IMA Journal of Numerical Analysis (1981) **1**, 253–266

Product Approximation for Non-linear Problems in the Finite Element Method

I. CHRISTIE

Department of Mathematics, Napier College, Edinburgh

D. F. GRIFFITHS AND A. R. MITCHELL

Department of Mathematics, University of Dundee, Dundee, Scotland

AND

J. M. SANZ-SERNA

Department of Mathematics, Facultad de Ciencias, Valladolid, Spain

[Received 8 August 1980]

The term "product approximation" is used to refer to a finite element technique for non-linear problems which has appeared several times in the literature under different presentations. The aims of this paper are to give a unified approach to the product integration technique and to provide new evidence for the fact that it can be a good alternative to the standard Galerkin approximation in certain circumstances.

1. Introduction

THE TERM "product approximation" is used to refer to a finite element technique for *non-linear* problems which has appeared several times in the literature under different presentations (see end of this section). The aims of this paper are to give a unified approach to the product integration technique and to provide new evidence for the fact that it can be a good alternative to the standard Galerkin approximation (S.G.A.) for non-linear problems in some circumstances. Rather than start with the general situation let us describe the product approximation (P.A.) idea as applied to two particular problems.

As a first example consider the non-linear two point boundary value problem

$$-u''+f(u) = 0, \quad 0 < x < 1, \tag{1.1}$$

$$u(0) = u(1) = 0 \tag{1.2}$$

which is assumed to have a unique solution. We introduce a mesh

$$0 = x_0 < x_1 < \ldots < x_k < x_{k+1} = 1$$

and denote by $\phi_i(x)$, $i = 1, 2, \ldots, k$ the usual basis functions for the space of continuous piecewise linear functions satisfying the boundary conditions (1.2). The

0272–4979/81/030266+14 $02.00/0

S.G.A. to $u(x)$ is obtained from the equations

$$\Big(\sum_i U_i\,\phi_i',\,\phi_j'\Big)+\Big(f\big(\sum_i U_i\phi_i\big),\,\phi_j\Big)=0,\quad 1\leqslant j\leqslant k, \tag{1.3}$$

where U_i is the value of the Galerkin approximation at x_i, a dash denotes differentiation with respect to x, and $(\,,)$ denotes the L^2-inner product. The P.A. is defined by the equations

$$\Big(\sum_i U_i\phi_i',\,\phi_j'\Big)+\Big(\sum_i f(U_i)\phi_i,\,\phi_j\Big)=0,\quad 1\leqslant j\leqslant k, \tag{1.4}$$

where u and $f(u)$ have been approximated by independent piecewise linear functions. Since (1.4) can be written in the form

$$\sum_i U_i(\phi_i',\,\phi_j')+\sum_i f(U_i)(\phi_i,\,\phi_j)=0,\quad 1\leqslant j\leqslant k, \tag{1.5}$$

it is clear that in order to determine the P.A. solution we can first compute the inner products (ϕ_i, ϕ_j), (ϕ_i', ϕ_j') and then solve the resulting system of non-linear equations. On the other hand to obtain the S.G.A. solution, the computation of the inner products and the solution of the equations are not separate processes and *at each step* of the iteration, numerical quadrature must be used in order to evaluate the contribution of the non-linear term. Thus, in general, P.A. requires much less computational effort, which as we shall see in Section 3 may be achieved without sacrificing accuracy.

As a second example consider the non-linear conservation law

$$u_t+\left(\frac{u^2}{2}\right)_x=0,\quad -\infty<x<+\infty,\quad t>0 \tag{1.6}$$

together with square integrable initial data. We keep the time continuous and discretize in space by means of piecewise linear elements based on the nodes x_i, $i=0, \pm 1, \pm 2, \ldots$. The S.G.A. equations are

$$\Big(\sum_i \dot U_i(t)\phi_i,\,\phi_j\Big)-\tfrac{1}{2}\Big(\big(\sum_i U_i(t)\phi_i\big)^2,\,\phi_j'\Big)=0,\quad j=0,\pm 1,\pm 2,\ldots \tag{1.7}$$

(a dot denotes differentiation with respect to t), and the P.A. system is

$$\Big(\sum_i \dot U_i(t)\phi_i,\,\phi_j\Big)-\tfrac{1}{2}\Big(\sum_i U_i^2(t)\phi_i,\,\phi_j'\Big)=0,\quad j=0,\pm 1,\pm 2,\ldots. \tag{1.8}$$

When the elements are of uniform length h, (1.7) and (1.8) reduce to

$$G_h(U_j)=\frac{1}{6}(\dot U_{j+1}+4\dot U_j+\dot U_{j-1})+\frac{1}{6h}(U_{j+1}+U_j+U_{j-1})(U_{j+1}-U_{j-1})=0, \tag{1.9}$$

and

$$P_h(U_j)=\frac{1}{6}(\dot U_{j+1}+4\dot U_j+\dot U_{j-1})+\frac{1}{4h}(U_{j+1}^2-U_{j-1}^2)=0, \tag{1.10}$$

respectively where $j=0, \pm 1, \pm 2, \ldots$.

Before we compare (1.9) and (1.10) it is worth considering the linear analogue of (1.6)

$$u_t+u_x = 0. \tag{1.11}$$

With linear elements of uniform length h, the Galerkin method yields the system

$$L_h(U_j) = \frac{1}{6}(\dot{U}_{j+1}+4\dot{U}_j+\dot{U}_{j-1})+\frac{1}{2h}(U_{j+1}-U_{j-1}) = 0, \qquad j = 0, 1, 2, \dots \tag{1.12}$$

which enjoys two important properties.

(i) It is fourth-order accurate, for if $v(x, t)$ is a smooth function

$$\frac{1}{6}(\dot{V}_{j+1}+4\dot{V}_j+\dot{V}_{j-1}) = \dot{v}_j+\frac{h^2}{6}(\dot{v}_{xx})_j+O(h^4), \tag{1.13}$$

and

$$\frac{1}{2h}(V_{j+1}-V_{j-1}) = (v_x)_j+\frac{h^2}{6}(v_{xxx})_j+O(h^4), \tag{1.14}$$

so that for the solution u of (1.11)

$$L_h(u_j) = (u_t+u_x)_j+\frac{h^2}{6}[(u_t+u_x)_{xx}]_j+O(h^4) = O(h^4). \tag{1.15}$$

It should be emphasized that this high order local accuracy at the grid points is not accomplished by approximating each term u_t, u_x to fourth-order, but by approximating u_t, u_x to $O(h^2)$ in such a way that second-order terms cancel each other through application of the differential equation (cf. compact differencing).

(ii) It is conservative in the sense that

$$\frac{d}{dt}\int \Big(\sum_j U_j\phi_j\Big)^2\,dx = \frac{d}{dt}\sum_j (U_jU_{j+1}+4U_j^2+U_jU_{j-1}) = 0. \tag{1.16}$$

This conservation of a quadratic quantity follows directly from the Galerkin equations and it is desirable not only because it reproduces a property of the equation (1.11) but also because it ensures numerical stability. (For related material see Morton, 1977.)

We now return to the non-linear equation (1.6) and its discretizations (1.9) and (1.10) and ask whether properties (i) and (ii) hold. Looking first at accuracy (i.e. local accuracy at grid points) one has, replacing v by $v^2/2$ in (1.14),

$$\frac{1}{4h}(V_{j+1}^2-V_{j-1}^2) = \left[\left(\frac{v^2}{2}\right)_x\right]_j+\frac{h^2}{6}\left[\left(\frac{v^2}{2}\right)_{xxx}\right]_j+O(h^4), \tag{1.17}$$

which implies

$$P_h(u_j) = \left[u_t+\left(\frac{u^2}{2}\right)_x\right]_j+\frac{h^2}{6}\left\{\left[u_t+\left(\frac{u^2}{2}\right)_x\right]_{xx}\right\}_j+O(h^4), \tag{1.18}$$

so that the P.A. retains the accuracy of the linear case. This is a consequence of the fact that in the P.A., the term u^2 has been treated as a variable in its own right, so that the Taylor expansions of the linear case are carried over to the non-linear problem. On the other hand for the S.G.A., one has

$$\frac{1}{6h}(V_{j+1}+4V_j+V_{j-1})(V_{j+1}-V_{j-1}) = \left[\left(\frac{v^2}{2}\right)_x\right]_j + \frac{h^2}{6}(vv_{xxx}+2v_xv_{xx})_j+O(h^4). \tag{1.19}$$

Here the $O(h^2)$ terms do not reproduce a derivative of v^2, the cancellation cannot take place, and

$$G_h(u_j) = O(h^2). \tag{1.20}$$

The estimates (1.18) and (1.20) depend strongly on the fact that local accuracy is measured pointwise. Were it to be measured in a more global sense, using L_2 norms for example, P.A. would be only $O(h^2)$ accurate whereas S.G.A. would be $O(h^4)$ accurate. Which scheme is the more "accurate" therefore depends on the interpretation we give to the numerical solution (the best fit at nodes or the best L_2 fit by a piecewise linear function). These and related issues are discussed in some detail by Cullen and Morton (1980). Within the context of this paper, the methods and analysis are motivated by the pointwise properties of P.A. as it applies to the various problems. It seems more appropriate, therefore, that accuracy be measured pointwise and so this approach will be adopted for the remainder of the paper. On this basis, it may be argued that P.A. schemes more closely resemble finite difference than finite element methods.

Turning now to conservation properties, it is well known that $G_h(U_j)$ preserves the quadratic quantity in (1.16) whilst $P_h(U_j)$ does not do so. This loss of conservation properties results in non-linear instability of the scheme (1.10). In fact Fornberg (1973) proved that if the mass is "lumped" in (1.10) initial data can be found for which the numerical solution blows up in finite time. Recently Chin, Hedstrom & Karlsson (1979) have extended Fornberg's conclusions to the "unlumped" case. However, these instabilities occur only for values of u close to zero since the approximation of the non-linear term cannot distinguish between positive and negative data and can be removed by the use of artificial viscosity. Overall (1.10) can be a useful scheme as is shown by the numerical computations of Chin *et al.* (1979) and Turkel (1980).

The examples above illustrate why P.A. may be more advantageous than S.G.A. The primary reason is computational ease. However, it may occur (as for equation 1.6) that P.A. also leads to increased local accuracy. The idea behind P.A. is so simple that it is not surprising that many authors have used it without attempting a comparison with S.G.A. These authors include Swartz and Wendroff (1969, 1974*a*), Chin *et al.* (1979) for first-order hyperbolic equations, Fletcher (1979) and Lucchi (1980) for time-dependent problems in fluids and Sanz-Serna and Christie (1980) for non-linear dispersive waves.

An outline of the rest of the paper is as follows: Section 2 contains a more general description of the method, Section 3 deals with the elliptic problem where optimal order of convergence in the energy norm is proved. Sections 4 and 5 are concerned with Burgers and Korteweg de Vries equations, respectively.

2. Description of the Method

Returning to the first example considered in the introduction, it is convenient to reformulate Galerkin's method in a way which does not involve the nodal values U_i or the basis functions ϕ_i explicitly. If we denote by S the finite dimensional space spanned by the functions ϕ_i, the S.G.A.

$$U(x) = \sum_i U_i \phi_i(x),$$

where the values U_i are given by (1.3), solves the problem: find $U \in S$ such that $\forall\ \phi \in S$

$$(U', \phi') + (f(U), \phi) = 0. \tag{2.1}$$

To derive a similar expression for the P.A.

$$U^*(x) = \sum_i U_i \phi_i(x),$$

where the values U_i are now given by (1.4), let us note that $\sum_i f(U_i)\phi_i$ is the unique element in S which interpolates $f(\sum_i U_i\phi_i)$ at the nodes x_i, so that if we denote by Q_S the operator which maps each function onto its piecewise linear interpolant, U^* is a solution of the problem: find $U^* \in S$ such that $\forall\ \phi \in S$

$$(U^{*\prime}, \phi') + (Q_S f(U^*), \phi) = 0. \tag{2.2}$$

We are now in a position to treat the general case. Let us consider the partial differential equation

$$f(\mathbf{x}, u) + \sum_{i=1}^{r} \left\{ \frac{\partial}{\partial x_i} F_i(\mathbf{x}, u) - \frac{\partial}{\partial x_i}\left(a_i(\mathbf{x}) \frac{\partial}{\partial x_i} G_i(\mathbf{x}, u) \right) \right\} = 0 \tag{2.3}$$

where $\mathbf{x} = \{x_1, x_2, \ldots, x_r\}^T$, which is to hold in a region of $\mathbf{x}$ space with appropriate boundary conditions. In a weak form, for some appropriate function space X, the solution of (2.3) solves: find $u \in X$ such that $\forall \phi \in X$

$$(f(\mathbf{x}, u), \phi) + \sum_{i=1}^{r} \left\{ \left(\frac{\partial}{\partial x_i} F_i(\mathbf{x}, u), \phi \right) + \left(a_i(\mathbf{x}) \frac{\partial}{\partial x_i} G_i(\mathbf{x}, u), \frac{\partial \phi}{\partial x_i} \right) \right\} = 0. \tag{2.4}$$

Now let S be a finite dimensional subspace of X. Then the S.G.A. U to u solves: find $U \in S$ such that $\forall\ \phi \in S$

$$(f(x, U), \phi) + \sum_{i=1}^{r} \left\{ \left(\frac{\partial}{\partial x_i} F_i(\mathbf{x}, U), \phi \right) + \left(a_i(\mathbf{x}) \frac{\partial}{\partial x_i} G_i(\mathbf{x}, U), \frac{\partial \phi}{\partial x_i} \right) \right\} = 0. \tag{2.5}$$

Furthermore, if we assume that it is possible to define an operator Q_S which associates with each function an interpolant in S, we define the P.A., U^* to u, to be the solution of: find $U^* \in S$, such that $\forall\ \phi \in S$

$$(Q_S f(\mathbf{x}, U^*), \phi) + \sum_{i=1}^{r} \left\{ \left(\frac{\partial}{\partial x_i} Q_S F_i(\mathbf{x}, U^*), \phi \right) + \left(a_i(x) \frac{\partial}{\partial x_i} Q_S G_i(\mathbf{x}, U^*), \frac{\partial \phi}{\partial x_i} \right) \right\} = 0. \tag{2.6}$$

The extension of the method to cover semi-discrete schemes for time dependent problems offers no additional difficulties. Expressions of the form $Q_S F(\mathbf{x}, U^*)$ cause no difficulty when S is a finite element subspace of Lagrange or Hermite type and can be written in terms of the nodal values. The case of splines is more complicated, as the construction of $Q_S F$ involves the solution of a system of linear equations. This difficulty was overcome, however, by the observation of Swartz and Wendroff (1974*b*) that spline Galerkin methods for linear one-dimensional problems are equivalent to collocation methods which, of course, involve only nodal values. This simplified Galerkin technique is implemented by Chin *et al.* (1979).

We end this section with two remarks.

(i) It might happen for a given problem that the S.G.A. and the P.A. equations are identical. For example, if piecewise linear elements are used for the problem

$$[F(u)]_{xx} = g(x), \qquad u(0) = u(1) = 0, \tag{2.7}$$

it is easily seen that both procedures are equivalent. In fact in the term $[dF(U)/dx, d\phi/dx]$, the function $d\phi/dx$ is piecewise constant, and so the inner product can be expressed in terms of $F(U)$ evaluated at the nodes, where it equals $Q_S F(U)$. A non-linearity of the form (2.7) occurs for example in the Boussinesq equation.

(ii) The operator Q_S is chosen to be of interpolatory type on the grounds of computational simplicity. However, projections onto S of a different character have been used before. (See for example Cullen, 1974; Cullen & Morton, 1980).

3. The Elliptic Case

For simplicity of exposition we shall be concerned only with one-dimensional problems, although our results can be proved with small changes in higher dimensional cases.

Consider the two-point boundary value problem

$$L[u(x)] = f[x, u(x)], \quad 0 < x < 1, \tag{3.1}$$

with homogeneous boundary conditions

$$D^k u(0) = D^k u(1) = 0, \quad 0 \leqslant k \leqslant n-1, \quad D \equiv \frac{d}{dx}, \tag{3.2}$$

where

$$L[u(x)] = \sum_{j=0}^{n} (-1)^{j+1} D^j [p_j(x) D^j u(x)], \quad n \geqslant 1, \tag{3.3}$$

and the coefficients are real functions in $C^j[0, 1]$. Ritz–Galerkin methods for this problem have been analysed by Ciarlet, Schultz & Varga (1967, 1969), Perrin, Price & Varga (1969), and Herbold, Schultz & Varga (1969).

Denote by H^n the Sobolev space of real functions on $(0, 1)$ whose first n derivatives are square integrable, and by H_0^n the closure in H^n of the set of all infinitely differentiable functions with support in $(0, 1)$. The following hypotheses on the problem (3.1)–(3.3) are supposed to hold. First we assume that there exist two real constants β and K such that for all $w \in H_0^n$,

$$||w||_\infty \equiv \sup |w(x)| \leqslant K\left\{\int_0^1 \left[\sum_{j=0}^n p_j(x)[D^jw(x)]^2+\beta[w(x)]^2\right] dx\right\}^{\frac{1}{2}}. \tag{3.4}$$

This hypothesis is automatically fulfilled when L is strongly elliptic, i.e. when $p_n(x)$ is positive in $[0,1]$. Next we introduce the finite quantity (cf. Ciarlet *et al.*, 1967, lemma 1) given by

$$\Lambda = \inf_{\substack{w \in H_0^n,\\ w \not\equiv 0.}} \frac{\int_0^1 \left\{\sum_{j=0}^n p_j(x)[D^jw(x)]^2\right\} dx}{\int_0^1 [w(x)]^2\, dx}, \tag{3.5}$$

Λ is a lower bound for the eigenvalues of the associated eigenvalue problem

$$L[u(x)]+\lambda u(x) = 0, \quad 0 < x < 1,$$

subject to boundary conditions (3.2). We then assume that $f(x, u)$, $\partial f(x, u)/\partial u$, are real and continuous, and that the following monotonicity requirement holds: there exists a constant γ such that

$$\frac{f(x, u)-f(x, v)}{u-v} \geqslant \gamma > -\Lambda. \tag{3.6}$$

Finally it is easily proved that

$$||w||_\gamma = \left\{\int_0^1 \left[\sum_{j=0}^n p_j(x)[D^jw(x)]^2+\gamma[w(x)^2]\right] dx\right\}^{\frac{1}{2}} \tag{3.7}$$

is a norm on H_0^n which satisfies

$$||w||_\infty \leqslant K||w||_\gamma \quad \forall\ w \in H_0^n. \tag{3.8}$$

Now consider a finite dimensional subspace $S \subset H_0^n$ with associated projection operator Q_S, as in Section 2, and suppose that the S.G.A., $U \in S$ and the P.A., $U^* \in S$ exist.

THEOREM 1. Under the previous hypotheses

$$||U-U^*||_\gamma \leqslant K||Q_S f[x, U^*(x)]-f[x, U^*(x)]||_\infty, \tag{3.9}$$

where K is the constant of (3.4) and is independent of S.

Proof. U, U^* satisfy

$$\left(\sum_{j=0}^n p_j(x)D^jU(x), D^j\phi(x)\right)+(f[x, U(x)], \phi) = 0 \tag{3.10}$$

and

$$\left(\sum_{j=0}^{n} p_j(x)D^jU^*(x), D^j\phi(x)\right) + (Q_S f[x, U^*(x)], \phi) = 0 \tag{3.11}$$

respectively for all $\phi \in S$. Subtract and set $\phi = U - U^*$ to get

$$\left(\sum_{j=0}^{n} p_j(x)D^j[U(x) - U^*(x)], D^j[U(x) - U^*(x)]\right) +$$
$$(f[x, U(x)] - Q_s f[x, U^*(x)], U(x) - U^*(x)) = 0. \tag{3.12}$$

If we simplify the notation by setting

$$f[x, U(x)] \equiv f(U) \quad \text{and} \quad f[x, U^*(x)] \equiv f(U^*),$$

(3.12) becomes

$$\|U - U^*\|_\gamma^2 - \gamma \int_0^1 [U(x) - U^*(x)]^2\, dx + (f(U) - Q_S f(U^*), U - U^*) = 0. \tag{3.13}$$

Now

$$(f(U) - Q_S f(U^*), U - U^*) = (f(U) - f(U^*), U - U^*) + (f(U^*) - Q_S f(U^*), U - U^*), \tag{3.14}$$

and the hypothesis (3.6) implies that

$$(f(U) - f(U^*), U - U^*) \geqslant \gamma \int_0^1 [U(x) - U^*(x)]^2\, dx. \tag{3.15}$$

Taking (3.14) and (3.15) into account, (3.13) gives

$$\|U - U^*\|_\gamma^2 \leqslant |(f(U^*) - Q_S f(U^*), U - U^*)|. \tag{3.16}$$

We now bound the right-hand side of (3.16) to obtain

$$\|U - U^*\|_\gamma^2 \leqslant \|f(U^*) - Q_S f(U^*)\|_\infty \|U - U^*\|_\infty, \tag{3.17}$$

and the result follows from (3.8).

Let us now discuss the implications of the theorem. Suppose L is of second order ($n = 1$) and that S is the Lagrange space of continuous functions in $(0, 1)$ which vanish at $x = 0, 1$ and whose restrictions to the intervals $(0, h), (h, 2h), \ldots,$ are polynomials of degree $\leqslant m$. (Here $1/h$ is an integer N.) Then it is well known that in the norm $\|\cdot\|_\gamma$ the distance between the solution u of (3.1)–(3.3) and the S.G.A. is $\|u - U\|_\gamma = O(h^m)$ (i.e. optimal rate of convergence). Also, according to the standard theory of interpolation we have

$$\|f(U^*) - Q_S f(U^*)\|_\infty = O(h^{m+1})$$

provided that f is smooth and U^* has bounded derivatives within each element as $h \to 0$. We conclude from (3.9) that $\|U - U^*\|_\gamma = O(h^{m+1})$, which in turn implies $\|u - U^*\|_\gamma = O(h^m)$. Thus the optimal rate of convergence when using P.A. is retained and the distance between U and U^* is asymptotically negligible when compared with the distance between u and U. An additional feature of the result (3.9), which we have not exploited, is that it provides an *a posteriori* computable bound on the difference

TABLE 1

|Error| at $x = 0{\cdot}5$, linear elements

N	P.A.	S.G.A.
2	$2{\cdot}17 \times 10^{-3}$	$2{\cdot}06 \times 10^{-3}$
4	$5{\cdot}19 \times 10^{-4}$	$4{\cdot}98 \times 10^{-4}$
6	$2{\cdot}28 \times 10^{-4}$	$2{\cdot}20 \times 10^{-4}$
8	$1{\cdot}28 \times 10^{-4}$	$1{\cdot}23 \times 10^{-4}$
10	$4{\cdot}16 \times 10^{-5}$	$4{\cdot}02 \times 10^{-5}$
20	$2{\cdot}03 \times 10^{-5}$	$1{\cdot}97 \times 10^{-5}$

between U and U^*. Note that P.A. does not enjoy the "optimal approximation property in energy norm" that normally applies to Galerkin approximations of elliptic problems.

In order to assess the performance of the two techniques, we apply both to the problem

$$D^2u(x) = \exp[u(x)], \qquad u(0) = u(1) = 0 \tag{3.18}$$

with solution

$$u(x) = -\ln 2 + 2\ln\{c \sec[c(x-\tfrac{1}{2})/2]\}, \quad c \approx 1.33. \tag{3.19}$$

The bound (3.4) is valid with $K = \frac{1}{2}$, $\beta = 0$. Also $\Lambda = \pi^2$ and γ can be chosen as 0 or 1 to make $\|\cdot\|_\gamma$ a Sobolev norm. Equation (3.18) was solved by the S.G.A. and P.A. methods based on linear and quadratic Lagrange elements of uniform length $h = 1/N$. Although the theorem covers the energy norm, we quote the behaviour of the error at the nodes. Only the results at the node $x = 0{\cdot}5$ are displayed as other nodes follow the same pattern. For linear elements we see in Table 1 that both methods perform almost identically and the nodal errors are $O(h^2)$ as expected.

The results corresponding to quadratic elements are displayed in Table 2. It is convenient to separate the cases of $x = 0{\cdot}5$ being (a) an integer or (b) a half-integer node.

For integer nodes the error associated with the P.A. is four times that associated with the S.G.A. and both appear to be $O(h^4)$. For the half integer nodes the picture is reversed and the P.A. is more accurate. These differences may be attributed to the fact

TABLE 2

|Error| at $x = 0{\cdot}5$, quadratic elements

	(a) Integer node			(b) Half-integer node	
N	P.A.	S.G.A	N	P.A.	S.G.A.
2	$25{\cdot}2 \times 10^{-4}$	$4{\cdot}23 \times 10^{-4}$	3	$0{\cdot}645 \times 10^{-6}$	$3{\cdot}80 \times 10^{-6}$
4	$16{\cdot}5 \times 10^{-7}$	$4{\cdot}78 \times 10^{-7}$	5	$0{\cdot}159 \times 10^{-7}$	$4{\cdot}76 \times 10^{-7}$
10	$4{\cdot}27 \times 10^{-8}$	$1{\cdot}11 \times 10^{-8}$	7	$0{\cdot}0005 \times 10^{-7}$	$1{\cdot}24 \times 10^{-7}$

that, for small values of N, the error alternates in sign at the nodes. Close to $N = 7$ the error for P.A. changes sign at the half integer nodes explaining the low value in Table 2(b). For larger values of N the error increases initially although it is asymptotically $O(h^4)$ and consistently better than S.G.A.

4. Burgers Equation

In this section we shall concern ourselves with Burgers equation

$$u_t + uu_x - \varepsilon u_{xx} = 0, \quad \varepsilon > 0. \tag{4.1}$$

It is known that for small values of ε the solution develops steep fronts and numerical methods are likely to produce results including large non-physical oscillations unless the element size is unrealistically small. In finite difference methods upwinding has been the usual technique to avoid unwanted disturbances and Christie, Griffiths, Mitchell & Zienkiewicz (1976) have shown how upwinding can be simulated in the finite element method by choosing test functions which are different from the trial functions (Petrov–Galerkin). For instance in the steady linearized version of (4.1), Christie & Mitchell (1978) have shown that if u is approximated by

$$U(x) = \sum_{i=0}^{2N} B_{i/2}(x) U_{i/2} \tag{4.2}$$

on a grid of elements of size h, where the trial functions $B_{i/2}(x)$ are the usual Lagrange quadratics, then upwinding can be introduced into the Galerkin method by adding a cubic perturbation to the quadratic test functions to give the latter as

$$T_j(x) = B_j(x) + \alpha_1 \sigma\left(\frac{x}{h} - j\right) + \alpha_2 \sigma\left(\frac{x}{h} - j + 1\right) \tag{4.3}$$

at an integer node and

$$T_{j-\frac{1}{2}}(x) = B_{j-\frac{1}{2}}(x) - 4\alpha_3 \sigma\left(\frac{x}{h} - j\right) \tag{4.4}$$

at a half-integer node where

$$\sigma(s) = \begin{cases} -40s(s+\frac{1}{2})(s+1) & -1 \leqslant s \leqslant 0 \\ 0 & \text{elsewhere} \end{cases} \tag{4.5}$$

and $\alpha_1, \alpha_2, \alpha_3$ are the upwinding parameters. Full upwinding in the quadratic case is given by $\alpha_1 = \alpha_3 = 1$, $\alpha_2 = 0$.

We now use the trial and test functions (4.2)–(4.5) with values of $\alpha_1, \alpha_2, \alpha_3$ for full upwinding for the numerical solution of Burgers equation using S.G.A. and P.A. techniques. The latter is easily extended to the case of different trial and test functions. Two problems are chosen to illustrate the behaviour of each method for a range of values of the parameter ε. The first problem has initial and boundary values

$$\begin{aligned} & u(x,0) = \sin \pi x, && 0 < x < 1 \\ \text{and} \quad & u(0,t) = u(1,t) = 0, && t \geqslant 0 \end{aligned} \tag{4.6}$$

respectively. The theoretical solution of this problem given by Cole (1951) exhibits a steep front near $x = 1$ which broadens and dies out as t increases, leaving a "sine" wave of reduced amplitude. For the second problem we proceed in reverse order: via the Hopf transformation (Hopf, 1950) we construct a particular solution of (4.1) given by

$$u(x,t) = f(\xi), \qquad \xi = x - \mu t - \beta \tag{4.7}$$

where

$$f(\xi) = [\mu + \alpha + (\mu - \alpha)e^{\alpha\xi/\varepsilon}]/(1 + e^{\alpha\xi/\varepsilon}) \tag{4.8}$$

and α, β and μ are arbitrary constants. The initial data and the boundary conditions at $x = 0, 1$ are then taken from (4.7). The solution (4.7) represents a travelling wave front, positioned initially at $x = \beta$, travelling with speed μ and such that $u(x,t) \to \mu \mp \alpha$ as $x \to \pm\infty$ for any t. For our experiments we chose the constants to be $\alpha = 0{\cdot}4$, $\beta = 0{\cdot}125$ and $\mu = 0{\cdot}6$.

Tables 3 and 4 show the results for the first problem and Table 5 those for the second problem. In all the numerical experiments we used nine elements of uniform length and the integration in time is carried out using the trapezoidal rule with $\Delta t = 0{\cdot}001$ so that errors can be attributed only to the discretization in space. For comparison we have also included the results corresponding to the compact differencing technique of Hirsh (1975). (See also Mitchell & Griffiths, 1980.) This is probably the most successful of the finite difference techniques for solving Burgers equation. It may be that the compact differencing results could be improved by introducing a degree of upwinding. The odd-number nodes in Tables 3–5 are integer nodes and the even number nodes are half-integer nodes.

Numerical results were also obtained for the wave-front problem with $\varepsilon = 0{\cdot}0001$ and although the P.A. method was best the results were comparatively poor for all methods. In this case the grid size, $h = \frac{1}{9}$ is larger than the breadth of the wave front and so no numerical method can give accurate results in the vicinity of the wave front.

TABLE 3
Sin initial condition ($\varepsilon = 0{\cdot}01$, $t = 0{\cdot}5$)

		Petrov–Galerkin		
Node	Exact	S.G.A.	P.A.	Compact differencing
10	0·589	0·589	0·589	0·589
11	0·649	0·649	0·649	0·648
12	0·707	0·707	0·707	0·709
13	0·762	0·762	0·762	0·760
14	0·814	0·813	0·814	0·820
15	0·861	0·860	0·861	0·852
16	0·902	0·895	0·907	0·917
17	0·934	0·911	0·952	0·911
18	0·937	0·764	0·774	0·964
19	0	0	0	0

TABLE 4

Sin initial condition ($\varepsilon = 0{\cdot}0001$, $t = 0{\cdot}5$)

		Petrov–Galerkin		
Node	Exact	S.G.A.	P.A.	Compact differencing
10	0·595	0·594	0·595	0·619
11	0·656	0·656	0·656	0·621
12	0·715	0·715	0·715	0·764
13	0·772	0·772	0·772	0·718
14	0·826	0·826	0·826	0·887
15	0·876	0·876	0·876	0·819
16	0·921	0·906	0·921	0·978
17	0·959	0·902	0·960	0·908
18	0·959	0·835	0·854	1·046
19	0	0	0	0

It is worth pointing out that in the first problem (see Tables 3 and 4) the Petrov–Galerkin method, particularly the P.A. version, gives improved results as ε is reduced. This is because the method employs fully upwinded quadratics which become more appropriate as $\varepsilon \to 0$. With a piecewise quadratic approximant it is also unrealistic to

TABLE 5

Wave front ($\varepsilon = 0{\cdot}1$, $t = 0{\cdot}5$)

		Petrov–Galerkin		
Node	Exact	S.G.A.	P.A.	Compact differencing
1	1	1	1	1
2	1	1	1	1·030
3	1	1	1	0·990
4	1	1	1	0·973
5	1	0·998	0·999	1·009
6	0·998	0·991	0·997	1·004
7	0·980	0·970	0·982	0·986
8	0·847	0·862	0·850	0·696
9	0·452	0·461	0·444	0·360
10	0·238	0·159	0·171	0·228
11	0·207	0·300	0·286	0·203
12	0·2	0·194	0·197	0·2
13	0·2	0·213	0·211	0·2
14	0·2	0·211	0·210	0·2
15	0·2	0·188	0·190	0·2
16	0·2	0·201	0·207	0·2
17	0·2	0·191	0·193	0·2
18	0·2	0·203	0·202	0·2
19	0·2	0·2	0·2	0·2

expect an accurate nodal value at the half integer node 18 if the values are accurate at the integer nodes 17 and 19. (This can best be seen by considering the graph of the quadratic interpolant to the exact solution through these three nodes.) This is borne out by Table 4 for the P.A. There is little to say about Table 5 except that the Petrov–Galerkin methods are good ahead of the wave front and compact differencing at the rear. All methods would be improved, of course, by reducing the grid in the vicinity of steep gradients of the solution.

5. Korteweg de Vries and Related Equations

Our final section deals with another situation where P.A. results in an increase in the order of local accuracy of the scheme. Sanz-Serna & Christie (1979) solved the Korteweg de Vries equation

$$u_t + uu_x + \mu u_{xxx} = 0 \qquad (5.1)$$

using piecewise linear trial functions and cubic spline test functions. They proved that the resulting scheme, viewed as a continuous in time finite-difference replacement of (5.1), is fourth-order accurate with P.A. and second-order with S.G.A. Numerical experiments by these authors confirmed the improvement given by P.A.

According to Swartz & Wendroff (1974*a*) and as noted by Chin *et al.* (1979) for the linearized version of (5.1), the choice linear/cubic spline leads to the same system of ordinary differential equations as those obtained with quadratic spline/quadratic spline. Thus the fourth-order local accuracy was to be expected for the linearized equations in view of Thomee's superconvergence results (1973) for splines. As in the example considered in the introduction the high order of accuracy applies to the non-linear case if P.A. is used but not if we employ S.G.A. (Note that the term u_{xxx} cannot be approximated to fourth-order by a five-point difference scheme, so that again high accuracy is obtained through cancellations.)

Kuo Pen-Yu & Sanz-Serna (1981) have proved convergence of P.A. for the K.d.V. equation. Their analysis shows that here also it is possible to have numerical instability leading to blow-up in finite time. However, extensive numerical experimentation indicates that the scheme is remarkably stable so that we are led to believe that the blow-up, predicted by theory, takes place at a value of time so large as to be considered infinite for practical purposes.

Recently we have applied similar techniques to the equation

$$u_t + u^3 u_x + \mu u_{xxx} = 0$$

(see Jeffrey & Kakutani, 1972) so as to study the interaction of solitary waves, and once more we have found P.A. to perform satisfactorily.

References

Chin, R. C. Y., Hedstrom, G. W. & Karlsson, K. E. 1979 A simplified Galerkin method for hyperbolic equations. *Math. Comput.* **33**, 647–658.

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

CHRISTIE, I. & MITCHELL, A. R. 1978 Upwinding of high order Galerkin methods in conduction–convection problems. *Int. J. Num. Meth. Engng* **12,** 1764–1771.

CIARLET, P. G., SCHULTZ, M. H. & VARGA, R. S. 1967 Numerical methods of high order accuracy for nonlinear boundary value problems. I. One dimensional problem. *Num. Math.* **9,** 394–430.

CIARLET, P. G., SCHULTZ, M. H. & VARGA, R. S. 1969 Numerical methods of high order accuracy for nonlinear boundary value problems. V. Monotone operator theory. *Num. Math.* **13,** 51–77.

COLE, J. D. 1951 On a quasi linear parabolic equation occurring in aerodynamics. *Qt. App. Math.* **9,** 225–236.

CULLEN, M. J. P. 1974 A finite element method for a non-linear initial value problem. *J. Inst. Maths Applics* **13,** 233–248.

CULLEN, M. J. P. & MORTON, K. W. 1980 Analysis of evolutionary error in finite element and other methods. *J. Comput. Phys.* **34,** 245–267.

FLETCHER, C. A. J. 1979 A primitive variable finite element formulation for inviscid compressible flow. *J. Comput. Phys.* **33,** 301–312.

FORNBERG, B. 1973 On the instability of leapfrog and Crank–Nicolson approximations of a nonlinear partial differential equation. *Maths Comput.* **27,** 45–57.

HERBOLD, R. J., SCHULTZ, M. H. & VARGA, R. S. 1969 Quadrature schemes for the numerical solution of boundary value problems by variational techniques. *Aequ. Math.* **3,** 96–119.

HIRSH, R. S. 1975 Higher order accurate difference solutions of fluid mechanics problems by a compact differencing technique. *J. Comput. Phys.* **19,** 90–109.

HOPF, E. 1950 The partial differential equation $u_t + uu_x = \mu u_{xx}$. *Communs Pure Appl. Math.* **3,** 201–230.

JEFFREY, A. and KAKUTANI, T. 1972 Weak nonlinear dispersive waves: a discussion centred round the Korteweg de Vries equation. *SIAM Rev.* **14,** 582–642.

KUO PEN-YU & SANZ-SERNA, J. M. 1981 Convergence of methods for the numerical solution of the Korteweg de Vries equation. *IMA J. Num. Analysis* **1,** 215–222.

LUCCHI, C. W. 1980 Improvement of MacCormack's scheme for Burgers' equation using a finite element method. *Int. J. Num. Meth. Engng* **15,** 537–555.

MITCHELL, A. R. & GRIFFITHS, D. F. 1980 *The Finite Difference Method in Partial Differential Equations.* London: John Wiley and Sons.

MORTON, K. W. 1977 Initial-value problems by finite difference and other methods. In *The State of the Art in Numerical Analysis* (D. A. H. Jacobs, Ed.) Academic Press.

PERRIN, F. M., PRICE, H. S. & VARGA, R. S. 1969 On higher order methods for nonlinear two-point boundary value problems. *Num. Math.* **13,** 180–198.

SANZ-SERNA, J. M. & CHRISTIE, I. forthcoming Petrov–Galerkin methods for non-linear dispersive waves. *J. Comput. Phys.*

SWARTZ, B. K. & WENDROFF, B. 1969 Generalised finite difference schemes. *Maths Comput.* **23,** 37–49.

SWARTZ, B. K. & WENDROFF, B. 1974 The relative efficiency of finite difference and finite element methods. I. Hyperbolic problems and splines. *SIAM J. Num. Analysis* **11,** 979–993.

SWARTZ, B. K. & WENDROFF, B. 1974 The relation between the Galerkin and Collocation methods using smooth splines. *SIAM J. Num. Analysis* **11,** 994–996.

THOMEE, V. 1973 Convergence estimates for semi-discrete Galerkin methods for initial-value problems. *Springer Lecture Notes* 333. Berlin: Springer Verlag.

TURKEL, E. 1980 On the practical use of high-order methods for hyperbolic systems. *J. Comput. Phys.* **35,** 319–340.