COMPUTER METHODS IN APPLIED MECHANICS AND ENGINEERING 14 (1978) 39–64

AN ANALYSIS OF THE PETROV–GALERKIN FINITE ELEMENT METHOD

D.F. GRIFFITHS *
Department of Mathematics, The University of Dundee, Scotland

and

J. LORENZ *
Institut für Numerische und instrumentelle Mathematik, Universität Münster, Federal Republic of Germany

Manuscript received 23 May 1977

Petrov–Galerkin finite element methods, using different test and trial functions, are applied to the solution of an unsymmetric two-point boundary value problem intended to simulate certain aspects of convection-diffusion problems. For a specified space of trial functions we utilise an energy error bound to optimize this class of methods over a family of test spaces. The optimized method performs well provided the asymmetry in the differential operator does not lead to boundary layers in the solution. Following an analysis of the boundary layer behaviour of the continuous problem, L- splines are introduced, and, by studying their behaviour for coarse meshes, we are able to modify the original schemes to produce so-called "disconnected" finite element methods. Even for coarse meshes, when no nodes occur in the boundary layer, the accuracy at all nodal points is good. This would make them good candidates for application in more general situations.

1. Introduction

The current interest in the application of finite element methods to fluid dynamics has uncovered a number of challenging problems. Notable among these are boundary layer or convective-diffusion type flows governed by second order elliptic partial differential equations in which the coefficients of first derivatives become large relative to those of higher derivatives. Conventional Galerkin finite element methods applied to such problems result in physically unrealistic oscillatory solutions unless the mesh length h is excessively small (Lee et al. [1], Zienkiewicz [2]). This phenomenon has been well-known to finite difference practitioners for some time, and a comprehensive discussion may be found in Roache [3]. A popular finite difference method approximates first derivatives by first order one-sided differences, a process often referred to as "upwinding". Although this removes unwanted oscillations from the numerical solutions, the accuracy attained is often poor. In an earlier paper (Christie et al. [4]) it was shown how finite element schemes could be constructed so as to reproduce such finite difference methods. The construction was based on a Galerkin method using different test and trial functions; we shall follow Anderson and Mitchell [5] in calling these Petrov–Galerkin methods. In this paper we undertake a detailed analysis of the results of this earlier paper, from a new point of view, by utilizing a general error estimate established by Babuska and Aziz (Aziz [5]).

Following the introduction of a model problem in section 2, and the general Petrov–Galerkin

* Work performed whilst the authors were on leave at the University of Calgary, Alberta, Canada.

method in section 3, we then apply the Babuska–Aziz estimate. In section 4 we show that for a specified space of trial functions we can optimize the Petrov–Galerkin method with respect to a given family of test spaces. The L_2-convergence of such methods is discussed in section 5 by means of the well-known Aubin–Nitsche duality argument. After relating the Petrov–Galerkin method to finite difference schemes in section 6, sample numerical results are presented in section 7. These indicate that, although the Petrov-Galerkin method provides oscillation-free solutions, the accuracy attained, as for most common numerical methods is severely limited when the exact solution has a pronounced boundary layer behaviour. Accordingly, we devote part 2 of this paper to an analysis of this situation. In section 8 the continuous form of the model problem is studied and then, through the use of L-splines in section 9 we are able to identify specific forms of the Petrov–Galerkin method which behave well in regard to boundary layer problems. These methods suffer from being difficult to generalize to more complex problems, and so in section 10, we study modifications which, whilst retaining the superior properties of methods generated by L-splines, are relatively straightforward to implement and generalize; this leads to so-called "disconnected" Petrov–Gelerkin methods. The numerical results of section 11 show that these methods lead to significant reductions in the nodal errors even when the grid spacing is so coarse as not to place any nodes within the boundary layer.

Part 1

2. Model problem

In order that we may give a detailed analysis, our attention is devoted to a model problem similar to those discussed by Christie et al. [4] and Roache [3]:

$$Lu \equiv -D^2 u + k\,Du = f(x)\,, \qquad x \in (0,\,1)\,, \tag{2.1}$$

$$u(0) = u(1) = 0\,, \tag{2.2}$$

where $D \equiv \mathrm{d}/\mathrm{d}x$, $f \in L^2[0,1]$ and k is a positive constant. In later sections we shall assume $k \gg 1$.

We denote by $H^m[0,1]$ ($\overset{0}{H}{}^m[0,1]$) the completion of $C^\infty[0,1]$ ($C_0^\infty[0,1]$) with respect to the norm

$$\|u\|_m = \left\{ \int_0^1 \sum_{s \leqslant m} (D^s u)^2 \,\mathrm{d}x \right\}^{1/2}. \tag{2.3}$$

Recall that, for functions $u \in \overset{0}{H}{}^1[0,1]$, the norm (2.3) with $m = 1$ is equivalent to the seminorm

$$|u|_1 = \left\{ \int_0^1 (Du)^2 \,\mathrm{d}x \right\}^{1/2}. \tag{2.4}$$

The linear product and norm on $L^2[0,1]$ will be written $(\cdot,\cdot)$ and $\|\cdot\|_0$ respectively and C will denote a generic constant that is not necessarily the same at successive appearances. Associated with the operator L of (2.1), we have the unsymmetric bilinear form

$$a(u, v) = (u', v' + kv) . \tag{2.5}$$

With these notations the weak form of (2.1), (2.2) may be written:

$$\text{Find} \quad u \in \overset{0}{H}{}^1[0,1]$$

$$\text{so that} \quad a(u, v) = (f, v) \quad \forall v \in \overset{0}{H}{}^1[0,1] . \tag{2.6}$$

Existence and uniqueness of solutions to (2.6) follow from the generalized Lax–Milgram theorem due to Babuska and Aziz:

THEOREM 2.1. Given:
1. H_1 and H_2 are two real Hilbert spaces with inner products $(\cdot,\cdot)_{H_1}$ and $(\cdot,\cdot)_{H_2}$, respectively.
2. $a(u, v)$ is a bilinear form on $H_1 \times H_2$ with $u \in H_1$ and $v \in H_2$ such that

$$|a(u, v)| \leqslant C_1 \|u\|_{H_1} \|v\|_{H_2} ,$$

$$\inf_{u \in H_1} \sup_{v \in H_2} \frac{a(u, v)}{\|u\|_{H_1} \|v\|_{H_2}} \geqslant C_2 > 0 ,$$

and

$$\sup_{u \in H_1} |a(u, v)| > 0 \qquad \forall v \neq 0 ,$$

where $C_1 < \infty$.
3. $f \in H_2'$, i.e. f is a bounded linear functional on H_2.
Then there exists a unique element $u_0 \in H_1$ such that

$$a(u_0, v) = f(v) \qquad \forall v \in H_2$$

and

$$\|u_0\|_{H_1} \leqslant C_2^{-1} \|f\|_{H_2'} .$$

For a proof of this result the reader is referred to Aziz [6] (Theorem 5.2.1). Applications of this theorem in the present context is possible, because of the following result:

LEMMA 2.1. Let $a(\cdot,\cdot)$ be the bilinear form on $\overset{0}{H}{}^1[0,1] \times \overset{0}{H}{}^1[0,1]$ defined by (2.5). Then

$$|a(u, v)| \leqslant C_1 |u|_1 |v|_1 , \qquad \inf_{u \in \overset{0}{H}{}^1} \sup_{v \in \overset{0}{H}{}^1} \frac{|a(u, v)|}{|u|_1 |v|_1} = 1 , \tag{2.7}$$

where

$$C_1 = [1 + k^2/(4\pi^2)]^{1/2} . \tag{2.8}$$

Proof: We can define

$$C_1 = \sup_{u \in \overset{0}{H}{}^1} \sup_{v \in \overset{0}{H}{}^1} \frac{|a(u, v)|}{|u|_1 |v|_1} .$$

Introducing Lagrange multipliers λ and μ, we find that the supremum of the functional $a(u, v)$ constrained by $|u|_1 = |v|_1 = 1$ is attained at a stationary point of the functional

$$J[u, v, \lambda, \mu] = a(u, v) - \tfrac{1}{2}\lambda(|u|_1^2 - 1) - \tfrac{1}{2}\mu(|v|_1^2 - 1) .$$

The Euler equations for this functional are

$$\lambda D^2 u + L^* v = 0 , \tag{2.9}$$

$$Lu + \mu D^2 v = 0 , \tag{2.10}$$

$$|u|_1^2 = |v|_1^2 = 1 , \tag{2.11}$$

where $D^2 \equiv d^2/dx^2$, L is the operator defined by (2.1) and L^* is its formal adjoint: $L^* = -D^2 - kD$. Taking the inner product of (2.9) and (2.10) with u and v respectively, integrating by parts and applying (2.11) give $\mu = \lambda$. Eqs. (2.9) and (2.10) then provide a generalized eigenvalue problem for the eigenvalue λ. Its solution, subject to the boundary conditions $u(0) = u(1) = v(0) = v(1) = 0$, is

$$\lambda^2 = 1 + k/(4p^2\pi^2) , \qquad p = 1, 2, \ldots , \tag{2.12}$$

and in order that $a(u, v)$ attain a maximum, it is appropriate to choose $p = 1$ and λ as the positive root of (2.12). From (2.9) and (2.11)

$$a(u, v) = \lambda |u|_1^2 = \lambda ,$$

and so $C_1 = \lambda = [1 + k^2/(4\pi^2)]^{1/2}$. To compute the constant C_2, we note first that

$$C_2 = \inf_{u \in \overset{0}{H}{}^1} \sup_{v \in \overset{0}{H}{}^1} \frac{a(u, v)}{|u|_1 |v|_1} \geqslant \inf_{u \in \overset{0}{H}{}^1} \frac{a(u, u)}{|u|_1^2} = 1$$

since $a(u, u) = |u|_1^2$.

It remains to show that this lower bound cannot be improved. For this purpose we have

$$C_2 \leqslant \sup_{v \in \overset{0}{H}{}^1,\ |v|_1 = 1} a(U, v) , \tag{2.13}$$

where $U \in \overset{0}{H}{}^1$ is a chosen test function with $|U|_1 = 1$. The required result is obtained using

$$U = \alpha(1 - \cos 2p\pi x) ,$$

where $p \in \mathrm{N}$ and $\alpha = (\sqrt{2}p\pi)^{-1}$.

Again applying the idea of Lagrange multipliers, the supremum of (2.13) is attained when $v = z/|z|_1$, and z is defined by

$$-D^2 z = LU, \qquad z(0) = z(1) = 0 .$$

It holds that

$$\sup_{v \in \overset{0}{H}{}^1, |v|_1 = 1} a(U, v) = |z|_1 \to 1 \text{ as } p \to \infty .$$

3. The Petrov–Galerkin finite element method

Let π_h denote the subdivision of the interval [0,1] into $N+1$ subintervals $[x_j, x_{j+1}]$, $j = 0, 1, \ldots N$, with $x_0 = 0$, $x_{N+1} = 1$. For convenience, we shall consider only the case of uniform partition: $x_{j+1} - x_j = h$, $j = 0, 1, \ldots N$. Associated with π_h we have two finite dimensional subspaces Φ_h, Ψ_h of $\overset{0}{H}{}^1[0,1]$ with the properties

1. Φ_h and Ψ_h have equal dimension N.
2. The restriction of any function out of either Φ_h or Ψ_h to a subinterval of π_h is a polynomial.

(In part 2 we shall extend this definition to include L-spines.)

The subspaces Φ_h and Ψ_h are commonly referred to as spaces of trial and test functions, respectively. It is evident from experience gained by Christie [7] that the condition $\Psi_h \in \overset{0}{H}{}^1$ may be weakened. However, it is sufficient for our purposes to consider only conforming test functions.

The Petrov–Galerkin finite element approximation to (2.1) and (2.2) is then:

Find $u_h \in \Phi_h$

so that $a(u_h, v_h) = (f, v_h) \qquad \forall v_h \in \Psi_h$. (3.1)

Following Babuska and Aziz (Aziz [6, ch. 6], we have:

THEOREM 3.1. Given:

1. Φ_h and Ψ_h are finite dimensional subspaces of $\overset{0}{H}{}^1[0,1]$ such that

$$\inf_{u \in \Phi_h} \sup_{v \in \Psi_h} \frac{a(u, v)}{|u|_1 |v|_1} = c_2(h) > 0 , \tag{3.2}$$

and for every $0 \neq v \in \Psi_h$,

$$\sup_{u \in \Phi_h} |a(u, v)| > 0 . \tag{3.3}$$

2. $u \in \overset{0}{H}{}^1$ is the unique solution of (2.6).

Then

$$|u - u_h|_1 \leqslant (1 + C_1/c_2(h)) \min_{w_h \in \Phi_h} |u - w_h|_1 \ , \tag{3.4}$$

where u_h is the unique solution of (3.1), and C_1 is the continuity constant defined in theorem 2.1.

Proof: Since $\Psi_h \subset \overset{0}{H}{}^1[0,1]$, we have from (2.6)

$$a(u, v_h) = (f, v_h) \qquad \forall v_h \in \Psi_h \ ,$$

which together with (3.1) gives

$$a(u - u_h, v_h) = 0 \qquad \forall v_h \in \Psi_h \ . \tag{3.5}$$

Let $\hat{u}_h$ denote an arbitrary element of Φ_h; then

$$a(u_h - \hat{u}_h, v_h) = a(u - \hat{u}_h, v_h) \qquad \forall v_h \in \Psi_h \ ,$$

from which it follows that

$$|a(u_h - \hat{u}_h, v_h)| \leqslant C_1 |u - \hat{u}_h|_1 |v_h|_1 \qquad \forall v_h \in \Psi_h \ ,$$

where C_1 is the continuity constant for $a(\cdot,\cdot)$. We have immediately from this inequality that

$$\sup_{w \in \Psi_h} \frac{|a(u_h - \hat{u}_h, w)|}{|w|_1} \leqslant C_1 |u - \hat{u}_h|_1 \ ,$$

which then combines with (3.2) to yield

$$|u_h - \hat{u}_h|_1 \leqslant C_1/c_2(h) |u - \hat{u}_h|_1 \ .$$

The result (3.4) then follows from the triangle inequality

$$|u - u_h|_1 \leqslant |u_h - \hat{u}_h|_1 + |u - \hat{u}_h|_1$$

and by noting that $\hat{u}_h$ is an arbitrary element of Φ_h.

The bound (3.4) will play a central role in much of the remainder of this paper. It is clear that the spaces Φ_h and Ψ_h should be chosen so that $c_2(h)$ is strictly bounded away from zero in the limit $h \to 0$ (from lemma 2.1 we have immediately that $c_2(h) \leqslant C_1$). It is interesting to note that when this is the case, the limiting behaviour of the error in energy as $h \to 0$ depends only on the approximating ability of the spaces Φ_h.

Based on the assumption that (3.4) adequately reflects the behaviour of the error with k and h, we define an optimized Petrov–Galerkin method as one which incorporates a space of test functions Ψ_h which renders the constant $c_2(h)$ as large as possible for a given space of trial functions Φ_h. We anticipate that this approach will remove the possibility of numerical solutions with unwanted oscillatory properties, as these will necessarily lead to errors with large energy norms. Further, through the inequality

$$\|z\|_\infty \leqslant 2^{-1/2}|z|_1 \tag{3.6}$$

which holds for all $z \in \overset{0}{H}{}^1[0,1]$, where $\|\cdot\|_\infty$ denotes the usual maximum norm, we see that errors with small energy norms will lead to small pointwise errors. In order to follow this line, we require concrete examples of the spaces Φ_h and Ψ_h.

4. The finite element spaces Φ_h and Ψ_h

It is sufficient for our present purposes to take Φ_h as the space of piecewise linear functions on π_h having as basis the socalled pyramid (or hat) functions

$$\varphi_j(x) = \varphi(x/h - j)), \qquad j = 1, 2, \ldots N, \tag{4.1}$$

where

$$\varphi(s) = \begin{cases} 0, & |s| > 1, \\ 1+s, & -1 \leqslant s \leqslant 0, \\ 1-s, & 0 \leqslant s \leqslant 1. \end{cases}$$

Thus, for any function $w_h \in \Phi_h$, we have the representation

$$w_h(x) = \sum_{j=1}^{N} w_h(x_j)\varphi_j(x).$$

Note that $w_h(0) = w_h(1) = 0$ for all choices of nodal parameters $\{w_h(x_1), \ldots w_h(x_N)\}$.

For the test functions Ψ_h we adopt a family of spaces involving a parameter α and denote this family by $\Psi_{h,\alpha}$. For each value of α, $\Psi_{h,\alpha}$ has a basis $\{\psi_1(x), \ldots \psi_N(x)\}$, each having support on an interval of length $2h$ and defined by

$$\psi_j(x) = \varphi_j(x) + \alpha\sigma(x/h - j), \qquad j = 1, 2, \ldots N, \tag{4.2}$$

where $\varphi_j(x)$ is defined by (4.1), and $\sigma(s)$ is *any* odd function out of $\overset{0}{H}{}^1[-1,1]$ with $\int_0^1 \sigma(s)\,ds = -\frac{1}{2}$. A convenient choice for $\sigma(s)$ is

$$\sigma(s) = \begin{cases} 0, & |s| > 1, \\ -3s(1-s), & 0 \leqslant s \leqslant 1, \\ -\sigma(-s), & -1 \leqslant s \leqslant 0. \end{cases} \tag{4.3}$$

This results in one of the spaces discussed by Christie et al. [4]; alternate choices for $\sigma(s)$ would lead to the other test spaces discussed in that paper.

It will become apparent below that the sign of the parameter α of (4.2) is dictated by that of the constant k appearing in (2.1). According, without loss of generality, we can assume that $\alpha \geqslant 0$.

In order that theorem 3.1 may be invoked and, in particular, to study the effect of the parameter α on the bound (3.4), we have the following result:

LEMMA 4.1. With finite dimensional spaces $\Phi_h, \Psi_{h,\alpha}$ described above, the quantity $c_2(h)$ defined by (3.2) satisfies the inequality

$$C \leqslant c_2(h) \leqslant C\sqrt{1+\gamma^2 h}\,, \tag{4.4}$$

where $C = (1 + \frac{1}{2}kh\alpha)(1 + 3\alpha^2)^{-1/2}$, $\gamma = \frac{1}{2}kh/(1 + \frac{1}{2}kh\alpha)$, and the lower bound is attained when Φ_h (and $\Psi_{h,\alpha}$) has odd dimension. Furthermore,

$$\sup_{u_h \in \Phi_h} a(u_h, v_h) > 0 \qquad 0 \neq v_h \in \Psi_{h,\alpha}\,. \tag{4.5}$$

Proof: Let $u_h = \Sigma_{j=1}^{N} u_j \varphi_j(x) \in \Phi_h$ and $v_h = \Sigma_{j=1}^{N} v_j \psi_j(x) \in \Psi_{h,\alpha}$, then a straightforward computation yields

$$a(u_h, v_h) = \mathbf{v}^t \boldsymbol{A}_{h,\alpha} \boldsymbol{u}\,,$$

$$|u_h|_1^2 = \boldsymbol{u}^t \boldsymbol{A}_{\varphi,h} \boldsymbol{u}\,,$$

$$|v_h|_1^2 = (1 + 3\alpha^2)\,\mathbf{v}^t \boldsymbol{A}_{\varphi,h} \mathbf{v}\,,$$

where $\boldsymbol{u} = [u_1, u_2, \ldots u_N]^t$, $= [v_1, v_2, \ldots v_N]^t$,

$$\boldsymbol{A}_{h,\alpha} = h^{-1}\begin{bmatrix} 2 + kh\alpha & -1 + \frac{1}{2}kh(1-\alpha) & & \\ -1 - \frac{1}{2}kh(1+\alpha) & 2 + kh\alpha & -1 + \frac{1}{2}kh(1-\alpha) & \\ & \ddots & \ddots & \ddots \\ & & -1 - \frac{1}{2}kh(1+\alpha) & 2 + kh\alpha \end{bmatrix} \tag{4.7}$$

and

$$\boldsymbol{A}_{\varphi,h} = h^{-1}\begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}. \tag{4.8}$$

Thus,

$$c_2(h) = \min_{u \in \mathbb{R}^N} \max_{v \in \mathbb{R}^N} \frac{\mathbf{v}^t \boldsymbol{A}_{h,\alpha} \boldsymbol{u}}{((1+3\alpha^2)\mathbf{v}^t \boldsymbol{A}_{\varphi,h} \mathbf{v}\, \boldsymbol{u}^t \boldsymbol{A}_{\varphi,h} \boldsymbol{u})^{1/2}} \geqslant \min_{u \in \mathbb{R}^N} \frac{\boldsymbol{u}^t \boldsymbol{A}_{h,\alpha} \boldsymbol{u}}{(1+3\alpha^2)^{1/2} \boldsymbol{u}^t \boldsymbol{A}_{\varphi,h} \boldsymbol{u}}\,, \tag{4.9}$$

and, since

$$\boldsymbol{u}^{\mathrm{t}}\boldsymbol{A}_{h,\alpha}\boldsymbol{u} = \tfrac{1}{2}\boldsymbol{u}^{\mathrm{t}}(\boldsymbol{A}_{h,\alpha} + \boldsymbol{A}_{h,\alpha}^{\mathrm{t}})\boldsymbol{u} = (1 + \tfrac{1}{2}kh\alpha)\boldsymbol{u}^{\mathrm{t}}\boldsymbol{A}_{\varphi,h}\boldsymbol{u} ,$$

we have

$$c_2(h) \geqslant (1 + \tfrac{1}{2}kh\alpha)\,(1 + 3\alpha^2)^{-1/2} ,$$

and the lower bound of (4.4) is established.

For the upper bound, we have from (4.9) that

$$c_2(h) \leqslant \max_{\boldsymbol{v}\in\mathbb{R}^N} \frac{\boldsymbol{v}^{\mathrm{t}}\boldsymbol{A}_{h,\alpha}\boldsymbol{u}}{((1 + 3\alpha^2)\,\boldsymbol{v}^{\mathrm{t}}\boldsymbol{A}_{\varphi,h}\boldsymbol{v}\,\boldsymbol{u}^{\mathrm{t}}\boldsymbol{A}_{\varphi,h}\boldsymbol{u})^{1/2}} , \tag{4.10}$$

where $\boldsymbol{u}$ is now a selected vector. Two cases arise:

1. N, the dimension of $\boldsymbol{\Phi}_h$, is odd.

Select $\boldsymbol{u} = [1,0,1,0, \ldots 0,1]^{\mathrm{t}}$. Then, since $\boldsymbol{A}_{h,\alpha}\boldsymbol{u} = (1 + \frac{1}{2}kh\alpha)\boldsymbol{A}_{\varphi,h}\boldsymbol{u}$, we have

$$c_2(h) \leqslant \frac{1 + \frac{1}{2}kh\alpha}{(1 + 3\alpha^2)^{1/2}} \max_{\boldsymbol{v}\in\mathbb{R}^N} \frac{\boldsymbol{v}^{\mathrm{t}}\boldsymbol{A}_{\varphi,h}\boldsymbol{u}}{(\boldsymbol{v}^{\mathrm{t}}\boldsymbol{A}_{\varphi,h}\boldsymbol{v}\,\boldsymbol{u}^{\mathrm{t}}\boldsymbol{A}_{\varphi,h}\boldsymbol{u})^{1/2}} ,$$

and the maximum of the quantity on the right of this expression is easily computed to be unity, and is attained when $\boldsymbol{v} = \boldsymbol{u}$. Thus, in the case that N is odd,

$$c_2(h) = (1 + \tfrac{1}{2}kh\alpha)\,(1 + 3\alpha^2)^{-1/2} .$$

2. N is even.

We now select

$$\boldsymbol{u} = [1,0,1, \ldots 0,1,1,0, \ldots 0,1]^{\mathrm{t}} \text{ if } N/2 \text{ is odd}$$
$$\qquad\qquad\quad \uparrow\ \uparrow$$
$$\qquad\qquad N/2, N/2 + 1$$
$$\qquad\qquad\quad \downarrow\ \downarrow$$
$$\boldsymbol{u} = [1,0,1, \ldots 1,0,0,1, \ldots 0,1]^{\mathrm{t}} \text{ if } N/2 \text{ is even}$$

to maintain symmetry. Let $\boldsymbol{z}$ denote the solution of the system

$$\boldsymbol{A}_{\varphi,h}\boldsymbol{z} = \boldsymbol{A}_{h,\alpha}\boldsymbol{u} ; \tag{4.11}$$

then

$$\max_{\boldsymbol{v}\in\mathbb{R}^N} \frac{\boldsymbol{v}^{\mathrm{t}}\boldsymbol{A}_{h,\alpha}\boldsymbol{u}}{(\boldsymbol{v}^{\mathrm{t}}\boldsymbol{A}_{\varphi,h}\boldsymbol{v}\,\boldsymbol{u}^{\mathrm{t}}\boldsymbol{A}_{\varphi,h}\boldsymbol{u})^{1/2}} = \frac{(\boldsymbol{z}^{\mathrm{t}}\boldsymbol{A}_{\varphi,h}\boldsymbol{z})^{1/2}}{(\boldsymbol{u}^{\mathrm{t}}\boldsymbol{A}_{\varphi,h}\boldsymbol{u})^{1/2}} , \tag{4.12}$$

and the maximum is attained for $\boldsymbol{v} = \boldsymbol{z}$. With $\boldsymbol{u}$ chosen as above, (4.11) gives

$$\boldsymbol{z} = (1 + \tfrac{1}{2}kh\alpha)(\boldsymbol{u} + \gamma\hat{\boldsymbol{z}}) , \tag{4.13}$$

where $\gamma = \frac{1}{2}kh(1 + \frac{1}{2}kh\alpha)^{-1}$, $\hat{z}$ is the solution of

$$\boldsymbol{A}_{\varphi,h}\hat{\boldsymbol{z}} = h^{-1}\boldsymbol{b} , \tag{4.14}$$

and

$$\boldsymbol{b} = [0,0, \ldots 0,1,-1,0, \ldots 0]^t \text{ if } N/2 \text{ is odd}$$

or

$$\boldsymbol{b} = [0,0, \ldots 0,-1,1,0, \ldots 0]^t \text{ if } N/2 \text{ is even} ,$$

the nonzero entries occurring in the $N/2$ and $N/2 + 1$ positions. Solving (4.14) directly and substituting the result into (4.13), we find

$$\boldsymbol{z}^t\boldsymbol{A}_{\varphi,h}\boldsymbol{z} = (1 + \tfrac{1}{2}kh\alpha)^2 \ (\boldsymbol{u}^t\boldsymbol{A}_{\varphi,h}\boldsymbol{u} + \gamma^2 N) ,$$

which, after the computation $\boldsymbol{u}^t\boldsymbol{A}_{\varphi,h}\boldsymbol{u} = h^{-1}N$ and substitution into (4.12) and (4.10), yields the required upper bound (4.4). Finally, (4.5) follows from

$$\sup_{u_h \in \boldsymbol{\Phi}_h} a(u_h, v_h) = \max_{\boldsymbol{u} \in \mathbb{R}^N} \boldsymbol{v}^t\boldsymbol{A}_{h,\alpha}\boldsymbol{u} \geqslant \boldsymbol{v}^t\boldsymbol{A}_{h,\alpha}\boldsymbol{v}$$

$$= (1 + \tfrac{1}{2}kh\alpha)\,\boldsymbol{v}^t\boldsymbol{A}_{\varphi,h}\boldsymbol{v} = (1 + \tfrac{1}{2}kh\alpha)\,|v_h|_1^2 > 0 \qquad \forall v_h \neq 0 .$$

The convergence of the Petrov–Galerkin finite element method does not require knowledge as precise as that contained in lemma 4.1; it requires only that $c_2(h)$ is bounded away from 0 when h tends to 0. It may be possible to establish this result in a more general setting by application of results similar to those of Beyn and Lorenz [8]. The analysis of lemma 4.1 would be much more difficult for a more complex differential operator L, but we note that it may be possible to deduce the behaviour of $c_2(h)$ with h, k and α by taking only representative basis functions $\varphi_j(x)$ and $\psi_j(x)$ from $\boldsymbol{\Phi}_h$ and $\boldsymbol{\Psi}_{h,\alpha}$, respectively, and studying the quantity $\alpha(\varphi_j, \psi_j)/(|\varphi_j|, |\psi_j|)^{1/2}$. In most applications the factor $(1 + \gamma^2 h)^{1/2}$ appearing in (4.4) is close to unity, and therefore in the sequel we shall assume that

$$c_2(h) \approx (1 + \tfrac{1}{2}kh\alpha)(1 + 3\alpha^2)^{-1/2} . \tag{4.15}$$

We recall the well-known result that the best approximation to $u \in \overset{0}{H}{}^1[0,1]$, out of $\boldsymbol{\Phi}_h$, with respect to the norm $|\cdot|_1$ actually interpolates u at the nodes of $\boldsymbol{\pi}_h$. Thus, by choosing the parameter α so as to make the energy of the error of the finite element solution as small as possible, we anticipate that this will have a favourable effect on the nodal values.

The effect of the parameter α in the error bound (3.4) is contained solely in the constant $c_2(h)$; accordingly, in order to choose the space $\Psi_{h,\alpha}$ so as to make this bound as small as possible, we should choose α so as to make $c_2(h)$ as large as possible. This is achieved when $\alpha = \alpha^*$:

$$\alpha^* = kh/6 , \tag{4.16}$$

giving

$$c_2^*(h) = (1 + (kh)^2/12)^{1/2} . \tag{4.17}$$

Thus, the optimized Petrov–Galerkin finite element solution u_h^* is determined by

$$a(u_h^*, v_h) = (f, v_h) \qquad \forall v_h \in \Psi_{h,\alpha^*} . \tag{4.18}$$

Thus, from (2.8), (3.4), (4.16) and (4.17) u_h^* satisfies

$$|u - u_h^*|_1 \leqslant \left\{ 1 + \left[\frac{1 + k^2/(4\pi^2)}{1 + k^2h^2/12} \right]^{1/2} \right\} \min_{w_h \in \Phi_h} |u - w_h|_1 . \tag{4.19}$$

From this result we deduce that, when k and kh are small, the optimized Petrov–Galerkin method yields a solution u^* which is close to the optimum approximation to u out of Φ_h, i.e. to the interpolant to u out of Φ_h. For fixed h the ratio

$$|u - u_h^*|_1 / \min_{w_h \in \Phi_h} |u - w_h|_1$$

remains bounded as $k \to \infty$, a result which does not hold when the Petrov–Galerkin method is employed with a constant value of α. We shall see in section 6 that such constant values arise from standard finite difference schemes. Before comparing these methods with those of finite differences we look at the behaviour of the error in the L_2 norm.

5. Approximation error in $L_2[0,1]$

In order to apply a variant of the Aubin–Nitsche duality argument, we require information on the approximating abilities of the spaces $\Psi_{h,\alpha}$.

LEMMA 5.1. Given any function $u \in H^2[0,1] \cap \overset{0}{H}{}^1[0,1]$, there exist a function $w_h \in \Psi_{h,\alpha}$ and a constant C (independent of u, h, α) such that

$$|u - w_h|_1 \leqslant C\{(1 + \alpha)\, h|u|_2 + |u|_1\} . \tag{5.1}$$

Furthermore, the dependence on α in this bound is optimal in that, for $z = \frac{1}{2}x(1 - x)$, we have

$$\inf_{w_h \in \Psi_{h,\alpha}} |z - w_h|_1 \geqslant K|z|_1 , \tag{5.2}$$

where $K^2 = 3\alpha^2 \cdot (1 + 3\alpha^2)^{-1}$.

Proof: Let $w_h \in \Psi_{h,\alpha}$ and $z_h \in \Phi_h$ denote the functions interpolating $u \in H^2[0,1] \cap \overset{0}{H}{}^1[0,1]$ at the points $x_1, x_2, \ldots x_N$. Then, from (4.2),

$$w_h = z_h + \alpha s_h \ ,$$

where

$$s_h = \sum_{j=1}^{N} w_h(x_j)\, \sigma_j(x) = \sum_{j=1}^{N} z_h(x_i)\sigma_j(x) \ , \qquad \sigma_j(x) = \sigma(x/h - j) \ ,$$

and $\sigma(\cdot)$ is defined by (4.3).

Hence,

$$|u - w_h|_1 \leqslant |u - z_h|_1 + \alpha |s_h|_1 \ .$$

By direct computation (cf. lemma 4.1) we find

$$|s_h|_1^2 = 3|z_h|_1^2 \ ,$$

and so

$$|u - w_h|_1 \leqslant |u - z_h|_1 + 3^{1/2}\alpha |z_h|_1 \ . \tag{5.3}$$

From Schultz [9] we have the existence of a constant C (independent of h and u) such that

$$|u - z_h|_1 \leqslant Ch|u|_2 \ . \tag{5.4}$$

This result, together with

$$|z_h|_1 \leqslant |u - z_h|_1 + |u|_1 \ ,$$

and (5.3) results in the inequality (5.1).

The inequality (5.2) is derived by observing that the given function z satisfies

$$-z'' = 1 \ , \qquad x \in (0,1) \ ,$$

$$z(0) = z(1) = 0 \ ,$$

and that, if $z_h \in \Psi_{h,\alpha}$ satisfies

$$(z_h', \psi_h') = (1, \psi_h) \qquad \forall \psi_h \in \Psi_{h,\alpha} \ ,$$

then z_h minimizes the expression $|z - w_h|_1$ taken over all $w_h \in \Psi_{h,\alpha}$:

$$|z - z_h|_1^2 = |z - w_h|_1^2 - |w_h - z_h|_1^2 \qquad \forall w_h \in \Psi_{h,\alpha} .$$

The result (5.2) follows by comparing z_h with the best approximation to z out of Φ_h.

Thus a convergent approximation out of $\Psi_{h,\alpha}$ to a given function u is obtained in the energy norm only when α vanishes in the limit $h \to 0$. We note that this is the case with the test functions for the optimized Petrov–Galerkin method (4.18).

Convergence of the Petrov–Galerkin solution u_h to u in L_2 is a direct consequence of theorem 3.1, lemma 4.1 and the inequality $\|z\|_0 \leqslant \pi^{-1}|z|_1 \ \forall z \in \overset{0}{H}{}^1[0,1]$. Under favourable conditions, the Aubin–Nitsche duality argument (see for example Oden and Reddy [10]) may be used to improve the rate of convergence of this result:

LEMMA 5.2. Let $u \in H^2[0,1] \cap \overset{0}{H}{}^1[0,1]$ and $u_h \in \Phi_h$ denote the unique solutions to the problems (2.6) and (3.1) with spaces Φ_h and $\Psi_{h,\alpha}$ as described in section 4. Then there exists a constant C (independent of h, α and u) so that

$$\|u - u_h\|_0 \leqslant C(1 + C_1/c_2(h))\,((1+\alpha)h + \alpha)h|u|_2 . \tag{5.5}$$

Proof: Let $z \in \overset{0}{H}{}^1[0,1] \cap C^2[0,1]$ denote the weak solution of

$$L^* z = u - u_h , \qquad x \in (0,1) , \tag{5.6}$$

$$z(0) = z(1) = 0 ,$$

where $L^* = -D^2 - kD$ is the formal adjoint of the operator L of (2.1).

Then

$$a(w, z) = (u - u_h, w) \qquad \forall w \in \overset{0}{H}{}^1[0,1] . \tag{5.7}$$

In particular,

$$a(u - u_h, z) = \|u - u_h\|_0^2 \tag{5.8}$$

and using (3.5) we obtain, for an arbitrary function $z_h \in \Psi_{h,\alpha}$,

$$a(u - u_h, z - z_h) = \|u - u_h\|_0^2 .$$

The continuity of the bilinear form $a(\cdot,\cdot)$ then yields

$$\|u - u_h\|_0^2 \leqslant C_1 |u - u_h|_1 \, |z - z_h|_1 . \tag{5.9}$$

From lemma 5.1 we have for a suitable $z_h \in \Psi_{h,\alpha}$

$$|z - z_h|_1 \leqslant C\{(1+\alpha)h + \alpha\}\,\|z\|_2 . \tag{5.10}$$

Further, if z is the solution of (5.6), it can be shown using Rellich's embedding theorem (Agmon

[11]) that there is a constant C (independent of z) such that

$$\|z\|_2 \leqslant C\|u - u_h\|_0 \,. \tag{5.11}$$

Collecting the results (3.4), (5.4), (5.10), (5.11) and (5.9) then provides the required estimate (5.5).

Thus, even though convergence of u_h to u is $O(h)$ in the energy norm for all values of the parameter α, convergence in L_2 is only $0(h)$ unless α tends to zero as $h \to 0$. This establishes a convergence rate of $0(h^2)$ in L_2 for the optimized Petrov–Galerkin method of (4.16) and (4.18).

6. Relationship to finite difference methods

From (3.1), (4.1), (4.2) and the calculations of lemma 4.1, we see that the nodal parameters $\{u_1, u_2, \ldots u_N\}$ of the Petrov–Galerkin solution u_h satisfy the difference equation

$$h(\boldsymbol{A}_{h,\alpha}\boldsymbol{u})_j = [-1 - \tfrac{1}{2}kh(1+\alpha)]\,u_{j-1} + (2 + kh\alpha)u_j + [-1 + \tfrac{1}{2}kh(1-\alpha)]\,u_{j+1} = h(f, \psi_j)\,, \tag{6.1}$$

$$j = 1, 2, \ldots N\,,$$

$u_0 = u_{N+1} = 0$. The parameter α also occurs on the right of this expression through the basis function $\psi_j(x)$. Its effect can be seen most clearly by assuming that $f \in C^2(0,1)$ and employing a Taylor expansion. In this way we easily compute

$$(f, \psi_j) = h\left(f(x_j) - \frac{\alpha h}{6} f'(x_j) + \frac{h^2}{12} f''(x_j)\right) + O(h^4)\,. \tag{6.2}$$

Since $f = Lu$ from (2.1), the order of the method (6.1), in a finite difference sense, is the largest integer p for which

$$\max_{j=1,\ldots N} \{h^{-1}|(\boldsymbol{A}_{h,\alpha}\boldsymbol{u})_j - (Lu, \psi_j)|\} = O(h^p)\,, \tag{6.3}$$

where u is any sufficiently smooth function and $\boldsymbol{u} = [u(x_1), \ldots u(x_N)]^t$. In table 1 we give the order of the method for several different values of α, some of which reproduce standard finite difference approximations of the operator L. The method (6.1) will not however be identical to finite difference approximations of $Lu = f$ for these values of α, due to the different treatment of the source term f (unless of course f is constant). From table 1 it is seen that the orders of these methods are consistent with the L_2 estimates of the previous section except when $\alpha = kh/6$ when the order is 4. This suggests that, for linear problems with constant coefficients, we may

Table 1. Order of truncation error for methods based on different test spaces $\Psi_{h,\alpha}$

α	Name of method	Order
0	Central difference	2
1	Backward difference	1
−1	Forward difference	1
$kh/6$	Optimized Petrov–Galerkin	4

attain $O(h^4)$ convergence at the nodes. This is confirmed by the numerical results to be presented later. Such high-order results should not be expected for more complex problems.

The notions of order have importance only when h is sufficiently small; in the present context this requires that kh be small. From (4.7) the matrix $A_{h,\alpha}$ defining the difference eq. (6.1) is an M-matrix if and only if

$$\alpha \geqslant 1 - 2/kh \tag{6.4}$$

(Varga [12]).

If (6.4) is violated, it follows (see Gantmacher and Krein [13]) that the inverse matrix $A_{h,\alpha}^{-1}$ has the following sign pattern:

$$\operatorname{sgn}(A_{h,\alpha}^{-1})_{ij} = \begin{cases} (-1)^{i+j} & \text{for } i \leqslant j\,, \\ 1 & \text{for } i > j\,. \end{cases} \tag{6.5}$$

The elements $(A_{h,\alpha}^{-1})_{iN}$, $i = 1, 2, \ldots N$, being of alternating signs, lead to numerical solutions with oscillating errors when the truncation error in the last equation dominates. This is precisely the situation for boundary layer type solutions when the mesh length h is not sufficiently small. For $\alpha = 0$ (central differences) (6.4) leads to the restriction $kh \leqslant 2$. Inequality (6.4) is satisfied by $\alpha = 1$ (backward differences) and $\alpha = kh/6$ (optimised Petrov–Galerkin method) for all $h > 0$, and consequently both these methods will lead to oscillation-free solutions.

7. Preliminary numerical results

We shall present here only sample numerical results to indicate the degree to which the preceding analysis has been succesful and enable us in Part 2 to focus on areas where further investigation is required. The examples relate to the solution of (2.1) and (2.2) with a prescribed source term. We shall take the value $k = 60$ throughout.

Example 7.1. $\qquad f(x) = 4\pi^2 \sin 2\pi x + 2k\pi \cos 2\pi x$

Problem (2.1), (2.2) then has the exact solution $u = \sin 2\pi x$. In table 2, we present the values of maximum nodal errors for three of the methods described above. The superiority of the optimized Petrov–Galerkin method is expected due to its higher order, but it is interesting to note that the central difference method ($\alpha = 0$) is oscillation free in this example even when $kh > 2$. These results lend support to the technique by which we optimize the Petrov–Galerkin method.

Example 7.2. $\qquad f(x) = k$

This leads to the exact solution

$$u(x) = x - (e^{kx} - 1)/(e^k - 1)\,.$$

Table 2. Maximum nodal errors $\max_{1\leq j\leq N} |u(x_j) - u_h(x_j)|$ for the Petrov–Galerkin method with the optimal value $\alpha = kh/6$, $\alpha = 0$ (central differences) and $\alpha = 1$ (backward differences), applied to example 7.1

h \ α	$kh/6$	0	1
1/5	.315(−1)	.159(−0)	.812(−1)
1/10	.376(−2)	.370(−1)	.376(−2)
1/20	.254(−3)	.908(−2)	.103(−1)
1/40	.160(−4)	.226(−2)	.721(−2)
1/80	.100(−5)	.565(−3)	.408(−2)

The rapid change in the solution $u(x)$ near $x = 1$ makes this example much more difficult than the previous one; this is borne out by the much larger values of maximum nodal error listed in table 3 for the same three methods. The relative performance of the methods is unchanged when h is small, but there are two major overall differences. First, and most important, the optimized Petrov–Galerkin method performs rather poorly for $h = 1/5$, better results being obtained with backward differences ($\alpha = 1$). The errors for this latter method however deteriorate as h is halved. Secondly, as anticipated in the previous section, the central difference method now produces oscillatory solutions unless $kh < 2$.

We also present, in table 4 the quantities $|u - u_h|_1$ and $|u|_1 - |u_h|_1$ for these methods. Note that the optimized Petrov–Galerkin method does in fact produce the smallest value for $|u - u_h|_1$ but only when $kh < 6$. Moreover, the value of this quantity is larger than might be anticipated and only marginally smaller than that for the other methods; we return to this question in section 9. The difference $|u|_1 - |u_h|_1$ may be regarded as a measure of the numerical dissipation of a method (see Roach [3]). Whereas the other methods perform poorly in this regard, the central difference method as surprisingly good. This stems from the relations

$$|u|_1^2 = \tfrac{1}{2}k - u'(0)\,, \tag{7.1}$$

and

$$(1 + \tfrac{1}{2}kh\alpha)|u_h|_1^2 = \tfrac{1}{2}k - \tfrac{1}{2}kh(1 - u_h'(0)) - u_h'(0)\,(1 + \tfrac{1}{2}\alpha kh)\,, \tag{7.2}$$

which hold for this example, together with the observations that $u'(0) = 1 + O(ke^{-k})$ and $u_h'(0)$ rapidly approaches the value unity (for all α considered) even for coarse meshes. The superiority of the central difference method in this regard persists for other problems although at a reduced level.

Table 3. Maximum nodal errors for Petrov–Galerkin method applied to example 7.2 (cf. table 2)

h \ α	$kh/6$	0	1
1/5	.364	.587*	.769(−1)
1/10	.140	.504*	.140
1/20	.271(−1)	.250*	.200
1/40	.268(−2)	.803(−1)	.179
1/80	.161(−3)	.178(−1)	.103

Table 4. The error in energy $|u - u_h|_1$ (lower value) and numerical dissipation $|u|_1 - |u_h|_1$ (upper value) for Petrov–Galerkin method applied to example 7.2 (for an explanation of the starred quantities see eqs. (7.1) and (7.2))

h	$kh/6$	0	1
1/5	.433(+1) .510(+1)	.959 .601(+1)	.357(+1) .501(+1)
1/10	.284(+1) .452(+1)	−.544(−2) .549(+1)	.284(+1) .452(+1)
1/20	.137(+1) .345(+1)	−.705(−11)* .385(+1)	.207(+1) .362(+1)
1/40	.459 .214(+1)	−.636(−11)* .223(+1)	.137(+1) .249(+1)
1/80	.126 .115(+1)	−.534(−11)* .117(+1)	.822 .150(+1)

The two examples cited serve to illustrate that our analysis has been effective, even for large k, provided the solution to be approximated does not undergo a rapid variation near the boundary (as in example 7.2). It is with the aim of improving numerical techniques in these situations that we devote part 2 of this paper to a detailed analysis of the boundary layer phenomenon.

Part 2

8. Boundary layer behaviour

It is clear from the limited numerical results of part 1 that the accuracy attained deteriorates rapidly when k (and kh) is large for certain problems. Before attempting to adapt the numerical schemes to deal with this situation, it is appropriate to consider the nature of the continuous problem (2.1) as $k \to \infty$. To preclude the possibility that the solution of (2.1) becomes vanishingly small in the limit, we rescale the source term to give

$$Lu \equiv -u'' + ku' = kF(x), \qquad x \in (0,1), \tag{8.1}$$

$$u(0) = u(1) = 0,$$

where, for convenience of analysis, we further assume that $F(x)$ is independent of k. Our conclusions will remain valid even though $F(x)$ has a dependence on k through terms of the form k^{-1}.

We begin by decomposing (8.1) into two separate problems:

$$\text{a)} \qquad Lu = kF(x) - k\int_0^1 F(x)\,dx, \tag{8.2}$$

$$\text{b)} \qquad Lu = k\int_0^1 F(x)\,dx, \tag{8.3}$$

where U and V are both subject to homogeneous Dirichlet boundary conditions. The solution $u(x)$ of (8.1) is then the sum of the functions U and V.

Problem (a): If we allow $k \to \infty$ in (8.2), we obtain the so-called reduced problem

$$U'_\infty = F(x) - \int_0^1 F(s)\,\mathrm{d}s\,. \tag{8.4}$$

Integration reveals that if one boundary condition, say $U_\infty(0)$ is imposed, the other condition $U_\infty(1) = 0$ is automatically satisfied. This indicates that (8.2) is devoid of any "boundary-layer" behaviour; indeed, using Green's functions, the quantity $\|U(x) - U_\infty(x)\|_\infty$ is easily shown to be $O(k^{-1})$. Thus, provided F is not subject to rapid variations, the solution of (8.2) can be accurately described by any of the numerical methods discussed earlier. This is evidenced by the first numerical example of the previous section, even though the function $F(x)$ there has a weak dependence on k^{-1}.

Problem (b). We can, without loss of generality, assume that $\int_0^1 F(s)\,\mathrm{d}s = 1$. Thus (8.3) becomes

$$-V'' + kV' = k\,, \tag{8.5}$$

with $V(0) = V(1) = 0$.

A boundary-layer analysis is unnecessary for such an elementary problem; its exact solution is

$$V(x) = x - (\mathrm{e}^{kx} - 1)/(\mathrm{e}^k - 1)\,. \tag{8.6}$$

Allowing $k \to \infty$, then

$$V(x) \to V_\infty(x) = x\,, \qquad x \subset [0,1)\,, \tag{8.7}$$

where $V_\infty(x)$ is the solution of the reduced problem derived from (8.5):

$$\begin{aligned} &V'_\infty(x) = 1\,, \qquad x \in (0,1)\,, \\ &V_\infty(0) = 0\,. \end{aligned} \tag{8.8}$$

Since $V_\infty(x)$ clearly violates the right-hand boundary condition, the convergence in (8.7) is not in the $\overset{0}{H}{}^1[0,1]$ sense but only in $L_2[0,1]$ (or pointwise in $[0,1)$). This is a reflection of the singular nature of the perturbation (in k^{-1}) and contrasts with the regular behaviour of problem (a) with k^{-1} (see for instance Carrier and Pearson [5]).

For large k we may approximate (8.6) by

$$V(x) \approx \begin{cases} x & x \in [0,\, 1-\epsilon)\,, \qquad (8.9a) \\ 1 - \mathrm{e}^{-k(1-x)}\,, & x \in (1-\epsilon,\, 1]\,, \qquad (8.9b) \end{cases}$$

where ϵ is a notional value of the "boundary-layer thickness". Given a tolerance δ, the edge of the boundary layer is taken to be the ordinate at which the solutions of (8.5) and (8.8) differ by

the amount δ. Thus, if a precision of 5% is acceptable, we have $\epsilon \sim 3k^{-1}$, and this is adequate for computations at large values of kh with crude numerical methods. However, for high-precision methods (which we shall introduce below) a smaller tolerance must be selected, and, for $\delta = 10^{-m}$, it is easily found that the appropriate value for the thickness is approximately

$$\epsilon \sim \gamma m k^{-1} , \qquad \gamma = \ln 10 = 2.30 \ldots \tag{8.10}$$

Combining the results of problems (a) and (b), the solution $u(x)$ to (8.1) is seen to be well-behaved (and consequently accurately approximated by piecewise polynomials) on $[0, 1-\epsilon)$, but on $(1-\epsilon, 1]$ its behaviour is dominated by exponential terms of the form (8.9b). We note that these are solutions of the homogeneous problem $Lu = 0$. This will, for the present, provide sufficient motivation for a brief investigation of L-splines in section 9. The results of that section will however have important ramifications regarding the methods discussed in sections 4 and 5.

9. L-splines

We emphasise at the outset that, because of difficulties in generalizing to more complex problems, it is not our intention to recommend L-splines as a numerical technique for boundary-layer problems. Rather, we use their superior behaviour in the boundary-layer to give insight into how best to adapt the methods of sections 4–6 to such problems.

We shall require only a rudimentary knowledge of L-splines; more details may be obtained from Varga [15]. Specifically, we define a finite dimensional subspace L_h of $\overset{0}{H}{}^1[0,1]$, so that each function $z_h \in L_h$ is a solution of $Lu = 0$ locally on each subinterval of π_h. Then L_h has a basis $\{\chi_1, \chi_2, \ldots \chi_N\}$ of functions each of which has support on an interval of length $2h$ and is defined by

$$\chi_j(x) = \chi(x/h - j) , \qquad j = 1, 2, \ldots N , \tag{9.1}$$

where

$$\chi(s) = \begin{cases} 0 & |s| \geqslant 1 , \\ (e^{hks} - e^{-hk})/(1 - e^{-hk}) , & -1 \leqslant s \leqslant 0 , \\ (e^{hk} - e^{hks})/(e^{hk} - 1) , & 0 \leqslant s \leqslant 1 . \end{cases} \tag{9.2}$$

Now, employing the space L_h rather than Φ_h of section 4, we have a Petrov–Galerkin finite element method based on L-splines (which will be designated the PGL method) (cf. (3.1)):

$$\begin{aligned} &\text{Find} \quad l_h \in L_h \\ &\text{so that} \quad a(l_h, \psi_h) = (f, \psi_h) \qquad \forall \psi_h \in \Psi_{h,\beta} . \end{aligned} \tag{9.3}$$

Replacing Φ_h by L_h in theorem 3.1, we have, from (3.4),

$$|u - l_h|_1 \leqslant (1 + C_1/\bar{c}_2(h)) \min_{w_h \in L_h} |u - w_h|_1 , \tag{9.4}$$

where

$$\bar{c}_2(h) = \inf_{u_h \in L_h} \sup_{v_h \in \Psi_{h,\beta}} \frac{|a(u_h, v_h)|}{|u_h|_1 \, |v_h|_1} . \tag{9.5}$$

Analogous to lemma 4.1, we have:

LEMMA 9.1. With L_h and $\Psi_{h,\beta}$ finite dimensional spaces as described above, the quantity $\bar{c}_2(h)$ defined by (9.5) satisfies

$$\bar{C} \leqslant \bar{c}_2(h) \leqslant \bar{C}(1 + h \tanh^2 \tfrac{1}{2}kh)^{1/2} , \tag{9.6}$$

where

$$\bar{C} = [\tfrac{1}{2}hk \coth \tfrac{1}{2}hk/(1 + 3\beta^2)]^{1/2} .$$

Furthermore, the lower bound is attained when the dimension of L_h (and $\Psi_{h,\beta}$) is odd.

Proof: Let $\bar{A}_h$ and $A_{\chi,h}$ denote the $N \times N$ matrices with entries $(\bar{A}_h)_{ij} = a(\chi_j, \psi_i)$ and $(A_{\chi,h})_{ij} = (\chi_i', \chi_j')$, respectively. Then an easy calculation gives

$$\bar{A}_h = \frac{\frac{1}{2}k}{\mu - 1} \begin{bmatrix} 1+\mu & -1 & & \\ -\mu & 1+\mu & -1 & \\ & \ddots & \ddots & \ddots \\ & -\mu & 1+\mu & \end{bmatrix}, \tag{9.7}$$

and $A_{\chi,h} = \frac{1}{2}kh \coth \frac{1}{2}kh A_{\varphi,h}$, where $\mu = \exp(kh)$ and $A_{\varphi,h}$ is defined in lemma 4.1. With these results the derivation of (9.6) is essentially the same as that of (4.4) and the details will be omitted.

Following the rationale of section 4, we now choose the parameter β of the space $\Psi_{h,\beta}$ so as to optimize the bound (9.4).

Unlike the analysis of section 4 we can justify this optimization only to a limited extent. For, unlike the interpolation properties of the best approximation to $u \in \overset{0}{H}{}^1[0,1]$ out of Φ_h with respect to the norm $|\cdot|_1$, in the present context we cannot conclude that the corresponding best approximation out of L_h will interpolate the solution of the boundary layer problem (8.5). But the remarks made earlier regarding (3.6) remain valid.

The maximization of $\bar{c}_2(h)$ clearly leads to the value $\beta = 0$. Thus the best PGL method (in that it provides the smallest error bound (9.4)) results from taking $\Phi_h (\equiv \Psi_{h,0})$ as the space of test functions. With this value for β (9.4) becomes

$$|u - l_h|_1 \leqslant C \min_{w_h \in L_h} |u - w_h|_1 , \tag{9.8}$$

where

$$C = 1 + C_1(\tfrac{1}{2}hk \coth \tfrac{1}{2}hk)^{-1/2} . \tag{9.9}$$

By comparing the matrices $A_{h,\alpha}$ and $\bar{A}_h$ derived from the finite element methods (3.1) and (9.3), and displayed in (4.7) and (9.7), respectively, we see that they are identical when α has the value

$$\hat{\alpha} = \coth \tfrac{1}{2}hk - 2/kh \tag{9.10}$$

(see fig. 1). Investigating the asymptotic form of this parameter, we find

$$\hat{\alpha} = \begin{cases} kh/6 + O(k^3h^3) , & kh \text{ small} , \\ 1 - 2/kh + O(\mathrm{e}^{-kh}) , & kh \text{ large} , \end{cases} \tag{9.11}$$

and this method therefore approaches the optimized Petrov–Galerkin method of section 11.4 when kh is small, and the backward difference method ($\alpha = 1$) when kh is large. The equivalent finite difference method based on this value of the parameter has been known for some time (Miller [16], Christie et al. [4]). It is usually derived on the basis that, for homogeneous problems $Lu = 0$ with inhomogeneous Dirichlet boundary conditions, the finite difference method will exactly interpolate the exact solution at the nodes.

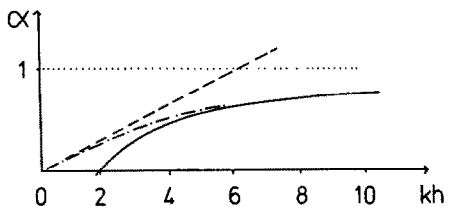

Fig. 1. Behaviour of the parameter α (eq. (9.10)).

If $\hat{u}_h \in \boldsymbol{\Phi}_h$ now denotes the finite element solution by the Petrov–Galerkin method with a space of test functions $\boldsymbol{\Psi}_{h,\hat{\alpha}}$, we have from (3.4) and (4.4)

$$|u - \hat{u}_h|_1 \leqslant C \min_{w_h \in \boldsymbol{\Phi}_h} |u - w_h|_1 , \tag{9.12}$$

where the constant C is the same as that appearing in (9.8). Our present concern with (9.8), (9.12) and indeed (4.19) is that, for boundary layer problems, the energy norms which appear are relatively large quantities (see table 4). This brings into question the validity of the assumption that minimizing these bounds will lead to good nodal values. We shall limit our remarks to the approximation of (8.5), for, in this case, all methods treat the source term in an identical manner. From the preceding discussion it is clear that the functions l_h and $\hat{u}_h$ of (9.8) and (9.12) have identical nodal values and, indeed, coincide with those arising from a composite method based on a space of trial functions consisting of L-splines in the boundary layer and piecewise linear functions elsewhere (provided obvious changes are made to the test functions). The solution of such a method also satisfies a bound of the form (9.8) and (9.12) in which, due to (8.9), the energy norms are small quantities (this deriving from the interpolation properties of best approcimation out of $\boldsymbol{\Phi}_h$ and L_h to this problem). Thus, by (3.6), small pointwise errors are ensured for $u - \hat{u}_h$

and $u - l_h$. We therefore conclude that the algebraic eqs. (6.1) which define the nodal values are to some degree insentitive to the actual trial functions chosen for the solution, but the latter have a significant effect when computing the energy of the solution and its error. We shall return to these composite methods and further adaptations in the next section, where we attempt to design efficient numerical algorithms for boundary layer problems.

10. Composite methods for boundary layer problems

In this section we shall bring together the results of the preceding sections. We begin by constructing a somewhat esoteric scheme which appears to be optimal for the model problem under discussion. Following a brief discussion of its properties, we shall turn to near-optimal methods which are easily generalized to more complex problems and higher dimensions.

We recall that in section 8.1 it was shown that the essential behaviour of the solution u in the boundary is described by (8.9b) and that u has a relatively uniform variation on the remainder of the interval. Let M denote the number of mesh points within the boundary layer (if the thickness ϵ of this layer is less than h, we take $M = 1$). We then define a composite Petrov–Galerkin method based on the space of test functions $\boldsymbol{\Psi}_h^c$:

$$\boldsymbol{\Psi}_h^c = \text{span}\{\psi_1^c, \psi_2^c, \ldots \psi_N^c\} ,$$

where

$$\psi_j^c \in \begin{cases} \boldsymbol{\Psi}_{h,\alpha^*} , & 1 \leqslant j \leqslant N - M , \\ \boldsymbol{\Psi}_{h,\hat{\alpha}} , & N - M < j \leqslant N , \end{cases} \tag{10.1}$$

and, from (4.16) and (9.10), $\alpha^* = kh/6$ and $\hat{\alpha} = \coth \frac{1}{2}kh - 2/kh$. Taking trial functions from the space of piecewise linear functions $\boldsymbol{\Phi}_h$, then $\boldsymbol{\Psi}_h^c$ generates the optimized PGL method in the boundary layer and the optimized Petrov–Galerkin method elsewhere; this is equivalent to the use of L-splines in the boundary layer. The numerical results of the next section reveal that this composite method does in fact inherit the best properties of its component parts and produces consistently accurate nodal values. A major disadvantage of this method arises from attempts to generalize the optimal value $\hat{\alpha}$ to other differential operators of the form (2.1) with variable coefficients. There is also the difficulty in properly estimating the boundary layer thickness (cf. (8.10)).

To overcome these objections, we observe from (9.11) that $\hat{\alpha}$ rapidly approaches the value

$$\tilde{\alpha} = 1 - 2/kh \tag{10.2}$$

as kh increases (for $kh = 6$ the discrepancy is less than 1%). From the matrix of coefficients (4.7) of the Petrov–Galerkin method we see that this value of α is characterized by the property that the algebraic equation applied at node j contains no information from the advanced node $j + 1$. In particular, when $j = N$, the method completely ignores the boundary condition $u(1) = 0$. For this reason we refer to this as a "disconnected" method; it is generated by the space of test functions $\boldsymbol{\Psi}_h^d$ (cf. (10.1)):

$$\Psi_h^d = \text{span}\{\psi_1^d, \ldots \psi_N^d\} ,$$

where

$$\psi_j^d \in \begin{cases} \Psi_{h,\alpha^*} , & 1 \leqslant j \leqslant N - M , \\ \Psi_{h,\tilde{\alpha}} , & N - M < j \leqslant N . \end{cases} \tag{10.3}$$

It is sufficient to choose $M = 1$, for, when kh is in the range where $\hat{\alpha} \sim \alpha^*$, the boundary layer behaviour is confined to only one mesh length. It is important to realize that the approximate solution thus produced does not converge to the solution of (9.1) as $h \to 0$ but to that of the problem

$$-u'' + ku' = f(x) , \quad x \in (0,1) , \tag{10.4}$$

with

$$u(0) = 0 , \quad ku'(1) = f(1) + k^{-1} f'(1) . \tag{10.5}$$

The second boundary condition (10.5) can be derived by applying (10.4) and its derivative at $x = 1$, eliminating $u''(1)$ and then neglecting the term $k^{-1}u'''(1)$. It can be shown that the amplitude of the boundary layer component in the solution of (10.4), (10.5) is of order $O(k^{-3})$ and is consequently negligible for large k.

In a similar manner, the "totally disconnected" method, based on (10.3) with $M = N$, approximates the differential equation

$$ku'(x) = f(x) + k^{-1} f'(x) , \quad x \in (0,1) , \tag{10.6}$$

together with the initial condition $u(0) = 0$, When k is large, this is seen to be close to the reduced problem of section 8.

In problems where the optimal value α^* is not easily determined, this value may be replaced by either $\alpha = 1$ or $\alpha = 0$ in (10.3) with a consequent reduction in accuracy (see section 11).

11. Numerical results and discussion

The linearity of the model problem (8.1) and the decomposition (8.2) and (8.3) allows a separate analysis of "smooth" and boundary layer components in the solution. Numerical results for the former have been presented in section 7. The performance of the methods we have described on the boundary layer component is typified by the results for:

Example 11.1.

$$-u'' + ku' = 3kx^2 , \quad x \in (0,1) , \qquad u(0) = u(1) = 0 . \tag{11.1}$$

Table 5. Maximum nodal errors for Petrov–Galerkin method applied to example 11.1 (best results for each h are underlined; the star denotes that the solution oscillates)

h \ α	$kh/6$	$1-2/kh$	0	1
1/5	.383	.933(−2)	.618*	.881(−1)
1/10	.148	.133(−2)	.523*	.148
1/20	.285(−1)	.520(−1)	.261	.210
1/40	.281(−2)	.234	.842(−1)	.185
1/80	.164(−3)	.496	.187(−1)	.108

Table 6. Maximum nodal errors for optimized Petrov–Galerkin method ($\alpha = \alpha^*$), optimized PGL method ($\alpha = \hat{\alpha}$) and composite Petrov–Galerkin method (10.1) utilizing a value of $M = M^*$ which produces smallest errors (results relate to example 11.1)

h	$\alpha = \alpha^*$	$\alpha = \begin{cases} \alpha^*, & 1 \leqslant j \leqslant N-M^* \\ \alpha, & N-M^* < j \leqslant N \end{cases}$	M^*	$\alpha = \hat{\alpha}$
1/5	.383	.363(−2)	1	.933(−2)
1/10	.148	.355(−3)	2	.147(−2)
1/20	.285(−1)	.259(−4)	3	.139(−3)
1/40	.281(−2)	.173(−5)	7	.102(−4)
1/80	.164(−3)	.109(−6)	14	.661(−6)

The solution is

$$u(x) = x^3 + 3k^{-1}x + 6k^{-2}x - (1 + 3k^{-1} + 6k^{-2})(e^{kx} - 1)(e^k - 1). \tag{11.2}$$

We shall again choose $k = 60$ throughout.

Table 5 contains the maximum nodal errors for the methods discussed in part 1 together with the totally disconnected Petrov–Galerkin method which utilizes the value $\alpha = \tilde{\alpha}$ (3.2) throughout the entire interval. For methods based on $\alpha = kh/6$, 0 and 1 the results are virtually the same as those of example 7.2, indicating that the error can be largely attributed to the boundary layer component. The totally disconnected method is particularly good when $kh \geqslant 6$ but diverges thereafter as $h \to 0$ (see (10.6) and the associated discussion). The energies are not tabulated for this example since they, too, are virtually identical to those of example 7.2 (table 4).

Turning next to the methods discussed in part 2, table 6 contains the maximum nodal errors of the composite method of section 10, which is generated by the test functions of (10.1). Results are presented for $M = 0$ (optimized Petrov–Galerkin method), $M = N$ (optimized PGL method) and the truly composite method which has the optimal value of M, say M^*, leading to smallest nodal errors. The quantity M^*h is comparable with the boundary layer thickness defined by (8.10), confirming our expectation that L-splines are required only within an appropriately defined boundary layer.

In table 7, the nodal errors are presented for the disconnected method generated by the test functions Ψ_h^d of (10.3) with $h = 1/5$ and $M = 1$. In addition to the value α^*, we also use $\alpha = 1 - 2/kh$, 0 and 1 for $1 \leqslant j \leqslant N-1$; the accuracy of all methods is seen to be significantly better than those used in table 5 for this same problem. It is interesting to note that the central difference

Table 7. Nodal values of the error $u(x) - u_h(x)$, when $h = 1/5$, for disconnected Petrov–Galerkin method ((10.3) with $M = 1$, and α^* replaced by different values of α) applied to example 11.1

x \ α	$kh/6$	$1-2/kh$	0	1
0.2	.117(−3)	.233(−2)	.461(−2)	.200(−2)
0.4	.433(−3)	.467(−2)	.776(−2)	.400(−2)
0.6	.129(−2)	.700(−2)	.129(−1)	.603(−2)
0.8	.362(−2)	.933(−2)	.153(−1)	.835(−2)

method ($\alpha = 0$) is oscillation-free even though $kh > 2$; this conforms with (6.5) and the subsequent discussion. It is observed from further numerical experimentation that allowing $h \to 0$, all disconnected methods have roughly the same maximum nodal error – i.e. approximately unity. This is to be expected since they all converge towards the solution of (10.4). They still, however, produce good approximations to (11.1) outside the boundary layer.

We conclude our numerical examples with a brief investigation of problems which have a pronounced variation on [0,1] in addition to a boundary layer component. Due to the linearity of our model problem, the performance of the methods of part 1 in these situations can be deduced by superposing the results of tables 2 and 3. By so doing we see that the additionof this component into the solution has little or no effect on the accuracy, this being dictated almost entirely by the boundary layer behaviour. To analyse the corresponding effect for disconnected methods, we apply them to:

Example 11.2. $$-u'' + ku' = f(x), \quad x \in (0,1),$$

$$u(0) = u(1) = 0$$

With $f(x) = 3kx^2 + 4\pi^2 \sin 2\pi x + 2k\pi \cos 2\pi x$, the solution is

$$u(x) = \sin 2\pi x + x^3 + 3k^{-1}x^2 + 6k^{-2}x - (1 + 3k^{-1} + 6k^{-2})(e^{kx} - 1)/(e^k - 1).$$

Table 8 lists the nodal errors when the disconnected method with $M = 1$ and $h = 1/5$ is applied with $\alpha = kh/6$, 0 and 1. We remark only that, although the errors are increased by an order of magnitude over those of table 7 for all three methods, the relative performance of the methods is unchanged.

Based on the above results and further experimentation, we can tentatively recommend an optimal strategy: for problems (9.1) which contain at most a mild boundary layer behaviour (e.g. $\int_0^1 f(x)\,dx = O(1)$ when $f = O(k)$) the optimized Petrov–Galerkin method should be used. When the boundary layer component is significant ($\int_0^1 f(x)\,dx = O(k)$ when $f = O(k)$), the optimized Petrov–Galerkin method should be coupled with the disconnected method (10.3) employing $M = 1$ if kh exceeds 4 or 5. Good results are also obtained by coupling central differences ($\alpha = 0$) or backward differences ($\alpha = 1$) with a disconnected method, although it cannot be guaranteed that the solution will be oscillation free when $\alpha = 0$ and $kh > 2$.

Table 8. Nodal values of error corresponding to the methods of table 7 applied to example 11.2

x \ α	$kh/6$	$1 - 2/kh$	0	1 1
0.2	−.267(−1)	−.396(−1)	−.112(0)	−.321(−1)
0.4	−.369(−1)	.236(−1)	−.559(−1)	.201(−1)
0.6	−.262(−1)	.105(0)	.164(0)	.870(−1)
0.8	−.356(−1)	.960(−1)	.154(0)	.777(−1)

References

[1] R.L. Lee, P.M. Gresho and R.L. Sani, A comparative study of certain finite-element and finite difference methods in advection-diffusion simulations (preprint).
[2] O.C. Zienciewicz, in: R.H. Callagher et al. (eds.), Proceedings of the Second International Symposium on Finite Element Methods in Flow Problems (Wiley-Interscience, New York, 1976).
[3] P.J. Roache, Computational fluid dynamics, 2nd ed. (Hermosa Publishers, Albuquerque, NM, 1976).
[4] I. Christie, D.F. Griffiths, A.R. Mitchell and O.C. Zienkiewicz, Finite element methods for second order differential equations with significant first derivatives, Inter. J. Numer. Meths. Eng. 10 (1976) 1379–1387.
[5] R. Anderson and A.R. Mitchell, Petrov–Galerkin methods, (Univ., Dundee, Numerical Analysis Rept. No. 17, 1976).
[6] A.K. Aziz (ed.), The mathematical foundations of the finite element method with applications to partial differential equations (Academic Press, New York, 1972).
[7] I. Christie, Ph.D. dissertation. Univ. Dundee, 1977 (in preparation).
[8] W.-J. Beyn and J. Lorenz., On convergence of finite element methods for non-coercive problems (Department of Mathematics and Statistics, Calgary, Research Paper No. 330, 1976).
[9] M.H. Schultz, Spline analysis (Prentice-Hall, Englewood Cliffs, NJ, 1973).
[10] J.T. Oden and J.N. Reddy, An introduction to the mathematical theory of finite elements (Wiley-Interscience, New York, 1976).
[11] S. Agmon, Lectures on elliptic boundary value problems (Van Nostrand, Princeton, NJ, 1965).
[12] R.S. Varga, Matrix iterative analysis (Prentice-Hall, Englewood Cliffs, NJ, 1962).
[13] F.R. Gantmacher and M.G. Krein, Oscillatory matrices and kernels and small vibrations of mechanical systems, 2nd ed. [Russian] (Moscow, 1937).
[14] G.F. Carrier and C.E. Pearson, Ordinary differential equations, (Blaisdell, Waltham, MA, 1968).
[15] R.S. Varga, Functional analysis and approximation theory in numerical analysis (SIAM Publications, 1971).
[16] J.J. Miller, A finite element method for a two point boundary value problem with a small parameter affecting the highest derivative (Trinity College. Dublin, Report. TCD-1955-11, 1975).