being given by $\gamma\alpha\beta=5/3$.

To understand the fact that these error bound arguments lead to $\alpha\to\infty$ as $\beta\to\infty$, instead of $\alpha\to 1$ as $\beta\to\infty$ in the Allen & Southwell scheme, we need to remember that they were based on the form of the problem (2.48) with homogenous boundary conditions and general data f. In the singular limit, the Allen & Southwell scheme drops the right-hand boundary condition and the data that goes with it and approximates $bu'=0$, $u(0)=0$ by $b\Delta_- U=0$, $U(0)=0$: and for a general piecewise linear data function F the Petrov-Galerkin method based on (3.12) with $\alpha=1$ will give

$$b\Delta_- U_i = \frac{h}{6}[6+\delta^2-3\Delta_0]F_i$$

i.e.
$$b(U_i-U_{i-1}) = \frac{h}{12}[-F_{i+1}+8F_i+5F_{i-1}] \quad , \tag{3.21}$$

not a very convincing approximation to $bu'=F$. On the other hand with $\alpha\to\infty$, the scheme for $U^0=U-x$ with homogeneous boundary conditions at each end becomes

$$b\delta^2 U_i^0 = h\Delta_0(F-b). \tag{3.22a}$$

Constant data clearly gives a null solution and this takes the place of a boundary condition being dropped: for one integration can be effected and (3.22a) reduces for U to

$$b(U_i-U_{i-1})=\tfrac{1}{2}h(F_i+F_{i-1}), \tag{3.22b}$$

a much more satisfactory approximation.

We should perhaps not consider these results for high β as too significant, for we have already seen that $||u_0-U_0^*||_{AC}$ for the optimal approximation U_0^* is very little reduced below $||u_0||_{AC}$. Thus, although (3.17) may seem heavily dependent on α, it is not surprising to find from (3.19) that

$$\min_\gamma ||\gamma R^*\psi_j-\phi_j||^2_{AC} = \left[1-\frac{(1+\frac{1}{2}\alpha\beta)^2}{1+3\alpha^2+\frac{1}{3}\beta^2+\frac{3}{10}\alpha^2\beta^2}\right]||\phi_j||^2_{AC} \tag{3.23}$$

which depends very little on α for large β. In the limit $\beta\to\infty$, the numerical factor in (3.23) quickly approaches 1/6 for any unbounded α and is 23/38 even for $\alpha=1$.

In addition to (3.12) with $\alpha=\xi$ and (3.10), any choice of test space that reproduces the exponentially-fitted Allen & Southwell scheme has the advantage that the corresponding discrete Green's function is exactly equal to that for the continuous problem with both arguments taken at node points: that is, in an obvious notation, $G_{jk}=G(jh,kh)$. Thus for the simplest such finite difference scheme

applied to (2.48) for u_0, the nodal errors are given by

$$u_0(jh)-U_j^0 = \int_0^1 G(jh,y)f(y)dy - h\sum_{k=1}^{J-1} G(jh,kh)f(kh) \tag{3.24}$$

and are therefore wholly attributable to the trapezoidal rule applied to the integral of $G(jh,\cdot)f(\cdot)$. Similarly, for such a Petrov-Galerkin scheme the nodal errors are given by

$$u_0(jh)-U_j^0 = \sum_{k=1}^{J-1} E_{jk} \tag{3.25a}$$

where for instance for $k\geq j$,

$$E_{jk} = b^{-1}\frac{e^{\beta j_{-1}}}{e^{b/a}-1}\int_0^1 [g_k(t)-g_k(0)\psi(t)-g_k(1)\psi(t-1)]f(kh+th)dt \tag{3.25b}$$

$$g_k(t) = e^{b/a\,-\,\beta(k+t)}-1. \tag{3.25c}$$

We can assume that $\psi(0)=1$, $\psi(\pm1)=0$ so that the kernel in (3.25b) is zero at the two ends of the range and the error depends on how well $\psi(t)$ matches $\exp(-\beta t)$ between these limits, (3.10) giving the perfect match.

In variable coefficient problemsthe choice of ψ, . and in particular of the parameter α, can be made locally and similar error estimates to (3.25) derived. In two dimensions, precise error estimation and selection of ψ is considerably more difficult but the extension of (3.12) to bilinear elements on rectangles is straightforward: as in this case the trial basis functions are given by $\phi_{ij}(x,y) = \phi_i(x)\phi_j(y)$, the test functions can be taken as

$$\psi_{ij}(x,y)=[\phi_i(x)+\alpha_1\sigma_i(x)][\phi_j(y)+\alpha_2\sigma_j(y)]\ . \tag{3.26}$$

where (α_1,α_2) are chosen relative to the two components of $\underline{b}=(b_1,b_2)^T$ and the mesh spacing in the x and y directions. With quadrilaterals one can use such product functions of the isoparametric co-ordinates.

3.3 Streamline diffusion methods

As has been remarked previously, the Allen & Southwell scheme can be interpreted as having had extra diffusion added before the Galerkin method is used. Thus with a enhanced by $\frac{1}{2}\alpha bh$ piecewise linear elements reproduce (3.14) and $\alpha=\xi$ gives the Allen & Southwell scheme, but of course with ψ replaced by ϕ in any inhomogenous terms and in the error expressions (3.25). To extend this to two dimensions with a scalar diffusion would lead to excessive "cross-wind diffusion", that is normal to the direction of flow $\underline{b}$. Hughes & Brooks (1979, 1981) have therefore used in extensive computations a tensor diffusion given as follows: in (1.9) we replace $-\underline{\nabla}\cdot(a\underline{\nabla}u)$ by

$$-\underline{\nabla}\cdot(\underline{\underline{A}}\underline{\nabla}u), \quad \text{where} \quad A_{\ell m}=a\delta_{\ell m}+\tilde{a}b_\ell b_m/|\underline{b}|^2. \tag{3.27}$$

On a uniform rectangular mesh with spacings h_1, h_2 the suggested choice of the parameter $\tilde{a}$ is

$$\tilde{a} = \tfrac{1}{2}(\xi_1 b_1 h_1 + \xi_2 b_2 h_2) \tag{3.28}$$

with $\xi_m = \coth(\tfrac{1}{2}b_m h_m/a) - (\tfrac{1}{2}b_m h_m/a)^{-1}, \quad m=1,2.$

Though this choice is rather arbitrary it seems to work well in practice.

In their more recent paper, Hughes & Brooks have put this scheme into a Petrov-Galerkin framework by noting that

$$(\underline{\underline{A}}\nabla v, \underline{\nabla}\phi) = (a\underline{\nabla}v, \underline{\nabla}\phi) + (\underline{b}\cdot\underline{\nabla}v, (\tilde{a}/|\underline{b}|^2)\underline{b}\cdot\underline{\nabla}\phi) \quad . \tag{3.29}$$

Thus, assuming $\underline{\nabla}\cdot\underline{b} = 0$, the scheme is equivalent to using test functions

$$\psi_{ij} = \phi_{ij} + (\tilde{a}/|\underline{b}|^2)\underline{b}\cdot\underline{\nabla}\phi_{ij} \tag{3.30}$$

on just the convection term. For most trial spaces these functions will be discontinuous, which is quite acceptable for the convection term, but with $\psi_{ij} \notin H^1$ use of such test functions leads to consideration of so-called external approximations which is beyond the scope of these lectures. It is enough to note here, however, that if a is constant and U is bilinear then $\underline{\nabla}\cdot(a\underline{\nabla}U) = a\nabla^2 U = 0$ on each element. Hence, with the proviso that the term $(a\nabla^2 U, \underline{b}\cdot\underline{\nabla}\phi_{ij})$ is evaluated in this way the streamline diffusion method defined from (3.27) can be regarded as a Petrov-Galerkin method using test functions given by (3.30).

Alternatively, Johnson & Nävert (1981) have analysed a modification of this scheme in a way related to that followed in the next section. Starting from the reduced problem, (1.9) with $a = 0$, they use the fact that its solution also satisfies

$$(1-\delta\underline{b}\cdot\underline{\nabla})\ (\underline{b}\cdot\underline{\nabla}u+cu) = f-\delta\underline{b}\cdot\underline{\nabla}f. \tag{3.31}$$

Then the streamline diffusion method, with a modified right-hand side, is obtained by applying the Galerkin method to this equation with an appropriate choice of δ. They therefore obtain an error bound in a norm which depends on δ.

4. APPROXIMATE SYMMETRIZATION AND OPTIMAL APPROXIMATION

4.1 Motivation

The methods described in Section 3 were mainly motivated by the aim of high accuracy at nodal points or, equivalently for linear elements in one dimension, achieving a nearly optimal approximation in the $||\cdot||_{AC}$ norm. In addition, two of them involved a symmetrization of the problem: the Liouville transform did so directly; and, from the definition (2.3a) of the Riesz representer R and relation (3.11) which together imply

$$B(U,\psi_j) = (U,R^*\psi_j)_{AC} \propto (U,\phi_j)_{AC}, \tag{4.1a}$$

the use of exponential test functions leads to an approximation given by the symmetric system

$$(U,\phi_j)_{AC} = (f,R^{*-1}\phi_j) \qquad \forall\ \phi_j \epsilon S_0^h \ . \tag{4.1b}$$

The adherence to the norm $||\cdot||_{AC}$ throughout the discussion of the constant coefficient model problem was deliberate, though most authors adopt the equivalent H_0^1 norm: it emphasises the derivation of the norm from the original problem and hints at its deficiencies when $c=0$ and the singular limit $a \to 0$ is approached.

These deficiencies were apparent in the very small reduction from $||u||_{AC}$ to $||u-U^*||_{AC}$ achieved by the optimal approximation U^*. An alternative interpretation is as follows: from the optimal approximation one wants to deduce as much information as possible about u, a problem in optimal recovery (see Micchelli & Rivlin, 1976); but for a sharp exponential boundary layer as in the model problem, the point value one mesh spacing inside the boundary gives very little information. The difficulty can also be attributed to the fact that the coefficient of the dominant convection term does not appear in the norm. Thus this term has its effect only in the rather awkward exponential which appears either in the test function or in the operator R^{*-1} appearing in (4.1b).

However, when $c=0$ in the problem (1.9), the operator can be factored and a symmetrization effected in an alternative way which has been exploited by Barrett & Morton (1980, 1981). Denoting the operators $\underline{\nabla}$ and $a\underline{\nabla}-\underline{b}$ by L_1 and L_2 respectively, the operator in (1.9a) is $L_1^*L_2$ and (4.1b) is based on the identity

$$(L_1R_1v,L_1w) = (L_1v,L_1R_1^*w) = (L_2v,L_1w) \qquad \forall\ v,w\epsilon H_1, \tag{4.1c}$$

in which R_1 can be regarded as the Riesz representer of L_2 in a norm based on L_1 and defining the Hilbert space H_1. The alternative is to introduce a Riesz representer R_2^* for which

$$(L_2v,L_2R_2^*w) = (L_2v,L_1w) \qquad \forall\ v,w\epsilon H_2 \tag{4.2a}$$

and which can be regarded as the Riesz representer of L_1 in a norm based on L_2 and defining a Hilbert space H_2: then an approximation U is generated from a test space $T_0^h=R_2^{*-1}S_0^h$ giving

$$(L_2U,L_2\phi_j) = (f,R_2^{*-1}\phi_j) \qquad \forall \phi_j\epsilon S_0^h \ . \tag{4.2b}$$

Clearly if L_1w in (4.1c) spanned the same space as L_2v in (4.2a), we should have $R_2^*=R_1^{-1}$, $R_2^{*-1}=R_1$ so that in the model problem with R_1 given by (2.49) no exponentials would be involved. Unfortunately, such a relation does not hold exactly and some approximation is involved in aiming for optimality in the norm based on L_2 without the use of exponentials. In this section we consider how this is done first for problems in one dimension and then for those in two.

4.2 One dimensional problems

Consider the variable coefficient Dirichlet problem for u:

$$-(au')' + (bu)' = f \quad \text{on } (0,1) \tag{4.3a}$$

$$u(0) = g_L, \; u(1) = g_R . \tag{4.3b}$$

Then, from (2.25), we have for $v \in H^1$, $w \in H^1_0$

$$B(v,w) = (av',w') + ((bv)',w) = (av'-bv,w') . \tag{4.4}$$

We introduce a symmetric form with an arbitrary positive weighting function $\rho(x)$:

$$B_s(v,w) := (\rho a^2 v',w') + ([\rho b^2 + (\rho ab)']v,w) \tag{4.5a}$$

$$= (av' - bv, \rho[aw'-bw]) \qquad \forall\, v \in H^1,\ w \in H^1_0 . \tag{4.5b}$$

In addition to the usual assumptions on a and b, as in (2.23), we assume that ρ is normalised to have unit integral and is chosen so that on (0,1) we have:

$$\text{(i)} \quad p(x) := \rho(x)a^2(x) > 0, \tag{4.6a}$$

$$\text{(ii)} \quad q(x) := \rho(x)b^2(x) + (\rho ab)'(x) \geq 0, \tag{4.6b}$$

$$\text{(iii)} \quad \alpha(x) := \rho(x)b(x) + (\rho a)'(x) \Rightarrow \alpha(x)b(x) \geq 0. \tag{4.6c}$$

This is easily achieved by, for instance, taking $(\rho a)' = 0$ where $b' \geq 0$ and $(\rho ab)' = 0$ where $b' < 0$; then $q \geq \rho b^2$ and $\alpha b \geq \rho b^2$. These assumptions ensure that $B_s(\cdot,\cdot)$ is a coercive form and that if $B(\cdot,\cdot)$ is coercive relative to the $\|\cdot\|_{AC}$ norm then it is also coercive relative to $\|\cdot\|_s^2 := B_s(\cdot,\cdot)$.

Establishing the coercivity of $B(\cdot,\cdot)$ through (2.6) would require us to assume that there exists a $\delta > 0$ such that

$$(1-\delta)\int_0^1 av'^2 dx + \tfrac{1}{2}\int b'v^2 dx \geq 0 \qquad \forall\, v \in H^1_0. \tag{4.7}$$

However, we shall see in a moment that (2.1b,c) can be satisfied under much weaker conditions. Then we can apply either the Lax-Milgram lemma or Theorem 2.1 in respect of H_s, the Hilbert space formed from H^1_0 equipped with the $\|\cdot\|_s$ norm, and, retaining the notation of Barrett & Morton (1981), introduce a symmetrizing operator $N: H^1_0 \to H^1_0$ such that

$$B(v,Nw) = B_s(v,w) \qquad \forall\ v,w \in H^1_0 . \tag{4.8}$$

Indeed, it is not too difficult to construct N explicitly: we require from (4.4) and (4.5b) that

$$\int (av'-bv)\,[(Nw)' - \rho(aw' - bw)]dx = 0 \qquad \forall\ v \in H^1_0$$

and introduce $z = e^{-\lambda} v \in H^1_0$, where

$$\lambda(x) = \int_0^x (b/a)dt, \tag{4.9}$$

so that $av'-bv = az'e^{\lambda}$; then, as with (2.49), we have

$$(Nw)' = \rho(aw' - bw) + \text{const. } e^{-\lambda}/a \tag{4.10a}$$

from which a little manipulation gives

$$(Nw)(x) = (\rho aw)(x) + \int_x^1 (\alpha w - Ke^{-\lambda}/a)dy \tag{4.10b}$$

and the constant K is such that $(Nw)(0) = 0$. We have given the form which is appropriate for $\lambda > 0$, or $b \geq 0$, and it is also useful to note that N^+ its adjoint in the L_2 inner product is given by

$$(N^+f)(x) = (\rho af)(x) + \alpha(x)[F(x) - \tilde{F}] , \tag{4.11a}$$

where $F(x) := \int_0^x f(y)dy$

and $\tilde{F} \int_0^1 (e^{-\lambda}/a)dx := \int_0^1 (e^{-\lambda}/a)F dx.$ (4.11b)

It is clear that N and N^+ involve an exponential kernel $e^{-\lambda}/a$ unless $\alpha \equiv 0$, which would require instead that ρ be proportional to the same exponential kernel. It is also clear from this construction why only the positivity of a is necessary to establish the hypotheses of Theorem (2.1): for if in $B(v,w)$ we set $w = e^{-\lambda}v$ we have

$$\left\{\sup_w \text{ or } \sup_v\right\} B(v,w) \geq B(e^{\lambda}w,w) = \int_0^1 ae^{\lambda}(w')^2 dx. \tag{4.12}$$

An optimal approximation to u in the norm $||\cdot||_s$ can now be constructed using the symmetrizing operator N. If the trial space S^h is spanned by $\{\phi_j\}$, taking the test space as $T_0^h = NS_0^h$ in a Petrov-Galerkin method gives $U^* \epsilon S_E^h$ such that

$$B(U^*, N\phi_j) = (f, N\phi_j) \qquad \forall\ \phi_j \epsilon S_0^h. \tag{4.13}$$

Subtracting from a similar equation for u and using (4.8) establishes the optimality of U^*,

$$B_s(u - U^*, \phi_j) = 0 \qquad \forall\ \phi_j \epsilon S_0^h . \tag{4.14}$$

It is important to note too that the discrete equations for U^* only involve the operation of N and N^+ on the data and the test functions never need to be obtained explicitly: if ϕ_0 and ϕ_J are the basis functions corresponding to the data g_L and g_R on the left and right respectively, we have from (4.12) and (4.8)

$$B_s(U^* - g_L\phi_0 - g_R\phi_J, \phi_j) = (N^+f, \phi_j) - B(g_L\phi_0 + g_R\phi_J, N\phi_j) \qquad \forall\ \phi_j \epsilon S_0^h . \tag{4.15a}$$

This in turn can be reduced by (4.10) to

$$B_s(U^*, \phi_j) = (N^+f, \phi_j) - \alpha_j \ell_\lambda(g_L\phi_0 + g_R\phi_J) \qquad \forall\ \phi_j \epsilon S_0^h . \tag{4.15b}$$

where $$\alpha_j := \int_0^1 \alpha\phi_j dx \tag{4.16}$$

and $$\ell_\lambda(w) \int_0^1 (e^{-\lambda}/a)dx := \int_0^1 (e^{-\lambda}/a)(aw' - bw)dx. \tag{4.17}$$

In this form we can see that the exponential kernel is involved only in the calculation

of the averages $\tilde{F}$, $\ell_\lambda(\phi_0)$ and $\ell_\lambda(\phi_j)$. We can also regard the equation for U^* as obtained by operating on (4.3) with the symmetrizing operator N^+ and then using the Galerkin method: this can then be compared with the streamline diffusion method in the form (3.31).

In their consideration of one dimensional problems, Barrett & Morton (1980, 1981) eschew exponentials completely by approximating the averages (4.11b) and (4.17) by a weighting function $\varepsilon(x)$, normalised to unit integral, or a delta-function at $x = 0$: they denote the corresponding operator (4.10) by N_ε or N_0 and the corresponding linear functional (4.17) by ℓ_ε or ℓ_0. In the former case $N_\varepsilon : H_0^1 \to H_0^1$ gives a proper Petrov-Galerkin method with $T_0^h = N_\varepsilon S_0^h$: but the symmetrization (4.8) is not exactly achieved and instead of (4.14) we have for the approximation U^N,

$$B_s(U^N,\phi_j) + \alpha_j \ell_\varepsilon(U^N) = (N_\varepsilon^+ f,\phi_j) \qquad \forall\, \phi_j \in S_0^h \,. \tag{4.18}$$

When the delta function is used N_0 is defined by (4.10) with $K = 0$, for no value will ensure that $(N_0 w)(0) = 0$ if $w(0) = 0$, and N_0^+ by (4.11a) with $\tilde{F} = 0$. Thus the resulting method is not strictly of Petrov-Galerkin form but the approximation still satisfies equation (4.18), with ℓ_ε and N_ε^+ replaced by ℓ_0 and N_0^+, and indeed is the simplest to use and the most appropriate in the singular limit $b/a \to \infty$.

Introducing $V^* \in S_0^h$ such that

$$B_s(V^*,\phi_j) = \alpha_j \qquad \forall\, \phi_j \in S_0^h \,, \tag{4.19}$$

Barrett & Morton (1981) show that for a problem with no turning points, $b(x) > 0$, U^N is uniquely determined if ε is chosen to ensure that $1+\ell_\varepsilon(V^*) \neq 0$ and

$$\begin{aligned} ||u-U^N||_s^2 &= ||u-U^*||_s^2 + ||U^*-U^N||_s^2 \\ &= ||u-U^*||_s^2 + \left[\frac{||V^*||_s}{1+\ell_\varepsilon(V^*)}\right]^2 [\ell_\varepsilon(u-U^*)]^2 \,. \end{aligned} \tag{4.20}$$

This estimate also holds for problems with a homogeneous Neumann condition at $x = 1$ but for the Dirichlet problem it is easy to show that $||V^*||_s \leq 1$. The same result holds using N_0 and ℓ_0 and in that case the authors show that under quite general conditions $\ell_0(V^*) \geq 0$. Precise error bounds may then be derived: thus we have

$$||u-U^N||_s^2 \leq ||u-U^*||_s^2 + e^2(0)[u'(0) - U^{*\prime}(0)]^2 \tag{4.21}$$

and, for example with constant coefficients and linear elements on a uniform mesh,

$$|U_j^N - U_j^*| \leq \frac{3}{2}\frac{a}{b}\left|\frac{U_1^N - U_0^N}{h} - \frac{f(0)}{b}\right| + \frac{3}{2b}\left|\int_0^1 e^{-bx/a}\,[f(x)-f(0)]dx\right| + O(e^{-b/a}). \tag{4.22}$$

Similar results are given for variable coefficient problems. They show that when b/a or $\lambda(1)$ is large and $f(x)$ is nearly constant near $x = 0$ then U^N is very close to U^*, the optimal approximation in $||\cdot||_s$. Even some turning-point problems

can be approximated well in this way. For a single turning point at ξ with $b(x) \le 0$ for $x \le \xi$ and $b(x) \ge 0$ for $x \ge \xi$, the delta function is placed at ξ and comparable error bounds obtained.

To compare these methods with other Petrov-Galerkin methods, consider the constant coefficient problem with ρ also taken constant: then from (4.10) the test functions are given by

$$\psi_j(x) = (N_\varepsilon \phi_j)(x) = a\phi_j(x) + b\int_x^1 (\phi_j - \bar{\phi}_j \varepsilon)dy. \tag{4.23}$$

These are not localised functions though linear combinations of successive pairs can be localised. The choice $\varepsilon(x) \equiv 1$ corresponds to the H^{-1} least squares formulation of Bristeau et al. (1980) which, as Barrett (1980) shows, is not very accurate for the simple model problem when bh/a is moderately large. On the other hand, the localised upwind test functions (3.12) can also be related to (4.23): we have already noted that $\sigma' = 6(\bar{\phi} - \phi)$, so that these test functions can be written as

$$\psi_j(x) = \phi_j(x) + (6\alpha/h)\int_x^1 (\phi_j - \bar{\phi}_j/2h)dy. \tag{4.24}$$

That is, they correspond to taking a different weighting function for each ϕ_j, equal to $(1/2h)$ over the support of ϕ_j, if the parameter α is taken as $\alpha = \beta/6$

To conclude this section we consider a numerical example together with the resulting recovery problem. The example, taken from Barrett (1980), is for

$$-10^{-3}u'' + [(1.0-0.98x)u]' = 0, \qquad u(0) = 1 \tag{4.25}$$

with either the Neumann condition $u'(1) = 0$ or the Dirichlet condition $u(1) = 49.95$ which gives the same solution. The table below gives various approximations for each case using piecewise linear elements on a mesh with $h = 0.1$; only the last three nodal values are given. The Galerkin approximation is U^G and U^U is the Petrov-

	Dirichlet case			Neumann case		
Node j	8	9	10	8	9	10
u(jh)	4.73	9.31	49.95	4.73	9.31	49.95
U^G_j	5.73	18.41	"	3.85	1.76	25.34
U^U_j	4.70	5.71	"	4.70	5.70	50.03
U^E_j	3.57	5.35	"	3.57	5.35	12.64
U^N_j	4.75	7.01	"	4.85	6.77	45.29
$u_R(jh)$	4.57	9.83	"	4.85	8.60	50.03

TABLE Results from approximating (4.25) with linear elements.

Galerkin result using the upwind scheme of Heinrich et al. (1977), that is with test functions (3.12), and α given the nodal values of $\alpha_{crit} = 1-2/\beta$. The row labelled U^E corresponds to test functions $\psi_j = \phi_j e^{-\lambda}$, the case of exact symmetrization obtained from (4.10b) with $\alpha = 0$, $\rho \propto e^{-\lambda}$: this also corresponds to using an exponential weighting in the inner product as in Axelsson (1981). The penultimate row gives U^N obtained from (4.18) using N_0 and $\rho \equiv 1$ in the Dirichlet case but $\rho(x) = (1.0-0.98x)^{-1}$ in the Neumann case in order to give more weight to the right of the interval.

None of these nodal values is particularly accurate, except $U^U(1)$ in the Neumann case which results from U^U satisfying a simple flux conservation relation. However U^N does not purport to have accurate nodal values: it aims instead at being a nearly optimal fit to u in the $||\cdot||_s$ norm, which is close to a weighted L_2 norm in this case. This optimality property can be combined with any further a priori knowledge of u, such as smoothness, monotonicity, positivity, etc., to give more accurate estimates for u. For the present problem, in the Dirichlet case, we expect u to be well approximated by an exponential in the boundary layer near $x = 1$, of the form

$$u_R(x) = \lambda_1 e^{\lambda_2(x-1)} + \lambda_3 \tag{4.26}$$

for some constants λ_1, λ_2 and λ_3. One equation for determining these parameters is provided by the boundary condition $u_R(1) = 49.95$ and the other two can be obtained by assuming that locally U^N is a best fit to u_R and hence

$$B_s(u_R - U^N, \phi_j) = 0 \qquad j = J-2, J-1. \tag{4.27}$$

The result of this procedure is given in the table: it can be seen that the value for $u(0.9)$ is accurate to 5%; and the boundary layer half-width can be predicted as 0.0394 as compared with the exact value 0.0447.

In the Neumann case, in order to satisfy the boundary condition we take

$$u_R(x) = \lambda_1 e^{\lambda_2(x-1)^2} + \lambda_3. \tag{4.28}$$

One equation for the parameters is obtained from integrating (4.25) over $(x,1)$ to obtain a relation of the form

$$-au' + bu = \text{const.} = b(1)u(1), \tag{4.29}$$

and others can be obtained from equations of the form (4.27), including $j = J$. However, if only $u(1)$ is required, substitution from (4.29) directly into $B_s(u-U^N, \phi_J) = 0$ provides a good approximation: this is the value given in the table, together with interior values obtained for (4.28) and (4.27).

4.3 Two-dimensional problems

Before extending these techniques of approximate symmetrization to two dimensions, it is useful to place them more carefully in the abstract framework of Section 2. Leaving aside for the moment the case when $\varepsilon(x)$ is a delta function, we can work in H_s, that is H_0^1 equipped with the $||\cdot||_s$ norm, and define N_ε so that $T_0^h = N_\varepsilon S_0^h \subset H_s$. For the approximation U^N given by

$$B(U^N,W) = (f,W) = B(u,W) \qquad \forall\ W\in T_0^h\ , \tag{4.30}$$

we have, from defining R^* as in (2.3a),

$$B_s(u-U^N,R^*W) \equiv B(u-U^N,W) = 0 \qquad \forall\ W\in T_0^h\ . \tag{4.31}$$

Suppose now that the constant $\Delta\in[0,1)$ is such that

$$\inf_{W\in T_0^h} ||V-R^*W||_s \le \Delta\ ||V||_s \qquad \forall\ V\in S_0^h\ . \tag{4.32}$$

Then with U^* given by (4.14) and repeating the argument following (2.17), we have

$$\begin{aligned} ||U^*-U^N||_s^2 &= B_s(u-U^N,U^*-U^N) \\ &= B_s(u-U^N,U^*-U^N-R^*W) \qquad \forall\ W\in T_0^h \\ &\le \Delta||u-U^N||_s||U^*-U^N||_s\ . \end{aligned} \tag{4.33}$$

Thus from $||u-U^N||_s^2 = ||u-U^*||_s^2 + ||U^*-U^N||_s^2$ we obtain

$$||u-U^N||_s \le (1-\Delta^2)^{-\frac{1}{2}}||u-U^*||_s\ . \tag{4.34}$$

It is clear that a good approximation is obtained if in particular R^*N_ε is close to the identity: that is, comparing (4.31) to (4.8), we take N_ε to approximate $R^{*-1} = N$.

When using N_0 by taking $\varepsilon(x)$ as a delta function, we generate test functions W which for non-turning point problems are not in H_0^1 and for turning point problems may not even be in H^1: in two dimensions the corresponding test functions would not be zero on the inflow Dirichlet boundary $\Gamma_D\cap\Gamma^-$. Though it may be possible to extend the definition of R^* to such functions and hence to establish an approximation result like (4.32), it is difficult to do this so as to maintain (4.31). For $B(v,w)$ to be defined by (2.25) so that (2.24) holds one needs to have $\nabla^2 v\in L_2(\Omega)$: thus as a minimum in (4.31) one needs to assume greater smoothness on u; and to define R^* by (4.31) using the Riesz representation theorem one needs to work in smoother spaces than H_s. We shall therefore regard the approximation derived using N_0 as a limiting case of those obtained from N_ε. In the one dimensional problems treated above this mainly required establishing a uniform bound on $||K^{-1}||$ as $\ell_\varepsilon \to \ell_0$, where K is the stiffness matrix for the system in (4.18).

We consider then two dimensional problems covered by Theorem 2.3, with the added restriction that $c \equiv 0$, and in the space H_s with the definition of $B_s(\cdot,\cdot)$ extended from (4.5) to

$$B_s(v,w) := (\rho a^2 \underline{\nabla} v, \nabla w) + ([\rho b^2 + \underline{\nabla}\cdot(\rho a \underline{b})]v, w) \tag{4.35a}$$

$$= (a\underline{\nabla} v - \underline{b}v,\ \rho(a\underline{\nabla} w - \underline{b}w)) + (\underline{n}\cdot\underline{b}v,\ \rho a w)_{\Gamma_N} \quad \forall\ v\in H^1, w\in H^1_0. \tag{4.35b}$$

The assumptions made in (4.6) are generalised in a natural way with $\underline{\alpha} := \rho\underline{b} + \underline{\nabla}(\rho a)$ and $\underline{\alpha}\cdot\underline{b} \geq 0$. Then if all these coefficients are sufficiently smooth $B(\cdot,\cdot)$ satisfies the hypotheses of Theorem 2.1 in $H_s \times H_s$. Thus there exists a symmetrizing operator N satisfying (4.8) which requires explicitly that

$$(a\underline{\nabla} v - \underline{b}v,\ [\underline{\nabla}(Nw) - \rho(a\underline{\nabla} w - \underline{b}w)]) + (\underline{n}\cdot\underline{b}v,\ [Nw - \rho a w])_{\Gamma_N} = 0 \qquad \forall\ v\in H^1_0 . \tag{4.36}$$

In case that the vector field $\underline{b}/a$ is irrotational a scalar λ can be introduced, as in (4.9), such that $\underline{\nabla}\lambda = \underline{b}/a$ and with $\lambda = 0$ at some inflow point. Then introducing $z = e^{-\lambda}v$ and using the divergence theorem, the problem for Nw becomes

$$\underline{\nabla}\cdot(ae^{\lambda}\underline{\nabla}(Nw)) = \underline{\nabla}\cdot[a\rho e^{\lambda}(a\underline{\nabla} w - \underline{b}w)] \quad \text{in } \Omega \tag{4.37a}$$

$$a\frac{\partial}{\partial n}(Nw) + \underline{n}\cdot\underline{b}(Nw) = \rho a^2 \frac{\partial w}{\partial n} \quad \text{on } \Gamma_N , \quad Nw = 0 \quad \text{on } \Gamma_D . \tag{4.37b}$$

This is not particularly useful as a starting point for approximating $N\phi_j$ to obtain test functions in a Petrov-Galerkin method. One could introduce a stream function $\psi(x,y)$ to take the place of the constant in (4.10) and for which one would then have

$$\partial_x(Nw) = \rho(a\partial_x w - b_1 u) + (e^{-\lambda}/a)\partial_y\psi \tag{4.38a}$$

$$\partial_y(Nw) = \rho(a\partial_y w - b_2 u) - (e^{-\lambda}/a)\partial_x\psi , \tag{4.38b}$$

with the boundary conditions on ψ obtained from (4.37b). However, such an approach has not so far been followed up directly, though it does motivate one of the two approaches that have been used.

This, the most direct extension of the one-dimensional technique, was reported in Morton & Barrett (1980). As with the upwind method based on (3.26), it uses bilinear elements on rectangles and correspondingly generates test functions given by

$$(N_\varepsilon\phi_{ij})(x,y) = (N_\varepsilon^{(x)}\phi_i)(x).(N_\varepsilon^{(y)}\phi_j)(y), \tag{4.39}$$

where $N_\varepsilon^{(x)}$ is as in (4.23) but based on $a(x,y_j)$ and $b_1(x,y_j)$ with $N_\varepsilon^{(y)}$ defined similarly. The results for standard test problems in which $\underline{b}$ has a fixed direction are quite good but the method is not adequate when $\underline{b}$ corresponds to very curved flow lines.

Morton/Finite Element Methods

An alternative approach has been given in Barrett & Morton (1981). Corresponding to (4.14), (4.15) the exact solution u is easily seen to satisfy

$$B_s(u,\phi_j) = (\rho af,\phi_j) + (\underline{b}u - a\underline{\nabla}u, \underline{\alpha}\phi_j) \qquad \forall\ \phi_j \in S_0^h . \tag{4.40}$$

Thus suppose we introduce the flux function $\underline{v} = \underline{b}u - a\underline{\nabla}u$ and approximate it by $\underline{V}$. Then an approximation $U \in S_E^h$ to u can be obtained from

$$B_s(U,\phi_j) = (\rho af,\phi_j) + (\underline{V}, \underline{\alpha}\phi_j) \qquad \forall\ \phi_j \in S_0^h \tag{4.41}$$

and any approximation scheme for $\underline{V}$. Defining U^* now by (4.14) one obtains

$$B_s(U^*-U,\phi_j) = (\underline{\alpha}\cdot(\underline{v}-\underline{V}),\phi_j) \qquad \forall\ \phi_j \in S_0^h . \tag{4.42}$$

Moreover, if we define s as $\underline{\alpha}\cdot\underline{v}$ and denote by S^* its best L_2 fit in S^h such that $s = S^*$ on Γ_D, then

$$B_s(U^*-U,\phi_j) = (S^*-\underline{\alpha}\cdot\underline{V},\phi_j) \qquad \forall\ \phi_j \in S_0^h$$

and hence

$$||U^*-U||_s^2 \le |||\underline{\alpha}|^{-1}S^* - \hat{\underline{\alpha}}\cdot\underline{V}|| \cdot |||\underline{\alpha}|(U^*-U)|| ,$$

where $\hat{\underline{\alpha}} = \underline{\alpha}/|\underline{\alpha}|$. Introducing the constant γ such that

$$|\rho\underline{b} + \underline{\nabla}(\rho a)|^2 \le \gamma[\rho b^2 + \underline{\nabla}\cdot(\rho\underline{b}a)] \tag{4.43}$$

we obtain the following relationship between the deviation of U and V from their "optimal" approximations:

$$||U^*-U||_s \le (\gamma/|\underline{\alpha}|)||S^* - \underline{\alpha}\cdot\underline{V}|| . \tag{4.44}$$

Thus $\underline{V}$ should be constructed by approximating the equation $\underline{\nabla}\cdot\underline{v} = f$ in such a way that $\underline{\alpha}\cdot\underline{V}$ is close to S^*: boundary conditions can be obtained by setting $\underline{V} = \underline{b}U - a\underline{\nabla}U$ on the inflow boundary and U and $\underline{V}$ obtained from an alternating iterative procedure. One such scheme was given in Barrett & Morton (1981) but what is the most effective scheme is not yet clear.

To conclude, we believe that some sort of symmetrization is the most useful basic approach to approximating non-self-adjoint problems by finite element methods. Other more economic methods of adequate accuracy may then be derived from these. In Section 3 use of the exponential test functions turned out to be such a basic approach, based on the symmetric part of the operator L being used to define the norm $||\cdot||_{AC}$: then the Allen & Southwell difference operator can be regarded as a practical shortcut to forming the stiffness matrix and various other upwinded test functions as approximating the Green's function in (3.25) in order to model the effect of the inhomogeneous data f. In this last section a natural alternative symmetrization based on the norm $||\cdot||_s$ has been presented. It gives quite a different type of approximation, much

less closely linked to finite difference methods, and can yield so-called sub-gridscale information. The further development of these two approaches for two-dimensional problems should show which is the more useful.

It is a pleasure to acknowledge the valuable discussions with Dr. J.W. Barrett that have taken place during and prior to the preparation of these lecture notes.

5. REFERENCES

[1] Agmon, S., Douglis, A., and Nirenberg L, 1964. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions II. Comm. Pure Appl. Math, 17, 35-92.

[2] Allen, D., and Southwell, R., 1955. Relaxation methods applied to determine the the motion, in two dimensions, of a viscous fluid past a fixed cylinder. Quart. J. Mech. and Appl. Math., VIII, 129-145.

[3] Axelsson, O., 1981. Stability and error estimates of Galerkin finite element approximations for convection-diffusion equations. I.M.A. J. Numer. Anal. 1, 329-345.

[4] Babuŝka, I., and Aziz, A.K., 1972. Survey lectures on the mathematical foundations of the finite element method. The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (ed. A.K. Aziz), New York: Academic Press, 3-363.

[5] Barrett, J.W., 1980. Optimal Petrov-Galerkin methods. Ph.D. Thesis, University of Reading.

[6] Barrett, J.W. and Morton, K.W., 1980. Optimal finite element solutions to diffusion-convection problems in one dimension. Int. J. Num. Meth. Engng. 15, 1457-1474.

[7] Barrett, J.W., and Morton, K.W., 1981. Optimal Petrov-Galerkin methods through approximate symmetrization. I.M.A. J. Numer. Anal. 1, 439-468.

[8] Barrett, J.W., and Morton, K.W., 1981. Optimal finite element approximation for diffusion-convection problems. Proc. MAFELAP 1981 Conf. (ed. J.R. Whiteman).

[9] Barrett, K.E., 1974. The numerical solution of singular-perturbation boundary-value problems. J. Mech. Appl. Math., 27, 57-68.

[10] Barrett, K.E., 1977. Finite element analysis for flow between rotating discs using exponentially weighted basis functions. Int. J. Num. Meth. Engng., 11, 1809-1817.

[11] de Boor, C., and Swartz, B., 1973. Collocation at Gaussian points. SIAM J. Numer. Anal., 10, 582-606.

[12] Bristeau, M.O., Pirronneau, O., Glowinski, R., Periaux, J., Perrier, P., and Poirier, G., 1980. Application of optimal control and finite element methods to the calculation of transonic flows and incompressible viscous flows. I.M.A. Conf. Numerical Methods in Applied Fluid Dynamics (ed. B. Hunt), Academic Press, 203-312.

[13] Christie, I., Griffiths, D.F., Mitchell, A.R., and Zienkiewicz, O.C., 1976. Finite element methods for second order differential equations with significant first derivatives. Int. J. Num. Meth, Engng., 10, 1389-1396.

[14] Ciarlet, P.G., 1978. The Finite Element Method for Elliptic Problems. North-Holland (Amsterdam).

[15] Doolan, E.P., Miller, J.J.H., and Schilders, W.H.A., 1980. Uniform Numerical Methods for Problems with Initial and Boundary Layers. Dublin: Boole Press.

[16] Douglas, J., Jr., and Dupont, T., 1973. Superconvergence for Galerkin methods for the two point boundary problem via local projections. Numer. Math., 21, 270-278.

[17] Gresho, P.M., and Lee, R.L., 1979. Don't suppress the wiggles - they're telling you something. Finite Element Methods for Convection Dominated Flows (ed. T.J.R. Hughes) AMD Vol. 34, Am. Soc. Mech. Eng., 37-61.

[18] Griffiths, D., and Lorenz, J, 1978. An analysis of the Petrov-Galerkin finite element method. Comp. Meth. Appl. Mech. Engng., 14, 39-64.

[19] Guymon, G.L., Scott, V.H., and Herrmann, L.R., 1970. A general numerical solution of the two-dimensional diffusion-convection equation by the finite element method. Water Resources 6, 1611-1617.

[20] Guymon, G.L., 1970. A finite element solution of a one-dimensional diffusion-convection equation. Water Resources 6, 204-210.

[21] Heinrich, J.C., Huyakorn, P.S., Mitchell, A.R., and Zienkiewicz, O.C., 1977. An upwind finite element scheme for two-dimensional convective transport equations. Int. J. Num. Meth. Engng., 11, 131-143.

[22] Heinrich , J.C., and Zienkiewicz, O.C., 1979. The finite element method and 'upwinding' techniques in the numerical solution of convection dominated flow problems. Finite Element Methods for Convection Dominated Flows (ed. T.J.R. Hughes) AMD Vol. 34, Am. Soc. Mech. Eng., 105-136.

[23] Hemker, P.W., 1977. A numerical study of stiff two-point boundary problems. Thesis, Amsterdam: Math. Cent.

[24] Hughes, T.J.R., and Brooks, A., 1979. A multidimensional upwind scheme with no crosswind diffusion. Finite Element Methods for Convection Dominated Flows (ed. T.J.R. Hughes) AMD Vol. 34, Am. Soc. Mech. Eng., 19-35.

[25] Hughes, T.J.R., and Brooks, A., 1981. A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions: application to the streamline-upwind procedure. To appear in Finite Elements in Fluids Vol. 4 (ed. R.H. Gallagher), J. Wiley & Sons : New York.

[26] Johnson, C., and Nävert, U., 1981. An analysis of some finite element methods for advection-diffusion problems. Conf. on Analytical and Numerical Approaches to Asymptotic Problems in Analysis (eds. O. Axelsson, L.S. Frank and A. van der Sluis), North-Holland.

[27] Lesaint, P., and Zlamal, M., 1979. Superconvergence of the gradient of finite element solutions. R.A.I.R.O. Numer. Anal., 13, 139-166.

[28] Micchelli, C.A., and Rivlin, T.J., 1976. A survey of optimal recovery. Optimal Estimation in Approximation Theory (ed. C.A. Micchelli & T.J. Rivlin), Plenam Press : New York.

[29] Morton, K.W., and Barrett, J.W., 1980. Optimal finite element methods for diffusion-convection problems. Proc. Conf. Boundary and Interior Layers - Computational and Asymptotic Methods (ed. J.J.H. Miller), Boole Press : Dublin, 134-148.

[30] Oden, J.T., and Reddy, J.N., 1976. An introduction to the Mathematical Theory of Finite Elements, Wiley-Interscience : New York.

[31] Strang, G., and Fix, G.J., 1973. An Analysis of the Finite Element Method, Prentice Hall : New York.

[32] Thomée, V, and Westergren, B., 1968. Elliptic difference equations and interior regularity. Numer. Math. II, 196-210.

[33] Zienkiewicz, O.C., Gallagher, R.H. and Hood, P., 1975. Newtonian and non-Newtonian viscous incompressible flow, temperature induced flows : finite element solution. 2nd Conf. Mathematics of Finite Elements and Applications (ed. J.R. Whiteman), London : Academic Press.

[34] Zienkiewicz, O.C., 1977. The Finite Element Method, London : McGraw Hill.

[35] Zlamal, M., 1977. Some superconvergence results in the finite element method. Mathematical Aspects of Finite Element Methods. Springer-Verlag.

[36] Zlamal, M., 1978. Superconvergence and reduced integration in the finite element method. Math. Comp. 32, 663-685.