A Simplified Galerkin Method for Hyperbolic Equations*

By R. C. Y. Chin, G. W. Hedstrom and K. E. Karlsson

Abstract. We modify a Galerkin method for nonlinear hyperbolic equations so that it becomes a simpler method of lines, which may be viewed as a collocation method. The high order of accuracy is preserved. We present a linear wave analysis of the scheme and discuss some aspects of nonlinear problems. Our numerical experiments indicate that the addition of a proper artificial viscosity makes the method competitive and the common difference schemes, even when the solution has discontinuities.

1. Introduction. In considering Galerkin methods for the problem

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} f(u, x, t) \quad (-\infty < x < \infty, \ t > 0),$$

$$u(x, 0) = u_0(x),$$

we must first decide how to implement such a method. We might replace \( u \) by a spline \( v \) and solve the system of equations obtained from

$$\int_{-\infty}^{\infty} v_t w \, dx = -\int_{-\infty}^{\infty} f(u, x, t) w_x \, dx \quad (1.2)$$

by letting \( w \) run through a basis of spline functions. This approach was taken by Swartz and Wendroff [17] in their theoretical discussion, but for computation in nonlinear cases they used some simplified methods. In this paper we examine in detail the properties of one of these simplified Galerkin methods. This method may also be viewed as a collocation method.

We first describe the method, starting with a definition of the \( B \)-splines [13]. Let \( \phi_0 \) be defined by

$$\phi_0(x) = \begin{cases} 1 & (|x| < 1/2), \\ 0 & (|x| \geq 1/2). \end{cases}$$

If \( p \) is a natural number, let \( \phi_p = \phi_0 * \phi_{p-1} \), where \( * \) denotes convolution. The number \( p \) may be either even or odd, but it must be positive. We approximate \( u \) by

---

Received September 22, 1975; revised November 23, 1977.


*The work of the first two authors was supported by the U. S. Energy Research and Development Administration under contract No. W-7405-Eng-48, and performed by the Lawrence Livermore Laboratory.
where \( h \) is the mesh size and the \( c_k(t) \) are yet to be determined. Let \( F \) be the spline
\[
F(x, t) = \sum_k d_k(t) \phi_p(x/h - k),
\]
where the \( d_k(t) \) are such that
\[
F(jh, t) = f(u(jh, t), jh, t) \quad (j = 0, \pm 1, \ldots).
\]
It is known [16] that \( F \) exists and is unique if
\[
\sum_j |f(u(jh, t), jh, t)|^2 < \infty.
\]
Let \( a_j \) and \( b_j \) be defined by
\[
a_j = \phi_p * \phi_p(j),
\]
(1.4)\[
b_j = \phi'_p * \phi_p(j) \quad (j = 0, \pm 1, \ldots);
\]
and let \( A \) and \( B \) be doubly infinite matrices with components \( A_{jk} = a_{j-k}, B_{jk} = b_{j-k} \).

One simplified Galerkin method for (1.1) is
\[
(1.5) \quad A c'(t) = (1/h)Bd(t),
\]
where \( c \) and \( d \) are the vectors with components, respectively, \( c_k \) and \( d_k \). A further simplification is possible and desirable. For purposes of programming, Eq. (1.5) has the disadvantage that if \( p > 1 \), then \( d \) depends on \( c \) through (1.3) in a very complicated way. This complication may be easily avoided, because the argument of Swartz and Wendroff [18] shows that Eq. (1.5) is equivalent to a method of lines
\[
(1.6) \quad \sum_k a_{j-k} v'_k(t) = \sum_k (1/h)b_{j-k} f(v_k(t), x_k, t),
\]
where \( x_k = kh \) and \( v_k(t) = u(x_k, t) \). As in [18], one may also view (1.6) as collocation at the points \( x_k \) based on the splines \( \phi_{2p+1} \).

It is known from the work of Thomée and Wendroff [21] that for a linear equation (1.1) with constant coefficients the accuracy of (1.5) is \( O(h^{2p+2}) \). Nevertheless, we sketch here a simpler, direct proof that the order of accuracy of (1.6) is \( O(h^{2p+2}) \).

In terms of the vector \( F \) with \( F_k = f(u_k(t), x_k, t) \) \((k = 0, \pm 1, \ldots)\), we may write
(1.6) in the form \( A \tilde{u}' = h^{-1}B \tilde{F}. \) Note that the time dependence is irrelevant here; we have simply a discrete approximation to \( \partial v/\partial x. \) Following the ideas of [15] and [12, p. 68], we introduce the discrete Fourier transform
\[
(1.7) \quad \tilde{v}(\theta) = \sum_k v_k e^{-ik\theta}
\]
to get
\[
(1.8) \quad \tilde{a} \partial \tilde{v}/\partial t = h^{-1} \tilde{b} \tilde{F}.
\]
It is known from [15] and [19] that

\( \tilde{a}(\theta) = \sum_k q(\theta - 2\pi k) \),

(1.9)

\( \tilde{b}(\theta) = i \sum_k (\theta - 2\pi k)q(\theta - 2\pi k) \),

(1.10)

where

\( q(\theta) = ((2/\theta) \sin(\theta/2))^{2p+2} \).

(1.11)

A direct computation (see [19, p. 985] or (2.6) below) shows that

\( \tilde{a}(\theta) + i\tilde{b}(\theta)/\theta = O(\theta^{2p+2}), \quad \theta \to 0. \)

(1.12)

This shows that the truncation error of (1.6) is \( O(h^{2p+2}) \). A standard argument as in [21], based on the energy method [12] and the positive definiteness of \( A \), shows that for \( f(u, x, t) = \lambda(x, t)u \) with \( \lambda \) bounded and for solutions \( u \) of (1.1) with \( \lambda u \) in the Sobolev space \( H_{2p+3} \), the order of accuracy is \( O(h^{2p+3}) \).

For \( p = 1 \) (linear splines) both Eqs. (1.5) and (1.6) take the form

\( (u'_j + 4u'_j + u'_{j+1})/6 = (f_{j+1} - f_{j-1})/(2h) \),

(1.13)

where \( f_j = f(u_j, x_j, t) \). This equation could have been obtained by integrating (1.1) from \( x_{j-1} \) to \( x_{j+1} \) and using Simpson's rule. Hence, it is fourth-order accurate. For \( p = 2 \) (quadratic splines) Eq. (1.6) becomes

\[
(\begin{array}{c}
u'_{j-2} + 26v'_{j-1} + 66v'_j + 26v'_{j+1} + v'_{j+2} \\
= (f_{j+2} + 10f'_{j+1} - 10f_{j-1} - f'_{j-2})/(24h).\end{array})
\]

(1.14)

Similarly, for \( p = 3 \) (cubic splines) Eq. (1.7) becomes

\[
(\begin{array}{c}
u'_{j-3} + 120v'_{j-2} + 1191v'_{j-1} + 2416v'_j + 1191v'_{j+1} + 120v'_{j+2} + v'_{j+3} \\
= (f_{j+3} + 56f'_{j+2} + 245f'_{j+1} - 245f_{j-1} - 56f'_{j-2} - f'_{j-3})/(720h).\end{array})
\]

(1.15)

Note that we use the splines to generate the coefficients in a method of lines. After that the splines are forgotten. The reader who regards (1.6) as a collocation method will undoubtedly view this matter differently. We remark that Collatz [6, p. 528] presents (1.13), and he also gives difference schemes with the same bandwidths as (1.14) and (1.15) but with even higher orders of accuracy. The efficiency of all these schemes was compared in [19].

In the next section we do a linear wave analysis of (1.6) in the case when \( f(u, x, t) = -v \). In particular, we determine the group velocities of propagating waves, and we determine the behavior of the solution near a wave front. It is known from the theory of dispersive waves [3] that the group velocity (velocity of a wave packet) is the velocity which is observed. In the special cases \( p = 1 \) or \( 2 \) the group velocities are given in [22] and [23]. Several authors have investigated phase velocities, most notably, Swartz and Wendroff [19]. In Section 3 we prove the instability of (1.6) when \( f(u, x, t) = -u^2/2 \), and we show some computations which indicate the stabilizing effect of adding an artificial viscosity.
2. Linear Wave Analysis. The system (1.6) is ordinarily solved by some finite-difference method in time. In this section we discuss the stability of the resulting time-difference scheme when \( f(u, x, t) = -u \), and of more interest, we determine the group velocities of propagating waves. We then describe the behavior of the solution of (1.6) when the initial data has a step discontinuity.

For the equation \( u_t + u_x = 0 \) the method of lines (1.6) becomes
\[
Av' = -Bu/h, \tag{2.1}
\]
and by (1.8) the discrete Fourier transform of (2.1) is
\[
\tilde{a} \tilde{v}/\partial t = -\tilde{b} \tilde{v}/h. \tag{2.2}
\]
The amplification function or symbol \( G \) for (2.1) is, therefore, \( G = \exp\{-i\psi_p/\tau\} \), where \( \psi_p = -ib/\alpha_p \). Note that it follows from (1.9)–(1.11) that \( \psi_p(\theta) \) is a real-valued, odd function of \( \theta \). In the following theorem we determine some of the properties of the phase function \( \psi_p \) and show their relevance to numerical computation.

**Theorem 2.1.** The eigenvalues \( \lambda \) of \(-A^{-1}B\) coincide with the interval on the imaginary axis of values taken by the function
\[
\lambda = -i\psi_p(\theta) \quad (-\pi < \theta < \pi). \tag{2.3}
\]
If the component \( v_k(t) \) in (2.1) is considered as an approximation to \( u(kh, t) \), then the solution of Eq. (2.1) has propagating waves with group velocities given by the values of \( \psi'_p(\theta) \) for \( -\pi < \theta < \pi \).

**Remarks.** (i) The matrices \( A \) and \( B \) are infinite, but if instead of the Cauchy problem, we impose the boundary condition that the solution (and the splines) be periodic in \( x \), then \( A \) and \( B \) become finite cyclic matrices. In this case the eigenvalues are \( \lambda_k = -i\psi_p(2\pi k/n) \) \((k = 1, 2, \ldots, n)\).

(ii) The significance of the eigenvalues is that if we solve (2.1) by a difference scheme, the values of \( \lambda \Delta t/h \) must lie in the stability region of the difference scheme. See the books by Gear [8] and Lapidus and Seinfeld [11] for stability regions of common difference schemes. For example, if \( u' \) is replaced by a central difference \((u(t + \Delta t) - u(t - \Delta t))/(2\Delta t)\), the requirement for stability is \( \Delta t/h \leq 1/\max |\psi_p| \). If \( p = 1 \), we may use elementary calculus to show that \( \max |\psi_1| = \sqrt{3} \). For \( p = 2 \) or 3 we used Newton’s method to obtain
\[
\max |\psi_2| \approx 2.0781, \quad \max |\psi_3| \approx 2.2813.
\]
It follows from (1.9), (1.10), and (1.11) that
\[
\lim_{p \to \infty} \psi_p(\theta) = \theta \quad (|\theta| < \pi),
\]
so that because of the periodicity of \( \psi_p \) we have
\[
\lim_{p \to \infty} \max |\psi_p(\theta)| = \pi.
\]
We conjecture that this convergence is monotone increasing. The graphs of \( \psi_p \) for \( p = 1, 2, \) and \( 3 \) are shown in Figure 1.
(iii) We shall say more about the group velocities shortly. The graphs of $\psi'_p$ for $p = 1, 2, 3$ are shown in Figure 2. It seems curious that for these values of $p$ the range of $\psi'_p$ is $-2p - 1 < \psi'_p(\theta) < 1$.

**Proof of Theorem 2.1.** Suppose that $\lambda$ is an eigenvalue of $-A^{-1}B$ and that $w$ is the corresponding eigenvector, so that $Bw = -\lambda Aw$. Upon taking discrete Fourier transforms, we obtain $\tilde{b}w = -\lambda \tilde{a}w$, from which it follows that $\lambda$ has one of the values $\lambda = -\tilde{b}(\theta_0)/\tilde{a}(\theta_0) = -i\psi_p(\theta_0)$.

We now show that all values $\lambda$ in (2.3) are eigenvalues. For this purpose we take $w$ to be the vector with components $w_k = e^{-iks}$ for some fixed $s$. Then each component of $Bw + \lambda Aw$ is $\tilde{b}(\theta) + \lambda \tilde{a}(\theta)$, so that we have (2.3). As expected in the case of continuous spectra, our eigenvector is not in $l^2$ and is, therefore, an eigenvector only in a generalized sense.

In order to study the group velocities of propagating waves, we solve (2.2) explicitly,
\[ \widetilde{v}(\theta, t) = \exp\{-i(t/h)\psi_p(\theta)\} \tilde{v}(\theta, 0). \]

According to the definition of the discrete Fourier transform (1.7), \( \tilde{v} \) is the sum of a Fourier series. Hence, the Fourier coefficient \( v_k \) is given by

\[ v_k(t) = 1/(2\pi) \int_{-\pi}^{\pi} \tilde{v}(\theta, 0) \exp\{-i(t/h)\psi_p(\theta) + ik\theta\} \, d\theta. \]  

With the notation \( x_k = kh \) Eq. (2.4) may be written

\[ v_k(t) = 1/(2\pi) \int_{-\pi}^{\pi} \tilde{v}(\theta, 0) \exp\{(-t/h)(\psi_p(\theta) - x_k\theta)\} \, d\theta. \]

This integral has saddle points (also points of stationary phase) for real values of \( \theta \) such that

\[ t\psi_p'(\theta) - x_k = 0, \]

and the behavior of the integral as \( h \to 0 \) may be estimated by methods used in the theory of wave propagation [3]. It is clear from (2.5) that the speed of propagation is \( \psi_p(\theta) \). The amplitude depends on \( \tilde{v}(\theta, 0) \). This proves the theorem.

We close this section with a description of the behavior of \( v_k(t) \) as \( h/t \to 0 \) when the initial value \( v_k(0) \) is a step function \( v_k(0) = 2^{-1} \text{sgn}(k) \). It follows from [3] and [14] that wiggles move with the group velocities \( \psi_p'(\theta) \). In particular, if \( \psi_p''(\theta_0) \neq 0 \), then the wave packet with speed \( \psi_p(\theta_0) \) has amplitude

\[ O(h^{1/2}e^{-1/2} |\psi_p''(\theta_0)|^{-1/2}), \quad h/t \to 0. \]

The behavior of a wave packet corresponding to a multiple zero of \( \psi_p' \) is more complicated. The wave front itself is one such packet (\( \theta_0 = 0 \)), for we know that [20] as \( \theta \to 0 \) we have

\[ \psi_p(\theta) = \theta - c\theta^{2p+3} + O(|\theta|^{2p+5}) \]

for some \( c \neq 0 \).

The behavior of \( v_k(t) \) near the wave front depends on the value of \( c \), so we proceed to determine \( c \) using (1.9)--(1.11),

\[ \psi_p(\theta) = -i\hat{b}(\theta)/\tilde{a}(\theta), \quad \psi_p(\theta) = \theta - 2\pi \sum kq(\theta - 2nk)/\tilde{a}(\theta). \]

For \( k \neq 0 \) and \( \theta \to 0 \) we have

\[ q(\theta - 2nk) = \left(\theta/(2\pi k)\right)^{2p+2}(1 + (p + 1)\theta/(\pi k) + O(\theta^2)) \]

uniformly in \( k \). Thus, it follows that as \( \theta \to 0 \) we have

\[ 2\pi \sum kq(\theta - 2nk) = 4(p + 1)\theta^{2p+3} \sum_{k=1}^{\infty} (2\pi k)^{-2p-2} + O(\theta^{2p+4}) \]

and

\[ \tilde{a}(\theta) = \sum q(\theta - 2\pi k) = 1 + O(\theta^2). \]
Hence, we see that (2.6) holds with
\[ c = 4(p + 1) \sum_{k=1}^{\infty} (2\pi k)^{-2p-2} = \frac{|B_{2p+2}|}{(2p + 1)!}, \]
where \( B_{2p+2} \) is the Bernoulli number [4, p. 298]. We note in passing that we have also proven (1.12).

Now that the value of \( c \) in (2.6) is known, the results of [5] show that near the wave front \( v_k(t) \) is asymptotic to the integral of a generalized Airy function,
\[
(2.7) \quad v_k(t) \sim A_{2p + 3, 2}(0, y), \quad y = \omega \epsilon^{-1/((2p + 3)(t/h)(2p + 2)/(2p + 3)).}
\]
Here, \( \omega \) is the relative distance from the wave front, \( \omega = (x - t)/t \), and the estimate (2.7) is valid so long as
\[
|\omega| (t/h)(2p+2)/(2p+5) \to 0.
\]
Graphs and tables of \( A_{5, 2}(0, y) \) and \( A_{7, 2}(0, y) \) may be found in [5]. There are oscillations both ahead of and behind the jump. Note that if we were to add an artificial viscosity of order \( r \) to (1.6), the oscillations away from the wave front would be exponentially damped, and the behavior near the wave front would be
\[
v_k(t) \sim A_{2p + 3, r}(\alpha, y)
\]
for some positive \( \alpha \) depending on the amount of artificial viscosity.

3. A Nonlinear Equation. We begin this section with an example showing the instability of the simplified Galerkin scheme (1.6) for Burgers’ equation. This example is an adaptation of Fornberg’s example [7] showing the instability of the leapfrog scheme. Note that the true Galerkin method (1.2) is stable for this equation [17].

**Theorem 3.1.** Let the method of lines (1.6) be used for the problem
\[
(3.1) \quad u_t + (u^2/2)_x = 0, \quad u(x, 0) = u_0(x).
\]
Define constants \( \alpha \) and \( \beta \) by
\[
\alpha = \sum_{j=0}^{\infty} (a_{3j} - a_{3j+1}), \quad \beta = \sum_{j=0}^{\infty} (b_{-3j-1} - b_{-3j-2}).
\]
Suppose that \( \beta \neq 0 \), as is true if \( p = 1, 2, \) or \( 3 \). Let \( \epsilon \) be any positive number. Then for every fixed \( h \) there exists a periodic function \( u_0 \) with \( \max |u_0| = \epsilon \) such that the solution \( u_k(t) \) of (1.7) becomes infinite within the time \( 2\alpha h/(|\beta| \epsilon) \).

**Remark.** This behavior is disturbing because it is known [9] that every solution of (3.1) with \( u_0 \) periodic and locally of bounded variation decays as \( O(1/t) \) as \( t \to \infty \).

**Proof of the Theorem.** It is clear from the definition (1.4) of \( a_k \) and from the bell-shaped graphs of the \( B \)-splines [13] that \( a_{k+1} < a_k \) (\( k > 0 \)) and \( a_1 < a_0 \). Hence, \( \alpha > 0 \). It is also clear from (1.13), (1.14), and (1.15) that \( \beta > 0 \) if \( p = 1, 2, \) or \( 3 \). For larger values of \( p \), though, the behavior of \( \phi_p \) as \( p \to \infty \) [1] shows that \( \beta \) is no longer a sum of nonnegative and positive terms. We do not know whether \( \beta \) is ever zero.
If $\beta > 0$, we take $u_0$ such that
\[
\begin{align*}
    u_0(x_k) &= 0, & k &\equiv 0 \pmod{3}, \\
    u_0(x_k) &= -\varepsilon, & k &\equiv 1 \pmod{3}, \\
    u_0(x_k) &= \varepsilon, & k &\equiv 2 \pmod{3}.
\end{align*}
\]
Because $a_{-k} = a_k$ and $b_{-k} = -b_k$, the solution of (1.6) may be written
\[
\begin{align*}
    v_k(t) &= 0, & k &\equiv 0 \pmod{3}, \\
    v_k(t) &= -y(t), & k &\equiv 1 \pmod{3}, \\
    v_k(t) &= y(t), & k &\equiv 2 \pmod{3},
\end{align*}
\]
where $y$ is the solution to the initial-value problem,
\[
(3.2) \quad \alpha y' = \beta y^2/(2h), \quad y(0) = \varepsilon.
\]
The solution of (3.2) is
\[
y(t) = 2e^\alpha t/(2\alpha h - \beta t),
\]
thus proving the theorem for $\beta > 0$.

If $\beta < 0$, we replace $-u_0$ in the above argument by $-u_0$.

Finally, we close the paper with some graphs illustrating the effect of an artificial viscosity for the problem (3.1) with $p = 1$. In place of the method of lines (1.13) we used
\[
(3.3) \quad (v_{k+1}^p + 4v_k^p + v_{k-1}^p)/6 = (-v_{k+1}^2 + v_{k-1}^2)/(4h)
\]
\[
+ (\varepsilon/h)(v_{k+1} - v_k)(1 + \alpha(v_{k+1} - v_k)^2)
\]
\[
- (\varepsilon/h)(v_k - v_{k-1})(1 + \alpha(v_k - v_{k-1})^2).
\]
In effect, Eq. (3.3) is an approximation to
\[
(3.4) \quad u_t + (u^2/2)_x = \varepsilon h \partial/\partial x \{1 + \alpha h^2 u_x^2\} u_x.
\]
We used initial data
\[
\begin{align*}
    u_0(x) &= 0 \quad (|x| > 1/2), \\
    u_0(x) &= 1/2 \quad (|x| = 1/2), \\
    u_0(x) &= 1 \quad (|x| < 1/2)
\end{align*}
\]
in all our runs. We used boundary conditions $u(-1, t) = u(1, t) = 0$ and stopped the computation before the shock hit the right-hand boundary. The grid size used was $h = 1/200$. The system (3.3) was integrated using Hindmarsh’s version [10] of the Gear ordinary differential equation solver. The stiff version was used, but since the eigenvalues lie near the imaginary axis, we restricted the order to at most four. The graphs of the solutions are shown at time $t = .25$. 

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use
Figure 3. Solution to (3.3) with $\alpha = 0$ and $\epsilon = .01$

Figure 4. Solution to (3.3) with $\alpha = 0$ and $\epsilon = .1$
Figure 5. Solution to (3.3) with \( \alpha = 0 \) and \( \epsilon = 0.25 \)

Figure 6. Solution to (3.3) with \( \alpha = 59.26 \) and \( \epsilon = 0.01 \)
Figure 3 shows the case when \( \alpha = 0 \) and \( \epsilon = .01 \). The wiggles behind the shock are very clear, and it is also clear that some of them move to the left. Figure 4 has \( \alpha = 0 \) and \( \epsilon = .1 \) and shows damping of the wiggles. Figure 5 has \( \alpha = 0 \) and \( \epsilon = .25 \). In this case the solution mimics the solution of (3.4) (see [2]). Finally, Figure 6 illustrates the effect of \( \alpha \); we used \( \alpha = 59.26 \) and \( \epsilon = 0.01 \). Note that the wiggles in Figure 6 are much smaller than those in Figure 3, where the value of \( \epsilon \) is the same.


