MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 137
JANUARY 1977, PAGES 112–123

Functional Fitting—New Family of Schemes for Integration of Stiff O.D.E.*

By Arieh Iserles

Abstract. Following the ideas of Liniger and Willoughby (although acting along different lines), a new family of schemes, derived from the trapezoidal rule, is developed, which enables the fitting of the scheme to arbitrary sets of equations.

1. Introduction. The optimality of the trapezoidal rule, in terms of maximal order (among the multistep methods with constant coefficients) and minimal truncation error when applied to stiff ordinary differential equation (O.D.E) systems, proved by Dahlquist [1], is a well-known fact. Its simplicity and cheapness of maintenance do not need any additional proof. In spite of this, the trapezoidal rule (T.R.) has some serious disadvantages which influence the numerical solution:

a. The poor performance on the "transient" segments of the solution (a disadvantage common to almost all methods for stiff O.D.E.).

b. The overestimation of rapidly decaying components of the solution and the consequent lack of L-stability.

c. Oscillations of the solution, introduced by the alternating sign of the characteristic function, when applied to linear equations.

d. The low order.

In the present paper an attempt is made to overcome these disadvantages and to show that schemes *derived* from T.R. can be quite efficient, in spite of the basic deficiencies of the original scheme.

2. The Functional Fitting. Liniger and Willoughby [2] introduced the concept of exponential fitting and suggested three new A-stable schemes. These schemes are accurate for one or two particular and previously chosen scalar linear equations. In spite of several critical deficiencies of these schemes–lowering of the order or computation of the second derivative (and, consequently, also its Jacobian matrix $\partial f'/\partial x$), insufficiency of the fitting to only one or two exponentials and the arbitrariness in the selection of the parameters–the basic idea of Liniger and Willoughby is remarkably deep, and it can lead us to far better results. To sum up this idea in a sentence, we can assume that in the solution of stiff O.D.E, *the fitting to decaying functions which approximate the foreseen behavior of the solution is more important than the order*, i.e. the fitting to polynomials.

Extrapolation, either polynomial (Romberg) or rational (Bulrisch-Stöer) is a typical situation in which a numerical facility is applied to increase the order of the solution.

Received January 5, 1976; revised March 25, 1976 and May 27, 1976.
AMS (MOS) subject classifications (1970). Primary 65L05; Secondary 41A20.

*This paper is based on parts of the author's doctoral thesis written under the supervision of Professor Giacomo Della Riccia of the Ben-Gurion University, Beersheba, Israel.

Nevertheless, as stated above, this is of little help when we deal with stiff O.D.E. Moreover, the extrapolation often destroys many of the stability properties.

This property of the extrapolation schemes is not intrinsic, and one can hope to obtain new extrapolation techniques which will be more appropriate for stiff systems. In this paper we shall try to show that this hope is fully justified.

In other words, at every stage of extrapolation we have a set of extrapolants, which are simply approximations to the solution. The extrapolated value is obtained by some "averaging" of the extrapolants, either in a linear or in some nonlinear mode. The goal of the conventional "averaging" is to increase the order of the solution, or, anticipating the nomenclature used later in this paper, to fit the scheme to polynomials of order greater than the original order of the scheme. This goal must be changed when we solve stiff O.D.E., because polynomial behavior is highly uncharacteristic of the behavior of the solution of stiff systems, which usually decay asymptotically. Solving stiff systems, we must "average" the extrapolants in another mode, fitting the scheme to decaying functions which approximate the foreseen behavior of the solution.

Let us introduce a number of notations:

For a set of different and increasingly ordered integers $\{l_1, l_2, \ldots, l_m\}$ we denote by $\mathbf{x}^{(i)}$ the solution of

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \in E^N, \tag{2-1}$$

obtained in $t_{n+1} = t_n + h$, by l_i consecutive applications of the T.R. scheme:

$$\mathbf{x}_{n+k/l_i} = \mathbf{x}^{(i)}_{n+((k-1)/l_i)} + \frac{h}{2l_i}(\dot{x}^{(i)}_{n+(k-1)/l_i} + \dot{x}^{(i)}_{n+k/l_i}) \tag{2-2}$$

with the step length h/l_i.

Let us assume we have a set of scalar ordinary differential problems

$$\dot{z}^{(j)} = f^{(j)}(t, z^{(j)}), \quad z^{(j)}(t_n) = 1, \qquad 1 \leqslant j \leqslant M, \tag{2-3}$$

such that $\lim_{t\to\infty} z^{(j)}(t) = 0$ for every j. We denote by z_{ij} the solution of (2-3) for j in $t_{n+1} = t_n + h$, obtained by l_i consecutive applications of the T.R. scheme with the step length h/l_i.

Finally, let us denote by $\{\omega_p^{(k,m)}\}_{p=1}^{\binom{m}{k}}$ the set of all permutations of $\{1, 2, \ldots, m\}$ of length $k \leqslant m-1$, when $\omega_p^{(1,m)} = p$.

THEOREM 1. *If* $M = 2^{m+1} - m - 4$ *and* $\{\{\beta_{\omega_p^{(k,m)}}, \gamma_{\omega_p^{(k,m)}}\}_{p=1}^{\binom{m}{k}}\}_{k=1}^{m-1}$ *is any solution of the linear algebraic system*:

$$\sum_{i=1}^{m} \beta_{\omega_i^{(1,m)}} = 0, \tag{2-4}$$

$$\sum_{i=1}^{\binom{m}{k}} \beta_{\omega_i^{(k,m)}} = \sum_{i=1}^{\binom{m}{k-1}} \gamma_{\omega_i^{(k-1,m)}}, \qquad 2 \leqslant k \leqslant m-1, \tag{2-5}$$

$$\sum_{i=1}^{m} \gamma_{\omega_i^{(m-1,m)}} = 1, \tag{2-6}$$

$$\sum_{k=1}^{m-1}\left[\sum_{p=1}^{\binom{m}{k}}\left(\prod_{i\in\omega_p^{(k,m)}} z_{ij}\right)\beta_{\omega_p^{(k,m)}}\right] - z^{(j)}(t_{n+1})$$

$$\times \sum_{k=1}^{m-1}\left[\sum_{p=1}^{\binom{m}{k}}\left(\prod_{i\in\omega_p^{(k,m)}} z_{ij}\right)\gamma_{\omega_p^{(k,m)}}\right] = -\prod_{i=1}^{m} z_{ij}, \tag{2-7}$$

$$1 \leqslant j \leqslant M = 2^{m+1} - m - 4,$$

then the scheme

$$g_q(x_q^{(1)}, x_q^{(2)}, \ldots, x_q^{(m)})$$

$$= \frac{\sum_{k=1}^{m-1}\left[\sum_{p=1}^{\binom{m}{k}}\beta_{\omega_p^{(k,m)}}\prod_{i\in\omega_p^{(k,m)}} x_q^{(i)}\right] + \prod_{k=1}^{m} x_q^{(k)}}{\sum_{k=1}^{m-1}\left[\sum_{p=1}^{\binom{m}{k}}\gamma_{\omega_p^{(k,m)}}\prod_{i\in\omega_p^{(k,m)}} x_q^{(i)}\right]}, \qquad 1 \leqslant q \leqslant n, \tag{2-8}$$

is at least of order 2, *and it is fitted to the set* $\{f^{(j)}\}_{j=1}^{M}$; *i.e. it solves exactly Eqs.* (2-3) *in* t_{n+1}.

Proof. $\{\mathbf{x}^{(k)}\}_{k=1}^{m}$ are solutions obtained by application of the T.R., which is a second order scheme. A necessary and sufficient condition for this is that if $f(t, x) = Lt^{L-1}$, $L = 0, 1, 2$, and $x(t_n) = 0$ (which may be assumed without loss of generality) then $x^{(k)} \equiv h^L$, $k = 1, 2, \ldots, m$. By the same reasoning, this is the condition which ensures that (2-8) is of the second order. But

$$g(h^L, h^L, \ldots, h^L) = h^L, \qquad 0 \leqslant L \leqslant 2,$$

implies

$$\sum_{k=1}^{m-1}\left[\sum_{p=1}^{\binom{m}{k}}\beta_{\omega_p^{(k,m)}}\right]h^{kL} + h^{mL} = \sum_{k=1}^{m-1}\left[\sum_{p=1}^{\binom{m}{k}}\gamma_{\omega_p^{(k,m)}}\right]h^{(k+1)L}.$$

This equation must be valid for arbitrary h, and therefore, we can equate the coefficients of terms with the same powers of h. This implies immediately (2-4), (2-5) and (2-6). Therefore, every scheme (2-8) for which these m equations hold is at least of the second order.

There are $\binom{m}{k}$ permutations of $\{1, 2, \ldots, m\}$ of length k; and therefore, the set $\{\{\beta_{\omega_p^{(k,m)}}, \gamma_{\omega_p^{(k,m)}}\}_{p=1}^{\binom{m}{k}}\}_{k=1}^{m-1}$ contains $2(2^m - 2) = 2^{m+1} - 4$ different components. Equations (2-4), (2-5) and (2-6) define m of these; and thus, in order to define all components it is necessary to add $M = 2^{m+1} - m - 4$ more linear equations. Setting

$$g(z_{1j}, z_{2j}, \ldots, z_{mj}) = z^{(j)}(t_{n+1}), \qquad 1 \leqslant j \leqslant M,$$

where $z^{(j)}(t_{n+1})$ is the exact solution of $z^{(j)} = f^{(j)}(t, z^{(j)})$, $z^{(j)}(t_0) = 1$, we easily derive Eqs. (2-7). Q.E.D.

Applying similar reasoning, we can prove immediately the two following weaker (but more useful) results:

THEOREM 1A. *If* $M = 2^m - m - 1$ *and* $\{\{\delta_{\omega_p^{(k,m)}}\}_{p=1}^{\binom{m}{k}}\}_{k=1}^{m-1}$ *is any solution of the linear algebraic system*

$$\sum_{p=1}^{m} \delta_{\omega_p^{(k,m)}} = 1,$$

$$\sum_{p=1}^{\binom{m}{k}} \delta_{\omega_p^{(k,m)}} = 0, \qquad 2 \leqslant k \leqslant m-1,$$

$$\sum_{k=1}^{m-1} \sum_{p=1}^{\binom{m}{k}} \left(\prod_{i \in \omega_p^{(k,m)}} z_{ij} \right) \delta_{\omega_p^{(k,m)}} = z^{(j)}(t_{n+1}), \qquad 1 \leqslant j \leqslant M,$$

then the scheme

$$g_q(x_q^{(1)}, x_q^{(2)}, \ldots, x_q^{(m)}) = \sum_{k=1}^{m-1} \left[\sum_{p=1}^{\binom{m}{k}} \delta_{\omega_p^{(k,m)}} \prod_{i \in \omega_p^{(k,m)}} x_q^{(i)} \right], \tag{2-9}$$

$$\mathbf{x}_{n+1} = \mathbf{g},$$

is at least of order 2 *and is fitted to the set* $\{f^{(j)}\}_{j=1}^{M}$ *in* t_{n+1}.

THEOREM 1B. *If* $M = m - 1$ *and* $\{\eta_p\}_{p=1}^{m}$ *is any solution of the linear algebraic system*

$$\sum_{p=1}^{m} \eta_p = 1, \quad \sum_{p=1}^{m} \eta_p z_{pj} = z^{(j)}(t_{n+1}), \qquad 1 \leqslant j \leqslant M, \tag{2-10}$$

then the scheme

$$g_q(x_q^{(1)}, x_q^{(2)}, \ldots, x_q^{(m)}) = \sum_{p=1}^{m} \eta_p x_q^{(p)} \tag{2-11}$$

is at least of order 2 *and is fitted to the set* $\{f^{(j)}\}_{j=1}^{M}$ *in* t_{n+1}.

The apparent conclusion of the Theorem 1, 1A and 1B is that *we can apply an extrapolation technique in order to fit the scheme to an arbitrary set of scalar equations of type* (2-3).

In the following, we will devote our attention particularly to the scheme (2-11), the simplest one. The discussion will be restricted to the exponential fitting only, i.e. to fitting to linear equations, both because of the significance of these equations as the first approximation to any nonlinear system and of the considerable difficulty in any more sophisticated fitting.**

**The problem of fitting a scheme to nonlinear functions, based on a quite different idea, is treated in another paper [4].

The exponentially fitted scheme (2-11) has the following elegant form:

$$\sum_{p=1}^{m} \eta_p = 1,$$

$$\sum_{p=1}^{m} \left(\frac{l_p + \phi_j h/2}{l_p - \phi_j h/2} \right)^{l_p} \eta_p = e^{\phi_j h}, \qquad 1 \leqslant j \leqslant m-1, \tag{2-12}$$

when the set (2-3) is

$$\dot{z}^{(j)} = \phi^{(j)} z^{(j)}, \quad z^{(j)}(t_n) = 1, \quad \operatorname{Re} \phi^{(j)} \leqslant 0, \qquad 1 \leqslant j \leqslant m-1,$$

and $h = t_{n+1} - t_n$.

The scheme (2-11) resembles the familiar linear (Romberg) extrapolation scheme. The following lemma shows the connection between the two approaches:

LEMMA 2. *If* $\mathsf{F}(\phi_1, \ldots, \phi_{m-1})$ *is the space of solutions of* (2-12) *and* $\mathsf{F}_0 = \lim_{\phi_i \to 0;\, 1 \leqslant i \leqslant m-1} \mathsf{F}(\phi_1, \ldots, \phi_{m-1})$, *if* $\mathbf{a}^{(l)} = (a_1^{(l)}, \ldots, a_m^{(l)})$ *are the coefficients of the* l*th order odd-power linear extrapolation, then*

$$\mathsf{F}_0 = \left\{ (\eta_1, \eta_2, \ldots, \eta_m) : \sum_{p=1}^{m} \eta_p = 1 \right\} \cap \operatorname{Sp}\{\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(m-1)}\}.$$

Proof. Let

$$Q_k = \left(\frac{l_k + \tilde{t}/2}{l_k - \tilde{t}/2} \right)^{l_k}, \qquad 1 \leqslant k \leqslant m, \ |\tilde{t}| < 2l_1,$$

If we denote $l = l_k$, $s = \tilde{t}/2l_k$, then $Q_k = ((1+s)/(1-s))^l$ and for $|s| < 1$:

$$\begin{aligned} Q_k &= [(1+s)(1+s+s^2+\cdots)]^l = (1 + 2s + 2s^2 + 2s^3 + 2s^4)^l + O(s^5) \\ &= 1 + 2ls + 2l^2 s^2 + \frac{2}{3} l(2l^2+1)s^3 + \frac{2}{3} l^2(l^2+2)s^4 + O(s^5) \\ &= 1 + \tilde{t} + \frac{1}{2}\tilde{t}^2 + \frac{1}{6}\left(1 + \frac{1}{2l_k^2}\right)\tilde{t}^3 + \frac{1}{24}\left(1 + \frac{2}{l_k^2}\right)\tilde{t}^4 + \theta(\tilde{t}^5). \end{aligned}$$

Now, if we define

$$P_k = \frac{1}{\tilde{t}^3}(Q_1 - Q_k) \quad \text{and} \quad S = \frac{1}{\tilde{t}^3}(Q_1 - e^{\tilde{t}}),$$

then

$$P_k = \frac{1}{12}\left(\frac{1}{l^2} - \frac{1}{l_k^2}\right)(1 + \tilde{t}) + \theta(\tilde{t}^2), \qquad 1 \leqslant k \leqslant m,$$

and

$$S = \frac{1}{12}\frac{1}{l_1^2}(1 + \tilde{t}) + (\tilde{t}^2).$$

It is easy to observe that fitting to $\dot{x} = \tilde{t}x$, $x(0) = 1$, is equivalent to

$$\sum_{k=1}^{m} \eta_k = 1, \quad \sum_{k=1}^{m} P_k \eta_k = S.$$

Therefore, when $\tilde{t}$ tends to zero we obtain

$$\sum_{k=1}^{m} \left(\frac{1}{l_1^2} - \frac{1}{l_k^2} \right) \eta_k = \frac{1}{l_1^2}$$

or

$$\frac{1}{l_1^2} \left(\sum_{k=1}^{m} \eta_k - 1 \right) = \sum_{k=1}^{m} \frac{1}{l_k^2} \eta_k.$$

But $\Sigma_{k=1}^{m} \eta_k = 1$, and therefore $\Sigma_{k=1}^{m} \eta_k / l_k^2 = 0$. This equation is independent of $\tilde{t}$; and therefore, if all the ϕ_j tend to zero, Eq. (2-12) has the form

$$\sum_{k=1}^{m} \eta_k = 1, \quad \sum_{k=1}^{m} \frac{1}{l_k^2} \eta_k = 0. \tag{2-13}$$

Let $e^{(k)}$ be the error of the T.R. scheme in the solution of the equation $\dot{x} = 3t^2$, $x(0) = 0$, using l_k equal substeps of length h/l_k in $[0, h]$. Then, according to the well-known Euler-Maclaurin formula (Ralston, [3, p. 133]), $e^{(k)} = \frac{1}{2} h^3 / l_k^2$.

Applying any odd-power extrapolation of T.R., the error due to h^3 vanishes. Therefore,

$$\sum_{k=1}^{m} \frac{1}{l_k^2} a_k^{(l)} = 0.$$

Moreover, every extrapolation is order preserving (in fact–order increasing); and thus, $\Sigma_{k=1}^{m} a_k^{(l)} = 1$ and (2-13) holds.

If $\xi \in \{\eta: \Sigma_{i=1}^{m} \eta_i = 1\} \cap \mathrm{Sp}\{\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m-1)}\}$, then there exists $\lambda = (\lambda_1, \ldots, \lambda_{m-1})$ such that

$$\xi = \sum_{i=1}^{m-1} \lambda_i \mathbf{a}^{(i)} \quad \text{and} \quad \sum_{i=1}^{m-1} \lambda_i = 1$$

and so (2-13) holds for ξ and $\xi \in F_0$.

The vectors $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m-1)}$ are linearly independent, because the m by $m - 1$ matrix $[\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m-1)}]$ is triangular with nonvanishing diagonal elements (if the highest-index coefficient of certain extrapolation vanishes, then this extrapolation does not depend on the highest-order extrapolant; therefore, it must be of lower degree, which contradicts the improvement of the order). Therefore, these vectors span the whole space of solutions of $\Sigma_{k=1}^{m} \eta_k / l_k^2 = 1$; and so,

$$F_0 = \left\{ \eta: \sum_{i=1}^{m} \eta_i = 1 \right\} \cap \mathrm{Sp}\{\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(m-1)}\}. \quad \text{Q.E.D.}$$

It is evident from Lemma 2 that application of the scheme (2-11) to nonstiff systems resembles the conventional (Romberg) extrapolation technique. The user of the

O.D.E. solving programme usually either does not know or is not interested in such "technicalities" like stiffness, and he wants a universal O.D.E. solver. The consequence of Lemma 2, that is the scheme (2-11) adjusts itself to both stiff and nonstiff systems, is important from this point of view.

Stability Analysis. The scheme (2-11) is obviously A-stable, because if we make an attempt to solve the scalar equation $x = \lambda x$, fitting of the scheme to this particular equation itself is nothing but natural.

Let us consider the less trivial problem: given some particular choice of parameters for the scheme (2-11) and given the multidimensional differential equation

$$\dot{\mathbf{x}} = A\mathbf{x}, \quad \mathbf{x}(0) = \mathbf{x}_0 \in E^N,$$

when all the real parts of the eigenvalue of A are negative, if we proceed from $t_0 = 0$ with constant positive step h, is $\lim_{t\to\infty} \|\mathbf{x}(t)\| = 0$. Let us call a scheme for which this limit actually exists an *MA-stable* (abbreviation for *Matricial A-stable*) scheme.

LEMMA 3. *Suppose we have an MA-stable scheme and* $\mathbf{x}^{(i)}$ *is the solution obtained at* $t_{n+1} = t_n + h$ *by* l_i *consecutive applications of this scheme with the step length* h/l_i, $1 \leqslant i \leqslant m$. *Then the scheme* $\mathbf{x}_{n+1} = \Sigma_{i=1}^m \eta_i \mathbf{x}_i$, $\Sigma_{i=1}^m \eta_i = 1$, $\eta_i \in R$, $\forall 1 \leqslant i \leqslant m$, *is MA-stable if* $\eta_i \geqslant 0$, $1 \leqslant i \leqslant m$.

Proof. The scheme is *MA*-stable if and only if $\|\mathbf{x}_{n+1}\| < \|\mathbf{x}_n\|$ for every n. Then, for every i, $1 \leqslant i \leqslant m$, $\|\mathbf{x}^{(i)}\| < \|\mathbf{x}_n\|$. But if $\eta_i \geqslant 0$, $1 \leqslant i \leqslant m$, then

$$\|\mathbf{x}_{n+1}\| = \left\|\sum_{i=1}^m \eta_i \mathbf{x}^{(i)}\right\| \leqslant \sum_{i=1}^m \eta_i \|\mathbf{x}^{(i)}\| < \sum_{i=1}^m \eta_i \|\mathbf{x}_n\| = \|\mathbf{x}_n\|$$

that is, the condition $\eta_i \geqslant 0$, $1 \leqslant i \leqslant m$, is sufficient for the *MA*-stability of (2-11). Q.E.D.

LEMMA 4. *The scheme* $\mathbf{x}_{n+1} = \eta \mathbf{x}^{(1)} + (1 - \eta)\mathbf{x}^{(2)}$, *where* $\mathbf{x}^{(1)}$ *and* $\mathbf{x}^{(2)}$ *are solutions obtained by the application of T.R. with one step and two half-steps respectively, is MA-stable if and only if* $\eta \in [0, 1]$.

Proof. The trapezoidal rule is *MA*-stable, and therefore by Lemma 3 if $\eta \in [0, 1]$ then the scheme $\mathbf{x}_{n+1} = \eta \mathbf{x}^{(1)} + (1 - \eta)\mathbf{x}^{(2)}$ is *MA*-stable.

On the other hand, let us assume that the above scheme is *MA*-stable. In order to prove the necessity of $\eta \in [0, 1]$ it is sufficient to check the behavior of the scheme for the systems $\dot{\mathbf{x}} = D\mathbf{x}$, $\mathbf{x}(0) = \mathbf{x}_0$, where D is diagonal and $\operatorname{Re} D_{ii} < 0$, $i = 1, 2, \ldots, n$. This is equivalent to considering the scalar systems $\dot{x} = \lambda x$, $x(0) = x_0$ for arbitrary λ, $\operatorname{Re} \lambda < 0$, (in general, λ is not the parameter to which the scheme is fitted). Without loss of generality we may assume $x_n = 1$, and then

$$x^{(1)} = \frac{2 + h\lambda}{2 - h\lambda}, \quad x^{(2)} = \left(\frac{4 + h\lambda}{4 - h\lambda}\right)^2.$$

If we denote $z(\mu) = (2 + \mu)/(2 - \mu)$, then

$$x_{n+1} = \eta z(h\lambda) + (1 - \eta)\left(\frac{3z(h\lambda) + 1}{z(h\lambda) + 3}\right)^2.$$

The mapping $z(\mu)$ is a Möbius transformation of the complex left half-plane onto

the closed unit circle. Therefore, for particular μ there exist R and θ, $0 \leqslant R \leqslant 1$ and $0 \leqslant \theta < 2\pi$, such that $z(\mu) = Re^{i\theta}$. The mapping $x_{n+1} = x_{n+1}(z)$ is analytic for $|z| < 1$, and obviously is not constant; therefore, its absolute value attains its maximum on $|z| = 1$. But

$$|x_{n+1}|^2 = \eta^2 R^2 + (1-\eta)^2 \frac{81R^4 + 108\cos\theta R^3 + (36\cos^2\theta + 18)R^2 + 12\cos\theta R + 1}{R^4 + 12\cos\theta R^3 + (36\cos^2\theta + 18)R^2 + 108\cos\theta R + 81}$$

$$+ 2\eta(1-\eta) \frac{9\cos\theta R^5 + (48 + 12\cos^2\theta)R^4 + (114\cos\theta + 4\cos^3\theta)R^3 + (48 + R\cos^2\theta)R^2 + 9\cos\theta R}{R^4 + 12\cos\theta R^3 + (36\cos^2\theta + 18)R^2 + 108\cos\theta R + 81};$$

and if $R = 1$, then

$$|x_{n+1}|^2 = 2Q\eta(\eta - 1) + 1,$$

where

$$Q = 1 - \frac{24 + 33\cos\theta + 6\cos^2\theta + \cos^3\theta}{25 + 30\cos\theta + 9\cos^2\theta} = \frac{(1 - \cos\theta)^3}{(5 + 3\cos\theta)^2} \geqslant 0.$$

Therefore, $|x_{n+1}|^2 \leqslant 1$ implies $Q\eta(\eta - 1) \leqslant 0$; thus $\eta(\eta - 1) \leqslant 0$ and $\eta \in [0, 1]$. Q.E.D.

Lemma 3 supplied a sufficient condition for *MA*-stability. This condition had been proved to be necessary for the simplest case, but for more complicated cases it is possible to obtain schemes which are *MA*-stable despite the appearance of negative coefficients.

The necessary conditions for MA-stability of the scheme (2-11) for arbitrary $(l_1, \ldots, l_m)$ are complicated, and the author has not yet succeeded in developing such general criteria. What has been found is that the problem can be reduced to the following problem in complex function theory.

If

$$\sigma_k(t) = \left(\frac{l_k + it}{l_k - it}\right)^{l_k} \quad (|\sigma_k(t)| \equiv 1), \qquad 1 \leqslant k \leqslant m,$$

and

$$\sigma(t) = \sum_{k=1}^{m} \eta_k \sigma_k(t), \qquad \sum_{k=1}^{m} \eta_k = 1,$$

what are the conditions for $\max_{-\infty < t < \infty} |\sigma(t)| \leqslant 1$. ***

Furthermore, even if we obtain a set of complicated conditions on $(\eta_1, \ldots, \eta_m)$ which are necessary for *MA*-stability, the main problem will be still open: what are the $(m - 2)$-tuples of scalar linear equations to which we can fit the scheme (2-11) to achieve *MA*-stability? This problem has been solved here only for the simplest case:

*** The author wishes to thank Professor Giacomo Della Riccia from the Ben-Gurion University and Professor Itzhak Katzenelson from the Hebrew University for their generous help in his fruitless efforts to solve this problem.

LEMMA 5. *Exponential fitting of T.R. by the extrapolation of the solutions obtained by one step and two half-steps respectively, for real negative argument* λ *of the fitted equation* $\dot{x} = \lambda x$, *is MA-stable if and only if*

$$\chi^{(1)}(\mu) = \frac{2+\lambda}{2-\lambda} \leqslant e^{\lambda} \leqslant \left(\frac{4+\lambda}{4-\lambda}\right)^2 = \chi^{(2)}(\mu) \qquad (h=1) \tag{2-15}$$

i.e. $\mu \leqslant -4.798\ldots$.

Proof. Let us fit

$$\chi(\lambda) = \eta\chi^{(1)}(\lambda) + (1-\eta)\chi^{(2)}(\lambda) = \eta\frac{2+\lambda}{2-\lambda} + (1-\eta)\left(\frac{4+\lambda}{4-\lambda}\right)^2 = e^{\lambda};$$

and then

$$\eta = \left(e^{\lambda} - \left(\frac{4+\lambda}{4-\lambda}\right)^2\right) \Big/ \left(\frac{2+\lambda}{2-\lambda} - \left(\frac{4+\lambda}{4-\lambda}\right)^2\right), \qquad \lambda < 0.$$

It is easily verified that

$$\frac{2+\lambda}{2-\lambda} - \left(\frac{4+\lambda}{4-\lambda}\right)^2 < 0 \quad \text{for } \lambda < 0;$$

thus the necessary and sufficient condition of Lemma 4, $0 \leqslant \eta \leqslant 1$, implies immediately (2-15). Q.E.D.

The stability conditions for the scheme

$$x_{n+1} = \eta_1 x_{n+1}^{(1)} + \eta_2 x_{n+1}^{(2)} + (1-\eta_1-\eta_2)x_{n+1}^{(3)},$$

for the sequence $(l_1, l_2, l_3) = (1, 2, 3)$, have been compiled by a computer. If we fit to the arguments μ_1 and μ_2 $(h = 1)$, it has been proved that if $\max\{\mu_1, \mu_2\} > -5.03025$, then one of the coefficients η_1, η_2 or $1 - \eta_1 - \eta_2$ is outside the unit interval. On the other hand, if, say, $\mu_1 < -5.03025$ and $\mu_2 \in (\mu_1, \tilde{\mu}(\mu_1)]$, when $\tilde{\mu}(\mu_1) \leqslant -5.03025$ is a certain critical value, then all the coefficients are inside the unit interval and the scheme is *MA*-stable. These critical values are listed below:

μ_1	$\tilde{\mu}(\mu_1)$
−6	−5.99999
−10	−5.86986
−50	−5.09467
−100	−5.04999
−200	−5.03025

TABLE 1. *Critical values for fitting arguments*

3. The Matricial Functional Fitting. The functional fitting, described in the previous chapter, has three deficiencies:

a. It is necessary to solve a linear algebraic system in every step, in order to compute the values of the parameters.

b. If we want to fit a scheme to a considerable number of functions, we must perform an unnecessarily large number of function evaluations.

c. The selection of the functions to be fitted is not natural, and we must introduce additional assumptions in order to define the selection properly.

Here we follow the well-known mathematical rule: if you want to remove deficiencies you must introduce other deficiencies. We define a scheme which can be applied to a narrower set of problems and which needs more computation. On the other hand the selection of parameters is quite natural; we need less subdivisions of the steps and the scheme is *MA*-stable:

Let $\mathbf{x}_{n+1}^{(1)}$ and $\mathbf{x}_{n+1}^{(2)}$ be defined as in the previous chapter. We define

$$\mathbf{x}_{n+1} = P\mathbf{x}_{n+1}^{(1)} + (I - P)\mathbf{x}_{n+1}^{(2)}, \tag{3-1}$$

when P is an N by N matrix, $x \in E^N$.

Let be the N matrix equation

$$\dot{Z} = F(t, Z), \quad Z(t_n) = I, \tag{3-2}$$

whose solution is known.

LEMMA 6. *Let $Z_{n+1}^{(1)}$ and $Z_{n+1}^{(2)}$ be the solutions of* (3-2) *by one step and two half-steps, respectively, of the trapezoidal rule, and let $Z_{n+1}^{(1)} - Z_{n+1}^{(2)}$ be nonsingular and*

$$P = (Z_{n+1} - Z_{n+1}^{(2)})(Z_{n+1}^{(1)} - Z_{n+1}^{(2)})^{-1}. \tag{3-3}$$

Then the scheme

$$X_{n+1} = PX_{n+1}^{(1)} + (I - P)X_{n+1}^{(2)}$$

is at least of second order and is fitted to (3-2).

Proof. Similar to the proof of Theorem 1B.

The most natural choice of (3-2) is

$$\dot{Z} = AZ, \quad Z(t_n) = I, \tag{3-4}$$

when A is a negative-definite matrix. In this case

$$E = Z_{n+1}^{(1)} - Z_{n+1}^{(2)} = \frac{-h^3A^3/16}{(I + hA/2)(I + hA/4)^2}.$$

If $\{\lambda_i\}_{i=1}^N$ are the eigenvalues of hA, $\lambda_i < 0$, $\forall i$, and $\{\mathbf{v}_i\}_{i=1}^N$ the corresponding eigenvectors, then it is trivial to show that $\{\mathbf{v}_i\}_{i=1}^N$ are also the eigenvectors of E, with the eigenvalues

$$\mu_i = \frac{-\lambda_i^3/16}{(1 + \lambda_i/2)(1 + \lambda_i/4)^2} \neq 0 \qquad \forall 1 \leqslant i \leqslant N,$$

and the conditions of Lemma 6 hold.

The solution of (3-4) is known to be $e^{(t-t_n)A}$. In order to compute an expo-

nential of a matrix we must obtain its eigenvalues and eigenvectors, which is an extremely laborious task if performed at every step. It seems reasonable to reduce the amount of computation and introduce certain rational approximations to the exponential of the matrix. In the following we will apply the approximation

$$e^{hA} \simeq R(hA) = \frac{I + k_1 hA + k_2 h^2 A^2}{I + k_3 hA + k_4 h^2 A^2}. \tag{3-5}$$

If we substitute (3-5) into the formula (3-3), we obtain

$$P = 16\,\frac{\left(I - \frac{1}{2}hA\right)\left((k_1 - k_3 - 1)I - \left(\frac{1}{2}k_1 - k_2 + \frac{1}{2}k_3 + k_4\right)hA + \left(\frac{1}{16}k_1 - \frac{1}{2}k_2 - \frac{1}{16}k_3 - \frac{1}{2}k_4\right)h^2A^2 + \frac{1}{16}(k_2 - k_4)h^3A^3\right)}{h^2A^2(I + k_3 hA + k_4 h^2A^2)}.$$

The matrix A had been defined to be nonsingular, but it can be ill-conditioned and we must take careful precautions in order to avoid any trouble in its numerical inversion. If we equate

$$k_1 = 1 + k_3, \quad k_2 = k_3 + k_4 + \tfrac{1}{2},$$

then the lower order terms in the numerator of P vanish, and we obtain

$$P = -\frac{(I - \tfrac{1}{2}hA)\,((3 + 8k_3 + 16k_4)I - (k_3 + \tfrac{1}{2})hA)}{I + k_3 hA + k_4 h^2 A^2}$$

and

$$R(hA) = \frac{I + (1 + k_3)hA + (k_3 + k_4 + \tfrac{1}{2})h^2A^2}{I + k_3 hA + k_4 h^2 A^2}.$$

We can proceed here in three distinct ways:

a. To substitute $k_3 = -1/2$, $k_4 = 1/12$, and then to obtain the two-by-two diagonal Padé approximation

$$R(hA) = \frac{I + hA/2 + h^2A^2/12}{I - hA/2 + h^2A^2/12}$$

and

$$P = -\frac{1}{3}\,\frac{I - hA/2}{I - hA/2 + h^2A^2/12}.$$

The Padé approximation is simply a Hermite-type rational interpolation at the origin; and therefore, it is extremely accurate for the components of a solution forced by the close-to-zero eigenvalues of the linear system. Consequently, this approximation is particularly good in the "smooth" segments of the solution, in which the contribution of the large (in absolute value) eigenvalues, so-called parasitic roots, is negligible, in other words outside the boundary layers.

b. To interpolate the expression (3-5) for two particular real and negative values λ_1 and λ_2. That is,

$$k_3 = \frac{\dfrac{e^{\lambda_2} - 1 - \lambda_2 - \frac{1}{2}\lambda_2^2}{\lambda_2^2}(e^{\lambda_1} - 1) - \dfrac{e^{\lambda_1} - 1 - \lambda_1 - \frac{1}{2}\lambda_1^2}{\lambda_1^2}(e^{\lambda_2} - 1)}{\dfrac{e^{\lambda_1} - 1 - \lambda_1}{\lambda_1}(e^{\lambda_2} - 1) - \dfrac{e^{\lambda_2} - 1 - \lambda_2}{\lambda_2}(e^{\lambda_1} - 1)},$$

$$k_4 = \frac{\dfrac{e^{\lambda_2} - 1 - \lambda_2 - \frac{1}{2}\lambda_2^2}{\lambda_2^2} \times \dfrac{e^{\lambda_1} - 1 - \lambda_1}{\lambda_1} - \dfrac{e^{\lambda_1} - 1 - \lambda_1 - \frac{1}{2}\lambda_1^2}{\lambda_1^2} \times \dfrac{e^{\lambda_2} - 1 - \lambda_2}{\lambda_2}}{\dfrac{e^{\lambda_1} - 1 - \lambda_1}{\lambda_1}(e^{\lambda_2} - 1) - \dfrac{e^{\lambda_2} - 1 - \lambda_2}{\lambda_2}(e^{\lambda_1} - 1)}.$$

As a rule, this gives us a better approximation to the exponential curve for negative arguments and this approach can be considerably better than the previous one for "transient" segments, i.e. the boundary layers. This approach is particularly useful if we have any *a priori* knowledge about the loci of the eigenvalues, especially if they are located in two clusters.

c. To find universal coefficients k_3 and k_4, which give the "best" approximation to the exponential curve in $(-\infty, 0]$ for any suitable norm. The natural choice is the integral L_2 norm with a weight function, giving greater weight to closer-to-the-origin arguments. One particular norm is

$$\|f\| = \left\{\int_{-\infty}^{0} e^t f^2(t)dt\right\}^{1/2}.$$

Analytical computation with this norm, i.e., finding $\min_{k_3, k_4}\|e^t - R(t, k_3, k_4)\|$, involves calculus with exponential integrals and is not practical. On the other hand, the numerical computation is simple, using Gauss-Laguerre integration. This approach can be useful for "transient" segments, when no information about the loci of the eigenvalues is available.

When we actually solve a system $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$, we substitute in the formula for P the value of hJ, where J is the Jacobian matrix in t_n and apply (3-1).

It is possible to include in (3-1) larger combinations of $x^{(i)}$'s and to fit the scheme to more matricial equations. In order to perform this, the solutions of the matricial equations must commute.

Department of Mathematics
Ben-Gurion University
Beer-Sheva, Israel

1. G. G. DAHLQUIST, "A special stability problem for linear multistep methods," *BIT*, v. 3, 1963, pp. 27–43. MR 30 #715.

2. W. LINIGER & R. A. WILLOUGHBY, *Efficient Numerical Integration of Stiff Systems of Ordinary Differential Equations*, IBM Res. Report RL-1970, 1967.

3. A. RALSTON, *A First Course in Numerical Analysis*, McGraw-Hill, New York, 1965. MR 32 #8479.

4. A. ISERLES, *Nonexponential Fitting Techniques for Stiff O.D.E.*, Math. Tech. Report 154, Ben-Gurion University, 1976.