SIAM J. SCI. STAT. COMPUT.
Vol. 7, No. 3, July 1986

020

ON THE SCOPE OF THE METHOD OF MODIFIED EQUATIONS*

D. F. GRIFFITHS† AND J. M. SANZ-SERNA‡

Abstract. A rigorous analysis is presented for the method of modified equations whereby its range of applicability and its shortcomings are delineated. Numerous examples from different areas are presented and the theoretical findings are confirmed throughout by computational experiments.

Key words. modified equations, difference approximations, local truncation errors, stability

AMS (MOS) subject classifications. 65L05, 65L10, 65M05, 65M10

1. Introduction. Modified equations have been a commonly used tool in the study of difference schemes. Because of the lack of any theoretical foundation, this use has been accompanied by constant difficulties and results derived from modified equations have sometimes been regarded with apprehension. As a result a situation has arisen where authors either disregard entirely the technique or have an unjustified faith in its scope. The aim of the present paper is to investigate carefully the foundation and applicability of the method in the hope of clarifying the situation.

To our best knowledge the method of modified equations was first used by Garabedian [4] in the analysis of SOR iterations. Few papers have been devoted to studying the method (Hirt [11], Warming and Hyett [34], Wilders [35], also Morton [16]). On the other hand the technique has been extensively employed in the literature, see e.g. [1], [6], [7], [8], [10], [14], [19], [20], [26], [33] and [36]. By and large, applications have concentrated on the investigation of dispersive and dissipative properties of partial difference schemes. A nonstandard example is given by Duncan and Griffiths [3]. One of the referees has rightly pointed out the analogy between the idea of modified equation and the backward error analysis of Wilkinson.

A summary of the paper is as follows. The main ideas are introduced in § 2, in the context of a concrete example. This is followed in § 3 by the discussion of a wide range of applications to both ordinary and partial differential equations. The theoretical analyses are backed throughout by numerical illustrations. We place the method in a wider context in § 4, by making comparisons with other forms of analysis. Our findings are summarized in § 5.

In keeping with the aim of the paper, the examples included, mostly simple, have been chosen to provide insight into the various aspects of the method; the presentation of new real-life applications is completely outside the scope of the article.

2. Modified equations. This section introduces, in a rigorous way, the concept of modified equation. For simplicity, the ideas are presented in the case of a model problem which exhibits all the important features of the more general situation. In fact, it is not difficult to rewrite the material below in the language of any of the general discretization theories (e.g. [31], [29], [33], [23]) and in particular, [37 § 2.4]).

We consider the scalar initial value problem

(2.1a) $$u(0) = \eta,$$

(2.1b) $$\frac{du}{dt} = f(u), \qquad 0 \leqq t \leqq T,$$

* Received by the editors September 18, 1984, and in revised form, February 20, 1985.

† Department of Mathematical Sciences, The University, Dundee DD1 4HN, Scotland.

‡ Departamento de Ecuaciones Funcionales, Facultad de Ciencias, Universidad de Valladolid, Valladolid, Spain.

where $f(u)$ is smooth and Lipschitz continuous in $-\infty < u < \infty$, with Lipschitz constant L. These hypotheses ensure the existence and the uniqueness of a smooth solution. The problem (2.1) is discretized by means of Euler's rule

(2.2a) $$U_0 = \eta + \delta,$$

(2.2b) $$(U_{n+1} - U_n)/h = f(U_n), \qquad n = 0, 1, \cdots, N-1.$$

Here N is a positive integer, $h = T/N$ and δ caters for a possible error in the starting value. For simplicity, the effects of round-off errors are not considered in this paper.

Some of the basic, elementary steps of the analysis of (2.2) ([12], [9], [5]) will now be presented for later reference. A crucial part of the analysis is the estimation of the size of the global errors

(2.3) $$e_n = Y_n - U_n,$$

where $Y_n = u(t_n)$ is the value of the theoretical solution at the grid-point $t_n = nh$. In more concrete terms we are interested in the quantity

(2.4) $$e = \max\{|e_n|: n = 0, 1, \cdots, N\}.$$

Note that U_n, Y_n, e_n, e, δ depend on the parameter h but this dependence does not appear in the notation. The standard approach to the study of e is the following indirect one (and this includes both the derivation of bounds for e for a given, fixed h and the investigation of the behaviour of e as h tends to zero).

First the auxiliary *local truncation errors*

(2.5a) $$l_0 = Y_0 - (\eta + \delta),$$

(2.5b) $$l_{n+1} = (Y_{n+1} - Y_n)/h - f(Y_n), \qquad n = 0, 1, \cdots, N-1$$

are introduced. A simple Taylor expansion taking into account that $Y_n = u(t_n)$ reveals that, for $n > 0$, l_n can be bounded by $\frac{1}{2}hB_2$, where B_2 is a bound for $|u''(t)|$, $0 \leqq t \leqq T$. Thus

(2.6) $$l = \max\{|l_n|: n = 0, 1, \cdots, N\}$$

is $O(h+\delta)$ as $h \to 0$.

Then, the *stability* of the discretization is established, i.e. it is shown that

(2.7) $$e \leqq Cl,$$

where C is a postive constant which depends on T and L but not on h. This bound is derived by subtracting (2.5) from (2.2) and applying induction w.r.t. n. No property of Y_n is required in the derivation, i.e. the fact that $Y_n = u(t_n)$ is not used at this stage. From (2.7) e is also $O(h+\delta)$ and one says that (2.2) possesses first order rate of convergence.

Remark. Some authors [13] prefer to write (2.2b) in the undivided form

$$U_{n+1} - U_n = hf(U_n).$$

Accordingly they define the local truncation error for $n > 0$, to be

$$Y_{n+1} - Y_n - hf(Y_n)$$

rather than (2.5b). With this definition the local errors are $O(h^2+\delta)$ while the global errors, whose definition remains unchanged, are $O(h+\delta)$. In this paper, finite difference schemes are always written in divided form, i.e. in the form resulting from the replacement of derivatives by divided differences.

The introduction of modified equations aims at describing the behaviour of the numerical solution U_n. This will now be illustrated in the context of (2.1), (2.2). We consider the *modified* problem

$$z(0) = \eta + \delta, \tag{2.8a}$$

$$(1+\tfrac{1}{2}hf'(z))z' = f(z), \qquad 0 \leqq t \leqq T. \tag{2.8b}$$

(The derivation of modified problems is considered in the next section. No motivation for (2.8) will be provided at this stage.) The standard theory of continuous dependence on the parameters shows that, at least for h small, (2.8) has a unique solution $z(t)$. (Notice again that $z(t)$ depends on h.) We claim that $z(t_n)$ is a better approximation to the numerical solution U_n than $u(t_n)$. Our task is to bound the quantity e given by (2.4)-(2.3), where now $Y_n = z(t_n)$. In order to do so, we resort to the indirect approach above. We still define l_n, l by (2.5)-(2.6) with $Y_n = z(t_n)$ and observe that (2.7) is still valid, since, as noted before, the derivation of the stability bound does not use any information on Y_n. However, now

$$l_0 = Y_0 - U_0 = z(0) - U_0 = 0$$

while, for $n = 0, 1, \cdots, N-1$,

$$l_{n+1} = (Y_{n+1} - Y_n)/h - f(Y_n) = [z(t_{n+1}) - z(t_n)]/h - f(z_n)$$
$$= z'(t_n) + (h/2)z''(t_n) + (h^2/6)z'''(\theta_n) - f(z_n),$$

where $t_n < \theta_n < t_{n+1}$. On using (2.8b)

$$l_{n+1} = z'(t_n) + (h/2)z''(t_n) + (h^2/6)z'''(\theta_n) - z'(t_n) - (h/2)f'(z(t_n))z'(t_n)$$
$$= (h/2)[z''(t_n) - f'(z(t_n))z'(t_n)] + (h^2/6)z'''(\theta_n)$$

which, on using the equation obtained by differentiation of (2.8b), leads to

$$l_{n+1} = (h/2)^2[f'(z(t_n))z''(t_n) + f''(z(t_n))(z'(t_n))^2] + (h^2/6)z'''(\theta_n). \tag{2.9}$$

Now z, z', z'', z''' can be bounded independently of h because of the continuous dependence of the solutions of (2.8) on the parameter h. We conclude that now $e = O(h^2)$ and say that the modified problem (2.8) describes the behaviour of the solution of (2.2) with second order of correctness. This will be now illustrated by means of an example.

The problem (2.8) is easily integrated to yield

$$\int_{\eta+\delta}^{z(t)} \frac{dv}{f(v)} + \frac{h}{2} \ln \frac{|f(z(t))|}{|f(\eta+\delta)|} = t. \tag{2.10}$$

In what follows we set $f(u) = u^2$. This does not strictly satisfy the hypotheses above in that $f(u)$ is Lipschitz continuous for $-M < u < M$, M finite but not for $-\infty < u < \infty$, however this poses no difficulty (see e.g. [25, p. 24]). We further set $T = .99$, $\eta = 1$, $\delta = 0$ with theoretical solution $u(t) = 1/(1-t)$. From (2.10), the modified solution $z(t)$ is given by

$$1 - \frac{1}{z} + h \ln z = t.$$

Figure 1 depicts $z(t)$, $u(t)$ and the Euler points U_n when $h = T/4$, $T/16$. It is clear that the values computed by the difference scheme are much closer to the values $z(t_n)$ than to the values $u(t_n)$. Moreover the agreement between $z(t_n)$ and U_n is very good, even for the coarser grid.

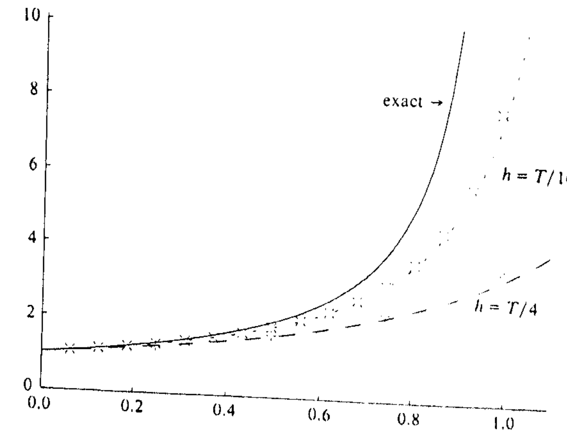

FIG. 1. *Exact solution of $u' = u^2$, $0 \leqq t \leqq T$ (= 0.99) (full line), together with solution of modified equation (broken lines) and numerical solution by Euler's method (+ and ×) for $h = T/4$ and $h = T/16$.*

It should be noted that the modified equation continues to describe the behaviour of the Euler solution even for $nh \geqq 1$, when the theoretical solution $u(t)$ ceases to exist (cf. [24]). This is illustrated in Fig. 2. One can actually derive bounds for $U_n - z(t_n)$, $nh \geqq 1$, but this point will not be pursued further. The following points summarize the main ideas and are useful in preventing the pitfalls which may arise from an indiscriminate application of modified problems.

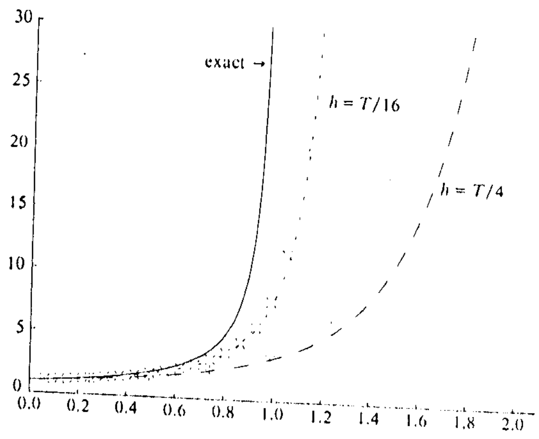

FIG. 2. *Solutions as in Fig. 1 but for an extended time interval.*

(i) A modified problem correct of order p is a problem depending on the parameter h with the property that its solution z has a local discretization error $O(h^p)$; i.e. it satisfies, except for $O(h^p)$ terms, the discrete equations defining the numerical method. In order to prove that the local discretization error l is $O(h^p)$, it is not enough to show that $l \leqq h^p B_p$, where B_p depends on the derivatives of z. In fact $z = z(h)$

also check that B_p remains bounded as $h\to 0$. This point was illustrated in the argument which follows (2.9).

It is perhaps worth pointing out that for a numerical method with qth order of convergence, the original problem being solved provides already a modified problem correct of order q.

(ii) It should be observed that even though it is customary in the literature to talk about modified *equations*, it is essential to consider modified *problems*, i.e. the modified equation should be supplemented by the necessary initial/boundary conditions (such as (2.8a) and care should be exercised in checking that the modified solution satisfies, except for $O(h^p)$ terms, the initial/boundary discrete equations (such as (2.2a)) which supplement the main scheme (such as (2.2b)).

(iii) The stability of the numerical method is an essential ingredient in guaranteeing the success of the method of modified problems. Without stability the bounds for local errors cannot be transferred to the global error $z-U$. The concept of stability used here refers to the $h\to 0$, nh fixed case (0-stability in ODEs [13], Lax stability in PDEs [19], [17], [18]) and not to the $n\to\infty$, fixed h case (weak stability in [13], contractivity [2]).

The importance of the points (i)-(iii) above will be borne out by the examples in the next section.

The idea of comparing the numerical solution U with a function close to but different from the theoretical solution u goes back to Strang [30]. See also [27], [22] and [37, Chap. 1].

3. The construction of modified problems: examples and counterexamples. In this section examples of modified problems are constructed, which illustrate the range of applicability of the technique.

(A) In our first example we return to (2.1)-(2.2). In order to construct a modified problem, correct of order two, the values of a smooth function $w(t)$ are substituted in (2.5):

$$l_0 = w(0)-(\eta+\delta),$$
$$l_{n+1} = (w(t_{n+1})-w(t_n))/h - f(w(t_n)), \qquad n=0,1,\cdots,N-1.$$

The possible dependence of $w(t)$ on h is not reflected in our notation. On Taylor expanding, we obtain

$$l_{n+1} = w'(t_n)+\frac{h}{2}w''(t_n)+\frac{h^2}{6}w'''(t_n)+\cdots-f(w(t_n))$$

and the requirement that $l=O(h^2)$ implies that $w(t)$ should satisfy

$$w(0)=\eta+\delta+O(h^2),$$
$$w'(t)+\frac{h}{2}w''(t)=f(w(t))+O(h^2).$$

In particular, the equations

$$w(0)=\eta+\delta, \tag{3.1a}$$

$$w'+\frac{h}{2}w''=f(w) \tag{3.1b}$$

appear to be good candidates for the role of modified problem with second order of correctness. However two difficulties have to be addressed. First the missing initial

value $w'(0)$ needs to be specified. Secondly, as $h\to 0$ the equation (3.1a) is singularly perturbed and there is a danger of w'' increasing without bound. Such a growth would destroy the $O(h^2)$ bound on l, as noted in § 2(i). The success of the modified problem approach depends on extracting a regularly perturbed problem from (3.1). A means of achieving this is by a suitable choice of $w'(0)$ to accompany (3.1). The difficulties inherent in this approach do not manifest themselves in this example and the study of this technique is deferred until the next example.

A second means of regularizing (3.1) is now presented.

Differentiation of (3.1b) leads to

$$w''+\frac{h}{2}w'''=f'(w)w'.$$

Upon eliminating w'' between this equation and (3.1b), we obtain

$$\left(1+\frac{h}{2}f'(w)\right)w'-\frac{h^2}{4}w'''=f(w).$$

The solutions of this equation we are interested in, namely those whose derivatives remain bounded as $h\to 0$, differ by $O(h^2)$ from those of

$$\left(1+\frac{h}{2}f'(z)\right)z'=f(z), \tag{3.2}$$

an equation which is not singularly perturbed. It was rigorously shown in § 2 that Euler's method provides a second order approximation to (3.2).

In practice, and for a more general problem, the steps leading up to a modified problem need not be performed rigorously. One would begin by replacing the grid values in the discrete equations by those of a smooth function w. Then, on performing a Taylor expansion and discarding powers of h higher than the pth, one would arrive at an equation involving high derivatives of w. Finally, and as far as possible, higher derivatives would be eliminated by combining this equation with those resulting from its differentiation (while systematically deleting terms which involve powers of h above the pth).

Once a candidate for a modified problem has been obtained by mere formal manipulation, the local error of its solution z should be rigorously shown to be small in order to conclude that z models the behaviour of the numerical solution provided by a stable scheme (cf. (i)-(iii) § 2).

An instance is provided by the equation

$$z'=\left(1-\frac{h}{2}f'(z)\right)f(z), \tag{3.3}$$

which results from formal inversion up to $O(h^2)$ of the factor $1+(h/2)f'(z)$ in (3.2). One easily shows that solutions of (3.3) with $z(0)=\eta+\delta$ possess a local error $l\le Ch^2$ (C independent of h), thereby providing a new modified problem for (2.2). This demonstrates that modified problems correct of order p are, by no means, unique.

(B) We retain the initial value problem (2.1), but this time discretize it by means of the backward Euler rule

$$U_0=\eta+\delta,$$
$$(U_{n+1}-U_n)/h=f(U_{n+1}), \qquad n=0,1,\cdots,N-1.$$

On proceeding as at the beginning of the previous example we arrive at the following analogue of (3.1)

$$w(0)=\eta+\delta, \tag{3.4a}$$

$$w'-\frac{h}{2}w''=f(w). \tag{3.4b}$$

We now discuss the regularization of (3.2) by means of a suitable choice of $w'(0)$. To avoid any unwelcome detail, we only consider the case $\eta=1$, $\delta=0$, $f(u)=\lambda u$. The family of solutions of (3.4a)-(3.4b) is given by

$$w(t)=(1+\alpha)\,e^{r_+t}-\alpha\,e^{r_-t},$$
$$r_\pm=(-1/h)[\pm\sqrt{1-2\lambda h}-1],$$

so that $r_+=\lambda+O(h)$, $r_-=2/h+O(1)$ and the derivatives of w will increase as $h\to 0$ unless the missing starting value $w'(0)$ is chosen to guarantee that $\alpha=0$, i.e. $w'(0)=r_+$. When $w'(0)\neq r_+$, solutions of (3.4) do not describe up to $O(h^2)$ the behaviour of the numerical solution, even though (3.4) was obtained by insisting that the expansion of the local error should only contain terms involving factors h^s, $s\geqq 2$ (cf. (i) of § 2).

This is illustrated numerically in Table 1, where $\lambda=1$, $t=\frac{1}{2}$ and $w'(0)=\lambda$ (a reasonable choice, since this coincides with $u'(0)$) and $w'(0)=r_+$. The theoretical solution has $u(\frac{1}{2})\simeq 1.649$.

TABLE 1

h	Numerical	Modified $w'(0)=\lambda$	Modified $w'(0)=r_+$
$\frac{1}{4}$	1.778	.93	1.796
$\frac{1}{8}$	1.706	−7.32	1.709
$\frac{1}{16}$	1.676	−5,907.49	1.676

Had Euler's rule been used, the roots r_+, r_- would have satisfied $r_+=\lambda+O(h)$, $r_-=-2/h+O(1)$ and then the study of the size of the derivatives of $\exp(r_-t)$ would have been rather delicate due to a boundary layer at $t=0$.

(C) This and the following example show the importance of considering modified *problems* rather than modified *equations*, i.e. proper account must be taken of all side conditions (§ 2(ii)).

We again consider the problem (2.1), but this time discretized by the leap-frog scheme

$$U_0=\eta, \tag{3.5a}$$

$$(U_1-U_0)/h=f(U_0), \tag{3.5b}$$

$$(U_{n+2}-U_n)/2h=f(U_{n+1}),\qquad n=0,1,\cdots,N-2, \tag{3.5c}$$

where the additional starting value U_1 is obtained by Euler's method. The scheme (3.5) possesses second order of convergence and therefore the original problem (2.1) provides a modified problem with second order of correctness. We now seek a modified problem of third order of correctness. On proceeding as in the derivation of (3.2), we

obtain the equation

$$\left[1+\frac{h^2}{6}(f''(z)f(z)+f'(z)^2)\right]z'=f(z) \tag{3.6}$$

whose solutions satisfy (3.5c) except for an $O(h^3)$ local discretization error. In this sense, (3.6) is a modified *equation* correct of *third* order for the leap-frog scheme. However solutions of (3.6) with $z(0)=\eta$ only satisfy (3.5b) with *second* order of correctness. Therefore a modified problem based on (3.6) cannot attain third order of correctness. A numerical example with $f(u)=u$, $u(0)=z(0)=1$, $t=1$ is presented in Table 2, which shows that the approximation provided by z has only second order of accuracy. In fact, no smooth function w of t and the parameter h can satisfy $w(t_n)-U_n=O(h^3)$, since the theory of asymptotic expansions of global errors [9] shows that $u(t_n)-U_n=h^2[\phi(t_n)+(-1)^n\psi(t_n)]+O(h^4)$, with ϕ and ψ smooth functions, which leads to a disparity between even and odd grid values of U_n. (This disparity is evident in the table.) A means of describing the behaviour of U_n may be found in [21] (cf. [28]).

TABLE 2

h	$(u-U)/h^2$	$(z-U)/h^2$
$\frac{1}{3}$	1.13	.69
$\frac{1}{4}$	0.99	.55
$\frac{1}{7}$	1.21	.76
$\frac{1}{8}$	1.03	.58
$\frac{1}{15}$	1.22	.77
$\frac{1}{16}$	1.04	.58

(D) The two-point boundary value problem

$$u(0)=0, \tag{3.7a}$$

$$-u''+u=0,\qquad 0\leqq t\leqq 1, \tag{3.7b}$$

$$u'(1)=1, \tag{3.7c}$$

is discretized by

$$U_0=0, \tag{3.8a}$$

$$-(U_{n-1}-2U_n+U_{n+1})/h^2+U_n=0,\qquad n=1,2,\cdots,N-1, \tag{3.8b}$$

$$(U_N-U_{N-1})/h=1, \tag{3.8c}$$

where $h=1/N$, N a positive integer. We observe that (3.8b) approximates (3.7b) with second order accuracy, while (3.8c) is only a first order accurate replacement of (3.7c). Consequently we obtain the following modified problem, which has second order of correctness

$$z(0)=0, \tag{3.9a}$$

$$-z''+z=0,\qquad 0\leqq t\leqq 1, \tag{3.9b}$$

$$z'(1)-(h/2)z(1)=1, \tag{3.9c}$$

where the last equation has been derived by Taylor expanding (3.8c) and using (3.9b) to eliminate z''. Table 3, which shows values at $t=1$, provides illustration of the fact that (3.8) is first order accurate, while (3.9) coincides with (3.8) up to second order

TABLE 3

h	(Exact-numerical)/h	(Modified-numerical)/h
$\frac{1}{4}$	$-.287$	$+.034$
$\frac{1}{8}$	$-.289$	$+.015$
$\frac{1}{16}$	$-.290$	$+.007$

(E) Our final example is given by the following periodic initial value problem for the heat equation:

$$u(x,0)=u_0(x), \quad -\infty<x<\infty, \tag{3.10a}$$

$$u(x,t)=u(x+1,t), \quad -\infty<x<\infty, \quad t>0, \tag{3.10b}$$

$$u_t=u_{xx}, \quad -\infty<x<\infty, \quad t>0, \tag{3.10c}$$

together with the discretization

$$U_j^0=u_0(jh), \quad j=0,\pm1,\pm2,\cdots, \tag{3.11a}$$

$$U_j^n=U_{j+J}^n \quad n=1,2,\cdots, \quad j=0,\pm1,\pm2,\cdots, \tag{3.11b}$$

$$(U_j^{n+1}-U_j^n)/k=(U_{j-1}^n-2U_j^n+U_{j+1}^n)/h^2, \quad n=0,1,\cdots, \quad j=0,\pm1,\pm2,\cdots. \tag{3.11c}$$

Here u_0 is 1-periodic, $h=1/J$, J a positive integer and $k=rh^2$, with r a positive parameter. We now present in detail the construction of a modified problem of second order of correctness (in k), so as to show the additional novelties involved in dealing with PDEs. A smooth function $w(x,t)$ is substituted in (3.11c) to yield

$$l_j^{n+1}=(w_j^{n+1}-w_j^n)/k-(w_{j-1}^n-2w_j^n+w_{j+1}^n)/h^2,$$

where $w_j^n=w(jh,nk)$. On Taylor expanding, we obtain

$$l_j^{n+1}=(w_t-w_{xx})+\frac{k}{2}\left(w_{tt}-\frac{1}{6r}w_{xxxx}\right)+\cdots \tag{3.12}$$

which leads to

$$w_t=w_{xx}-\frac{k}{2}\left(w_{tt}-\frac{1}{6r}w_{xxxx}\right) \tag{3.13}$$

as a candidate for modified equation. Again (3.13) contains a small parameter in front of the highest derivatives and, because of its high order, requires more side conditions than can be derived from (3.11a)-(3.11b). Differentiation of (3.13), first with respect to t and then with respect to x twice, yields

$$w_{tt}=w_{xxt}-\frac{k}{2}\left(w_{ttt}-\frac{1}{6r}w_{xxxxt}\right),$$

$$w_{xxt}=w_{xxxx}-\frac{k}{2}\left(w_{xxtt}-\frac{1}{6r}D_x^6w\right).$$

These equations can now be used to eliminate w_{tt} from (3.13) and, on discarding terms involving k^2, we arrive at the equation

$$z_t=\left(1-\frac{k}{2}\left(1-\frac{1}{6r}\right)D_x^2\right)z_{xx}. \tag{3.14}$$

We notice in passing that the form

$$\left(1+\frac{k}{2}\left(1-\frac{1}{6r}\right)D_x^2\right)z_t=z_{xx} \tag{3.15}$$

resulting from formally inverting the operator in brackets on the right of (3.14), may also be considered. This alternative form seems advantageous in the case of initial boundary value problems, since it does not increase the number of required boundary conditions. See below.

We now discuss whether (3.14), together with

$$z(x,0)=u_0(x), \quad -\infty<x<\infty, \tag{3.16a}$$

$$z(x+1,t)=z(x,t), \quad -\infty<x<\infty, \quad t>0, \tag{3.16b}$$

provides a modified problem for the study of (3.11). The Fourier transform of (3.14) is given by

$$(d/dt)\hat{z}(m,t)=-\left(1+\frac{k}{2}\left(1-\frac{1}{6r}\right)4m^2\pi^2\right)4m^2\pi^2\hat{z}(m,t)$$

where m is the wave number ($m=0,\pm1,\pm2,\cdots$). This leads to

$$\hat{z}(m,t)=\hat{z}(m,0)\exp[\sigma(m)t] \tag{3.17}$$

where $\sigma(m)$ is the symbol of (3.14)

$$\sigma(m)=-\left(1+\frac{k}{2}\left(1-\frac{1}{6r}\right)4m^2\pi^2\right)4m^2\pi^2.$$

Three ranges of the parameter r should be studied separately.

(i) $\frac{1}{6}\leqq r\leqq\frac{1}{2}$. When r has been fixed within this range the exponential term in (3.17) is bounded for all $t\geqq0$ uniformly in m and k. Therefore the solutions of (3.14)-(3.16) are bounded together with their derivatives uniformly in k. This fact combined with the stability of the scheme (3.11) allows us to conclude that we are dealing with a problem of second order of correctness. (Note that for $r=\frac{1}{6}$ (3.14) reduces to the original equation (3.10c), in agreement with the fact that, for this value of r, (3.11) is convergent for order $O(k^2)$ [15].) Table 4 provides a numerical illustration of the approximation at

$$x=\tfrac{1}{2}, \quad t=\tfrac{1}{4}, \quad u_0(x)=\sum_{l=1}^{\infty}\frac{(-1)^l}{l^6}\cos 2\pi lx, \quad u(\tfrac{1}{2},\tfrac{1}{4})=5.172\times10^{-5}$$

when $r=\frac{1}{4}$.

TABLE 4

h	Numerical $\times 10^5$	Modified $\times 10^5$
$\frac{1}{4}$	26.327	1.875
$\frac{1}{8}$	4.351	4.013
$\frac{1}{16}$	4.851	4.854
$\frac{1}{32}$	5.091	5.091

(ii) $\frac{1}{2}<r$. In this range the exponential term in (3.17) is still bounded. However the scheme is now *unstable* and bounds for the local discretization error do not lead

to bounds for global errors. As a result (3.14)–(3.16) do not model the behaviour of (3.11). This is borne out by Table 5, which is analogous to Table 4 except for the fact that now $r = \frac{2}{3}$.

TABLE 5

h	Numerical $\times 10^5$	Modified $\times 10^5$
$\frac{1}{4}$	-2.05×10^4	0.012
$\frac{1}{8}$	-3.13×10^7	1.129
$\frac{1}{16}$	-4.57×10^{20}	3.535
$\frac{1}{32}$	1.55×10^{39}	4.703

(iii) $0<r<\frac{1}{6}$. In this range the scheme is stable, but the exponential term in (3.17) cannot be bounded uniformly in m, k. Thus (3.14)–(3.16) is not well-posed uniformly in k. (Even for a fixed $k>0$, (3.14)–(3.16) cannot be solved for arbitrary initial data, due to the unboundedness of the exponential term as m varies, a situation similar to that for the backward heat equation. In particular (3.14)–(3.16) does not possess a solution for the initial datum employed in Tables 4–5.)

When attention is restricted to initial data u_0 containing only a prescribed finite number M of harmonics, (3.14), (3.16) may still be of some value, since the exponential in (3.17) is bounded for $m \leqq M$ and k sufficiently small $k \leqq k_0(M)$. This remark has often been expressed in the literature by saying that modified equations are valid only when the product mk is small [34], [16], [33], [20].

Table 6 refers to the initial condition

$$u_0 = \sum_{l=1}^{M} \frac{(-1)^l}{l^6}\cos 2\pi l x,$$

$x=\frac{1}{2}$, $t=\frac{1}{4}$, $r=\frac{1}{8}$, $M=5$, $u(\frac{1}{2},\frac{1}{4}) = 5.172\times 10^{-5}$. As M is increased, the value of h must be decreased accordingly in order to attain a prescribed level of accuracy.

TABLE 6

h	Numerical $\times 10^5$	Modified $\times 10^5$
$\frac{1}{4}$	34.474	2.277×10^{31}
$\frac{1}{8}$	5.927	5.872
$\frac{1}{16}$	5.342	5.339
$\frac{1}{32}$	5.214	5.213

To conclude this example we point out that the alternative modified problem (3.15)–(3.16) is uniformly well-posed, as $k \to 0$, if and only if r lies in the range $0<r\leqq\frac{1}{6}$. Therefore (3.14), (3.15) complement each other and allow a study of the scheme in the entire stable range $0<r\leqq\frac{1}{2}$.

4. Related techniques. The modified equation approach is closely related to other commonly employed means of analysis. We first consider the use of variational equations to study the behaviour of the global error $u-U$. For the sake of simplicity, attention is restricted to the model situation (2.1)–(2.2) with $\delta=0$. It is well known [9] that $U_n = u(t_n) + hv(t_n) + O(h^2)$, where the function $v(t)$ does not depend on h

and satisfies

$$v(0)=0, \tag{4.1a}$$

$$v' = f'(u(t))v - \tfrac{1}{2}u''(t), \qquad 0\leqq t\leqq T. \tag{4.1b}$$

Thus $y(t) = u(t) + hv(t)$ provides a model for the description of the Euler solution accurate to $O(h^2)$. However the determination of $y(t)$ requires successively the solution of the original problem (2.1) and that of the variational problem (4.1). The modified problem approach, on the other hand, involves only the solution of the single problem (2.8). This latter approach is therefore more convenient in practice, where often only qualitative information on the behaviour of U is of interest. Nevertheless the two approaches are closely related, as borne out by the fact that (2.8) can be rigorously derived from (4.1) as follows. On using (2.1b), we can rewrite (4.1b) as

$$v' = f'(u)v - \tfrac{1}{2}f'(u)u'.$$

Hence, since $y = u + hv$,

$$\begin{aligned} y' &= u' + hv' = f(u) + hf'(u)v - \frac{h}{2}f'(u)u' \\ &= f(y) - \frac{h}{2}f'(y)y' + O(h^2), \end{aligned} \tag{4.2}$$

where in the final step we have made use of the smoothness of u and v. Deletion of the $O(h^2)$ remainder in (4.2) can only lead to an $O(h^2)$ change in the solutions and yields (2.8).

This close relationship between the variational and modified equation approaches merely reflects the fact that they are based on the same information, namely the leading terms in the expansion of the local error. This remark applies equally to any strongly stable [29] linear multistep method. For stable linear multistep methods having roots $r\neq 1$, $|r|=1$ the situation is more delicate [9] due to the effect of choice of starting values. (See example C) above.)

The observation that modified problems make use only of the leading terms of the expansion of the local error applies generally, and is primarily responsible for restricting the scope of the method. A further illustration is given in the context of the heat equation example in the previous section. The scheme (3.11) was used there only insofar as to derive (3.12). In turn the modified equations (3.14), (3.15) were based solely on the terms displayed in (3.12); consequently they would serve as modified equations for any scheme that gave rise to the same terms. On the other hand the amplification factor [15]

$$\xi(m) = 1 - 4r\sin^2 m\pi h,$$

where the wave number m is an integer, provides a complete characterization of th scheme and may therefore be used to deduce all its properties. In particular the firs order of local accuracy ($r\neq\frac{1}{6}$) is a consequence of the expansion

$$\frac{\xi - \exp[-(2m\pi)^2 k]}{k} = -\frac{k}{2}(2m\pi)^4\left(1-\frac{1}{6r}\right) + O(k^2 m^6), \tag{4.3}$$

This expression is simply the Fourier transform of (3.12) when w is a solution ((3.10c). The $O(k)$ term in the right of (4.3) is the Fourier transform of the leadir term of the local truncation error, which is the only information required to constru the modified problems. This is reinforced by noting that

$$-\frac{k^2}{2}(2m\pi)^4\left(1-\frac{1}{6r}\right)=\exp\left[\frac{k^2}{2}(2m\pi)^4\left(1-\frac{1}{6r}\right)\right]-1+O(k^4m^8)$$

$$=\exp\left[-(2m\pi)^2k\right]\left\{\exp\left[\frac{k^2}{2}(2m\pi)^4\left(1-\frac{1}{6r}\right)\right]-1\right\}$$

$$+O(k^3m^6),$$

which, together with (4.3), leads to

$$\frac{\xi-\exp(\sigma k)}{k}=O(k^2m^6),$$

where $\sigma=\sigma(m)$ is the symbol of the modified equation (3.14). In other words the symbol $\sigma(m)$ and consequently the modified equation itself, can be derived from the terms displayed in (4.3) without having to resort to the original difference equations.

The study of the stability of the scheme (both for $k\to 0$, t fixed and k fixed, $t\to\infty$) requires complete knowledge of $\xi(m)$, $|hm|\leqq\pi$, information which cannot be deduced from the leading terms of the expansion of $\xi(m)$ around $mh=0$. Consequently, properties such as stability cannot be ascertained from a study of modified problems (cf. Table 5). Therefore the cases reported in the literature where analysis of a modified problem has resulted in the correct stability limits must be regarded as coincidence. These attempts have, by and large, been restricted to cases where the stability had previously been analysed by different means and the stability limits were thus known beforehand.

It may be useful to point out that although the derivation of modified equations only takes into account the behaviour of the numerical scheme for mh small, *well-posed* modified problems describe accurately the numerical solution provided by (Lax) *stable* schemes even if the solution contains all wave numbers (cf. Table 4). The reason for this is that in any initial datum in (say) L^2 the high frequencies are represented with amplitudes which tend to zero as the wave number increases. "It does no harm for these higher harmonics to be falsified" both by the scheme and by the modified equation "provided only that they do not become amplified to such an extent as to be no longer negligible" (see [19, p. 11]. The quoted sentences have been taken from this reference.)

5. Conclusions. The following conclusions have emerged from our study of the method of modified problems.

(i) The construction of a modified problem correct of order p may be undertaken in a purely formal manner. Having arrived at a suitable candidate it is necessary to verify that its solution satisfies the discrete equations except for an $O(h^p)$ remainder. In doing so it is imperative to ensure that any derivatives appearing in the remainder are bounded as $h\to 0$ (cf. Table 1).

(ii) Side conditions in both the original problem and its discretization must be incorporated into the analysis (cf. examples C) and D)).

(iii) Stability as $h\to 0$ of the discrete method being analyzed is an essential prerequisite for the success of the analysis. Without stability, estimates of the local truncation error do not imply estimates of the global error (cf. Table 5).

(iv) Since only a limited amount of information on the scheme is used in constructing a modified problem, such problems cannot provide a full description of the scheme. In particular stability properties, both for h fixed, $t\to\infty$ and $h\to 0$, t fixed, cannot be deduced from a modified problem.

(v) It has often been asserted in the literature that modified partial differential equations provide a valid description of the numerical solution only when the product (wave number)$\times h$ is small. However our analysis has revealed that this is not necessarily the case and that solutions to initial data containing all harmonics can be described, provided that the candidate modified problem satisfies (i)-(iii) above (cf. Table 4 and last paragraph of § 4).

Acknowledgment. The authors would like to express their gratitude to the British Council in Spain for their financial assistance.

REFERENCES

[1] R. C. Y. CHIN AND G. W. HEDSTROM, *A dispersion analysis of difference schemes: Tables of generalized Airy functions*, Math. Comp., 32 (1978), pp. 1163-1170.
[2] K. DEKKER AND J. G. VERWER, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, North-Holland, Amsterdam, 1984.
[3] D. B. DUNCAN AND D. F. GRIFFITHS, *The study of a Petrov-Galerkin method for first order hyperbolic equations*, Comp. Meth. Appl. Mech. Eng., 45 (1984), pp. 147-166.
[4] P. R. GARABEDIAN, *Estimation of the relaxation factor for small mesh sizes*, Mathematical Tables and other Aids for Computation, 10 (1956), pp. 183-185.
[5] C. W. GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
[6] D. F. GRIFFITHS, *Explicit methods for multidimensional advection-diffusion problems*, Proc. VI Congreso de Ecuaciones Diferenciales y Aplicaciones, Universidad de Zaragoza, 1983, pp. 64-80.
[7] A. HARTEN, J. M. HYMAN AND P. D. LAX, Appendix by B. KEYFITZ, *On finite-difference approximations and entropy conditions for shocks*, Comm. Pure Appl. Math., 29 (1976), pp. 297-322.
[8] G. W. HEDSTROM, *Models of difference schemes for $u_t+u_x=0$ by partial differential equations*, Math Comp., 29 (1975), pp. 969-977.
[9] P. HENRICI, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley, New York, 1963
[10] A. HINDMARSH, P. M. GRESHO AND D. F. GRIFFITHS, *The stability of explicit Euler time-integration for certain finite difference approximations of the multi-dimensional advection-diffusion equation*, Int J. Numer. Meth. Fluids., 4 (1984), pp. 853-897.
[11] C. W. HIRT, *Heuristic stability theory of finite difference equations*, J. Comp. Phys., 2 (1968), pp. 339-35
[12] E. ISAACSON AND H. B. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1966.
[13] J. D. LAMBERT, *Computational Methods in Ordinary Differential Equations*, John Wiley, London, 197
[14] A. LERAT, *Numerical shock structure and nonlinear corrections for difference schemes in conservation form*, Proc. 6th International Conference on Numerical Methods in Fluid Dynamics, H. Cabann et al., eds., Springer-Verlag, Berlin, 1979.
[15] A. R. MITCHELL AND D. F. GRIFFITHS, *The Finite Difference Method in Partial Differential Equation* John Wiley and Sons, Chichester, 1980.
[16] K. W. MORTON, *Initial value problems by finite difference and other methods*, State of the Art Numerical Analysis, D. A. H. Jacobs, ed., Academic Press, London, 1977, pp. 699-756.
[17] C. PALENCIA AND J. M. SANZ-SERNA, *Equivalence theorems for incomplete spaces: an apprais* IMA J. Numer. Anal., 4 (1984), pp. 109-115.
[18] ———, *An extension of the Lax-Richtmeyer theory*, Numer. Math., 44 (1984), pp. 279-283.
[19] R. D. RICHTMEYER AND K. W. MORTON, *Difference Methods for Initial-Value Problems*, John Wile Interscience, London, 1967.
[20] P. J. ROACHE, *Computational Fluid Dynamics*, Hermosa, Albuquerque, NM, 1976.
[21] J. M. SANZ-SERNA, *Studies in numerical nonlinear instability I: Why do leap-frog schemes go unstab* this Journal, 6 (1985), pp. 923-938.
[22] ———, *Convergence of the Lambert-McLeod trajectory solver and of the CELF method*, Numer. Mat 45 (1984), pp. 173-182.
[23] J. M. SANZ-SERNA AND C. PALENCIA, *A general equivalence theorem in the theory of discretizat methods*, Math. Comp., to appear.
[24] J. M. SANZ-SERNA AND J. G. VERWER, *A Study of the recursion $y_{n+1}=y_n+\tau y_n^m$*, Centrum v Wiskunde en Informatica, Report NM-R8403, Amsterdam, 1984.
[25] L. F. SHAMPINE AND M. K. GORDON, *Computer Solution of Ordinary Differential Equations*, W. Freeman, San Francisco, CA, 1975.

[26] Y. I. SHOKIN, *The Method of Differential Approximation*, Springer-Verlag, Berlin, 1983.
[27] M. N. SPIJKER, *Equivalence theorems for non-linear finite-difference methods*, Lecture Notes in Mathematics, 395, R. Ansorge and W. Tornig, eds., Springer-Verlag, Berlin, 1974, pp. 109-122.
[28] H. J. STETTER, *Symmetric two-step algorithms for ordinary differential equations*, Computing, 5 (1970), pp. 267-280.
[29] ———, *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag, Berlin, 1973.
[30] G. STRANG, *Accurate partial difference methods* II. *Nonlinear problems*, Numer. Math., 6 (1964), pp. 37-46.
[31] F. STUMMEL, *Diskrete Konvergenz linearer Operatoren*, I, Math. Ann., 190 (1970), pp. 45-92.
[32] G. VAINIKKO, *Funktionalanalysis der Diskretisierungsmethoden*, Teubner, Leipzig, 1976.
[33] R. VICHNEVETSKY AND J. B. BOWLES, *Fourier Analysis of Numerical Approximations of Hyperbolic Equations*, SIAM Studies in Applied Mathematics, 5, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982.
[34] R. F. WARMING AND B. J. HYETT, *The modified equation approach to the stability and accuracy analysis of finite difference methods*, J. Comp. Phys., 14 (1974), pp. 159-179.
[35] P. WILDERS, *Minimization of dispersion in difference methods for hyperbolic conservation laws*, Ph.D. thesis, Mathematisch Centrum, Amsterdam, 1983.
[36] A. LERAT AND R. PEYRET, *Propriétés dispersives et dissipatives d'une classe de schémas aux différénces pour les systèmes hyperboliques non linéaires*, Rech. Aérosp., 2 (1975), pp. 61-79.
[37] J. M. SANZ-SERNA, *Stability and convergence in numerical analysis* I: *linear problems, a simple comprehensive account*, preprint (available from the author).

SIAM J. SCI. STAT. COMPUT.
Vol. 7, No. 3, July 1986

021

AN EXTERIOR POISSON SOLVER USING FAST DIRECT METHODS AND BOUNDARY INTEGRAL EQUATIONS WITH APPLICATIONS TO NONLINEAR POTENTIAL FLOW*

DAVID P. YOUNG†‡, ALEX C. WOO‡¶, JOHN E. BUSSOLETTI‡§ AND FORRESTER T. JOHNSON‡§

Abstract. A general method is developed combining fast direct methods and boundary integral equation methods to solve Poisson's equation on irregular exterior regions. The method requires $O(N \log N)$ operations where N is the number of grid points. Error estimates are given that hold for regions with corners and other boundary irregularities. Computational results are given in the context of computational aerodynamics for a two-dimensional lifting airfoil. Solutions of boundary integral equations for lifting and nonlifting aerodynamic configurations using preconditioned conjugate gradient are examined for varying degrees of thinness.

Key words. partial differential equations, fast direct methods, boundary integral equations, fast Poisson solvers, preconditioned conjugate gradient, transonic potential flow

1. Introduction. Fast direct methods have been used extensively in recent years to solve Poisson's equation on rectangular and other separable domains [1], [2]. Much work has been devoted to extending these methods to other elliptic partial differential equations and/or nonseparable domains. In particular, for irregular geometries the analogy of capacitance matrices with potential theory has been exploited by Proskurowski and Widlund [3], [4]. In this paper, we show how a consistent, second-order boundary integral discretization can be implemented using fast direct methods. The starting point is the classical theory of double- and single-layer potentials. If N is the number of grid points, our discretization enables a solution of Poisson's equation in $O(N \log N)$ opeations which retains the spectral properties of the boundary integral formulation. This discretization has certain advantages with regard to conditioning of the matrices, flexibility in boundary discretization, and computation of quantities such as surface pressures.

In § 2, we outline the hybrid method in the context of boundary integral (panel) methods. Section 3 explains how the boundary integral problem is approximated using an exterior fast solver and an error estimate is given. Section 4 presents some two-dimensional computational results. Section 5 explains some iterative techniques for solving the linear system resulting from the approximations given in § 3 (such techniques are necessary in three dimensions). It also contains the results of some numerical experiments with preconditioned conjugate gradient. In § 6 we put our work in the context of previous work in this area.

2. The hybrid method. Because of the extreme sensitivity of airfoil problems to small perturbations in geometry [5], [6], [7], panel (boundary integral) methods with their accurate representation of surfaces have long been standard for linear potential flow calculations. But implementations of these methods have not been optimal computationally. Fast direct methods and boundary integral methods can be combined for the Poisson problem with advantages over either method alone. Consider the boundary

* Received by the editors August 2, 1983, and in revised form February 1, 1985.
† Boeing Computer Services Company, Tubwila, Washington 98188. The work of this author was partially supported by the National Science Foundation under grant MCS 80-12220.
‡ The work of this author was partially supported by NASA Ames Contract NAS2-9830 (PAN AIR Development).
§ Boeing Military Airplane Co., Seattle, Washington 98124.
¶ NASA/Ames Research Center, Moffett Field, California 94035.