Some stability techniques for multistep methods

by F. Odeh

A partial survey is given of sundry results in the stability theory of multistep formulae when these are used to integrate time-varying or nonlinear stiff systems, with emphasis on systems arising in circuit analysis.

1. Introduction

This paper is a short, informal, and quite partial survey of a few of the results in the stability of difference methods obtained by various people in or closely associated with the Mathematical Sciences Department at IBM in Yorktown. These results are mainly concerned with the stability of methods obtained by applying linear multistep formulae (LMF) to the numerical integration of initial value problems for systems of ordinary differential equations which may be characterized as being *stiff*. Following what is by now a time-honored tradition, no precise definition of this class of equations will be given. Loosely speaking, it usually denotes a system which is stable in some sense but exhibits a wide difference in the behavior of its individual solutions. For example, in linear systems, $\dot{x} = Ax$, this happens when the time constants, i.e., the reciprocals of the eigenvalues of A, of the system are very widely different. Such behavior is frequently encountered in the analysis of simple linear circuits, where the different sizes of the electrical components, e.g., large resistances and moderate capacitances and inductances, would lead to many orders of magnitude differences in the circuits' time-constants.

The classical theory of numerical integration methods for systems, $\dot{x} = f(t, x)$, is quite well established and is described in, for example, the textbooks of Henrici [1] and Stetter [2]. It is based on assuming that one chooses time-steps, h, so that the norm of hJ, where J is the Jacobian of f, is rather small. There are basically two reasons why the classical theory is inappropriate for stiff systems. First, the time-step has to be small with respect to the reciprocals of the *large* eigenvalues of J. In a standard RLC circuit with the typical values of $R = 1000\ \Omega$, $L = 1$ nH, $C = 1$ pF, this would restrict the time-step to be of the order of 10^{-3} ns even though the time variations of the current and voltage—apart from a short initial phase—are of the order of 1 ns. Second, the bounds on the numerical error provided by the classical theory are of the order of exp (LT), where $(0, T)$ is the relevant time of numerical integration and L is the Lipschitz constant of f in some domain surrounding the solution and is therefore always quite large for a stiff system. One of the objectives of the results which are highlighted below is to derive methods and stability and error-estimation techniques which are more or less free of the above difficulties. This, of course, can be accomplished only by restricting the class of differential systems, e.g., to those satisfying one-sided Lipschitz conditions such as negative monotone (dissipative) systems and their relatives. This class seems adequate to describe most of the systems arising in circuit applications.

We now outline the paper's plan. In Section 2, some preliminaries are given about multistep methods, their asymptotic stability regions, and Dahlquist's famous restriction on the accuracy of A-stable methods, i.e., those whose stability region includes the left-half complex plane. Section 3 is concerned with the fixed-h stability of A-stable and related methods when applied to various classes of nonlinear monotone systems. Here, we introduce the multiplier technique, related in spirit to the corresponding technique in control theory, for the study of global error estimation of multistep methods. The behavior, in the nonlinear regime, of methods based on $A(\alpha)$-stable formulae is also described. In Section 4, two Liapunov-like techniques which are appropriate for studying, or minimizing, the effects of time variations in the system and/or the size of the time-step are given. One is a method for constructing *polyhedral norms* for studying the stability of such systems, and the other is a method for constructing *methods* which are quite robust under time variations. Section 5, which is concerned with applications to circuit analysis, deals with the behavior of the global Picard-like iteration method known as Waveform Relaxation and the convergence of its discretized version.

2. Preliminaries

A linear multistep formula (LMF) is given by

$$Lx(t) \equiv \sum^{k} \alpha_j x_{n+j} - h \sum^{k} \beta_j \dot{x}_{n+j} = 0, \tag{1}$$

where h is the time-step, k is the step number, and the dot denotes differentiation with respect to time. If E denotes translation by h and the usual polynomials (ρ, σ) are defined by

$$\rho(\zeta) = \sum \alpha_j \zeta^j, \; \sigma(\zeta) = \sum \beta_j \zeta^j, \tag{2}$$

then (1) has the form

$$Lx(t) \equiv \left[\rho(E) - h\sigma(E)\frac{\partial}{\partial t}\right] x(t) = 0. \tag{3}$$

The order of accuracy, p, of L is defined by applying L to smooth functions $\psi(t)$ and expanding in powers of h to obtain

$$L\psi(t) = 0(h^{p+1}) \sim \text{local truncation error.} \tag{4}$$

Combining (3) with a smooth system

$$\dot{x} = f(t, x), \tag{5}$$

one obtains the algebraic system

$$N(x_n) \equiv \rho(E)x_n - h\sigma(E)f(t_n, x_n) = 0 \tag{6}$$

for the numerical solution x_n which approximates the exact solution $x(t = nh) \equiv x(n)$. Since the exact solution is assumed smooth, it satisfies

$$N(x(n)) \sim 0(h^{p+1}). \tag{7}$$

It is clear that if the local errors do not get unduly amplified, i.e., the method is stable, the global error $e(n) = x \; - x(n)$ should converge to zero like h^p as $h \to 0$. This is in fact the case if $\| hf_x \|$ is small, as shown by Dahlquist when the method is small-h stable, which may be characterized algebraically by the condition

$$\rho(\zeta_i) = 0 \Rightarrow |\zeta_i| \le 1, \quad \text{and the roots on the unit circle are simple.} \tag{8}$$

It is noteworthy that even this (relatively weak) stability requirement restricts the order of accuracy to about half what one would expect from counting the degrees of freedom in (1), i.e., to $p \sim k$ instead of $p \sim 2k$.

Since for stiff systems one does not wish to so severely restrict h to render $\| hf_x \|$ small, another stability concept, introduced again by Dahlquist [3], is more appropriate. Consider the linear model problem

$$\dot{x} = \lambda x, \tag{9}$$

where λ is a complex constant. Then the region of (absolute) stability, S, of a numerical method is the set of $h\lambda$ such that all the numerical solutions of the model problem with fixed step h will remain bounded when $n \to \infty$. Since the model problem is stable for λ in the left-half plane C^-, it is desirable that S contain as much of C^- as possible. A method is called A-stable if S contains C^-; a simple and widely used example is the trapezoidal rule. A-stability, which seems a natural, albeit rather extreme, requirement, severely restricts multistep methods (1) to be *implicit* $\beta_k \neq 0$ and to have low order, as the following barrier result shows:

Barrier: The order of accuracy

of an A-stable LMF (1) cannot exceed two. (10)

The original proof of this pretty result, given in [3], depends on two small calculations which algebraically quantify the accuracy and stability requirements. First apply L to $\psi(t) = e^t$ as in (4) to obtain

$$\frac{\rho(\zeta)}{\sigma(\zeta)} - \log \zeta \sim c(\zeta - 1)^{p+1} \text{ near } \zeta = 1.$$

For convenience, introduce the Greek-Roman transformation

$$\zeta = (z + 1)(z - 1)^{-1},$$

then

$$\frac{\rho}{\sigma}(\zeta) \to \frac{r}{s}(z), \tag{11}$$

and the accuracy requirement reads, near $z = \infty$,

$$\frac{r}{s} \sim \log \frac{z + 1}{z - 1} - c\left(\frac{2}{z}\right)^{p+1},$$

and hence, for $p > 2$,

$$z\frac{r}{s} \sim 2 + \frac{2}{3} z^{-3} + \cdots. \tag{12}$$

On the other hand, A-stability requires that r/s be regular and have positive real part for $z \in C^+$, and therefore can be represented, by the Riesz-Herglotz theorem, as the transform of a nonnegative measure

$$\frac{r}{s}(z) = \int \frac{1}{z - it}\, d\omega(t).$$

Hence

$$x\frac{r}{s}(x) = \int \frac{x^2}{x^2 + t^2}\, d\omega(t). \tag{13}$$

The behaviors, as $x \to \infty$, of the right sides of (12), (13) contradict each other since the integrand in (13) is nondecreasing in x for every fixed t. A more natural proof—a counting argument—was given much later using the order-stars theory [4].

Naturally, there have been lively activities in circumventing the above barrier. One approach is to relax the stability requirement so that S contains a reasonable part, but not all, of C^-. For example, in $A(\alpha)$-stable methods, S contains a wedge of angle α around the negative x-axis, and backward differentiation formulae (BDF) are a prime example thereof. Other approaches use higher-order derivatives in (1), or postprocessing of solutions obtained by A-stable methods [5], or nonlinear methods of integration, but such approaches have not been as popular, at least in the U.S., as the simple LMF.

3. Nonlinear stability and multipliers

The asymptotic stability concept described in Section 2 via the model problem (9) is adequate for linear systems $\dot{x} = Ax$ where A is a constant matrix. In that case, by the spectral theorem, the numerical solutions (and errors) corresponding to (6) decay (or stay bounded) if the eigenvalues of A lie within the stability region S. Intuitively, one suspects that for "reasonable" nonlinear systems, the numerical solution will be stable if the spectrum of the Jacobian f_x is always inside S. The reason is that the error would satisfy a more-or-less-linear time-varying variational equation, and hence the error growth could be controlled. This approach involves considerable care, since it is easy to obtain exponentially increasing solutions for systems which are quite stable when the coefficients are frozen, e.g, in the damped Mathieu equation. Liapunov-like methods (G-stability) were used by several authors to obtain global error estimates, but these methods are most applicable to A-stable, hence only low-order, methods. In this section we briefly describe [6] an approach for *directly* estimating such errors which also works for high-order methods.

From (6), (7) the global error satisfies

$$\rho(E)e_n + h\sigma(E)F_n = hp_{n+k}, \tag{14}$$

where $F_n = f(nh, x(nh) + e_n) - f(nh, x(nh))$, x is the exact solution, and where we have replaced f by $-f$ for notational convenience. Under mild restrictions on (ρ, σ) this may be written as

$$\gamma * e_n + hF_n = hq_n, \tag{15}$$

where γ is the ℓ_1-sequence defined by $\rho\sigma^{-1}(\zeta) = \Sigma\gamma_j\zeta^{-j}$ and $*$ denotes convolution. If the nonlinearity f is monotone,

$$\langle f(x) - f(y), x - y\rangle \geq \mu|x - y|^2, \qquad \mu > 0, \tag{16}$$

or satisfies similar circle conditions, then a useful device for studying the stability of A-stable methods is to scalar multiply (15) by e_n and use the monotonicity (16) together with the A-stability—which implies the positivity of the quadratic form $\Sigma\langle e, \gamma * e\rangle$—to obtain error bounds. For $A(\alpha)$-stable methods,

$$\alpha < \frac{\pi}{2},$$

this quadratic form could become negative, since the root-locus, i.e., the image of the unit circle under $\rho\sigma^{-1}$, intersects the left-half plane. One, however, can obtain *weighted* equalities by scalar multiplying by $\mu * e_n$, where μ is an ℓ_1 sequence, to obtain

$$\sum\langle \mu * e_n, \gamma * e_n\rangle + \sum\langle \mu * e_n, F_n\rangle = \sum\langle \mu * e_n, q_n\rangle. \tag{17}$$

If $\hat{\mu}(\tau) = \Sigma\mu_n e^{-in\tau}$, then μ is called a *multiplier* for (ρ, σ) if $\Sigma\mu_j\zeta^{-j}$ is rational, Re $\hat{\mu}(\tau) > 0$, and

$$\mathrm{Re}\,\{\bar{\hat{\mu}}(\tau)\rho\sigma^{-1}(e^{i\tau})\} \geq 0 \qquad \text{for all } \tau. \tag{18}$$

Then, by Parseval, the first term on the left side of (17) is nonnegative. If the method does possess a multiplier which is in some sense simple, then, under some monotonicity conditions stronger than (16), one can show that the second term is positive, and (17) could be thought of as a weighted energy inequality which may be used to obtain global bounds on e_n. The analysis naturally separates into three parts: the relation between methods and "their" multipliers, the relation between such multipliers and the nonlinearities, and finally the error behavior; we say a few words about each part. First, one can show the existence of a multiplier for any $A(\alpha)$-stable method. More precisely, one has Theorem 1.

- *Theorem 1*

If a method (ρ, σ) is $A(\beta)$-stable, then, for any $0 < \alpha < \beta$, there exists a multiplier of finite support, $\mu = \{\mu_j\}_0^M$, for (ρ, σ) such that

$$|\arg \hat{\mu}(\tau)| < \frac{\pi}{2} - \alpha. \tag{19}$$

Since it is easy to see that (19) implies $A(\alpha)$-stability, Theorem 1 may be interpreted as saying that the linear stability of an $A(\alpha)$-stable method can be "seen" through the multiplier. This multiplier is produced by judiciously modifying and truncating a fractional power of $\rho\sigma^{-1}(e^{i\tau})$

which is chosen so that (18) holds. Thus, such a multiplier would in general be a complicated sequence and seems to be useful for investigating the stability of linear problems only. For in such cases, assuming $f = A$ with $|\arg$ spectrum of $A|$ $< \alpha$, one can find a scalar product such that $|\arg \langle v, Av\rangle|$ $< \alpha$. The positivity of the second term of (17) then follows from (19). To treat nonlinear problems, it turns out that multipliers should have additional properties in the time domain, e.g.,

1. $\mu_j \le 0 \qquad$ for $j > 0$,

and

2. $\mu_0 \ge \omega \sum_{j=1} |\mu_j| \qquad$ with a large $\omega \ge 1$. (20)

Thus the best of all multipliers would be $\mu = \{1, -\eta\}$ with very small $\eta > 0$. This is the analogue of the Popov multiplier in control theory, and a method with such a simple multiplier enjoys very good stability properties in the nonlinear regime. A graphical method can be devised to check whether (ρ, σ) has such multipliers μ, and one finds that the BDF of orders two to five have such μ with $\eta = 0$, 0.0836, 0.2878, 0.8160, while the BDF of order 6 needs a more complicated μ to exhibit its linear stability.

Another use of this multiplier language is to derive a quantitative form of the barrier result (10) in the form of an *uncertainty principle* showing the incompatibility of extreme stability and accuracy of an LMF. If L_h denotes the linear functional (1), then the *Peano-kernel* of the method (ρ, σ) is given by

$$K_q(s) = L_1 \frac{s_+^{q-1}}{(q-1)!} \quad 2 \le q \le p + 1,$$

the local truncation error may be written as

$$L_h x(t) = h^q \int_{-k}^{0} K_q(s) x^{(q)}(t - hs)ds, \tag{21}$$

and one may measure this error by the size of $\|\hat{K}_q\|$; for p accurate methods choose $q = p + 1$. The departure from A-stability may be measured by the size of η, and one has Theorem 2.

- *Theorem 2*

For every k, there exists $C_k > 0$ such that, if $\{1, -\eta\}$ is a multiplier for the k-step method (ρ, σ), then, as $\eta \to 0$,

$$\|\hat{K}\| \ge C_k \eta^{2-p}, \tag{22}$$

which shows the blowup of the error if $p > 2$ and $\eta \to 0$. The main ideas in proving this result are that 1) $\|K\|$ is as large as the largest a_j where $\rho(\zeta) \sim r(z) \equiv \Sigma a_j z^j$; this is proved by soft arguments. Then, some detailed function theory arguments show that 2) a large amount of stability (η small) forces a_j to grow so that

$$a_{p-1} \ge \left(\frac{1 - \eta}{6} \eta\right)^{p-2} a_1.$$

The interaction between the multiplier and the nonlinearity is given by the second term in (17), and one wishes this term to be positive. This requirement is a sort of *generalized correlation inequality* and can be shown to hold under a variety of restrictions on the nonlinearity and on the sign/size of μ. Some examples follow.

- *Example 3*

Let $\varphi = \varphi(v)$ be convex and nonnegative and $\varphi(0) = 0$. Assume $\mu_j \le 0$ for $j > 0$ and let P_μ be the sequence of partial sums of μ. Then

$$\sum \langle \mu * v_n, \text{grad } \varphi_n\rangle \ge (P\mu * \varphi)_N, \tag{23}$$

and hence the sum is nonnegative if $\mu_0 \ge -\Sigma\mu_j$.

- *Example 4*

If f is σ-angle bounded, i.e., satisfies

$$\langle f(x) - f(y), y - z\rangle \le \sigma\langle f(x) - f(z), x - z\rangle,$$

and $f(0) = 0$, then

$$\sum \langle \mu * v_n, f_n\rangle \ge 0 \tag{24}$$

if $\mu_j \le 0$ for $j > 0$ and $\mu_0 \ge (1 + \sigma)\Sigma\mu_j$. Such functions are basically gradient-like monotone functions, e.g., 3-cyclic monotones which satisfy

$$\sum_{i=1}^{3} \langle x_i - x_{i-1}, f(x_i)\rangle \ge 0, \qquad x_0 = x_3.$$

A more interesting case relates the multiplier to the asymmetry and the variability of the problem. If one measures these two effects by K_1, K_2 defined by

$$\langle u, Jv\rangle \le \frac{K_1}{2} \{\langle u, Ju\rangle + \langle v, Jv\rangle\} \tag{25}$$

and

$$\langle u, J(x)u\rangle \le K_2 \langle u, J(y)u\rangle, \tag{26}$$

then one has Example 5.

- *Example 5*

The nonlinear energy term in (17) is nonnegative if

$$K_1\left(\frac{1 + K_2}{2}\right) \sum_{j=1} |\mu_j| \le \mu_0. \tag{27}$$

In studying the convergence behavior of the numerical method one needs (26) for $|x - y| = 0(h)$ and thus $K_2 \sim 1 + 0(h)$. For example, if $\mu = (1, -\eta)$, then (27) reduces to $K_1\eta \lesssim 1$, which shows that as the asymmetry of the problem increases one has to decrease η and thus use more stable methods. A strategy for changing the order of the integration method to keep the asymmetry (just) under control can thus be devised which together with local error control would ensure the convergence of the numerical solution obtained by high-order methods; see [6] for details.

Finally, error bounds are obtained by combining the above positivity results with the G-stability theory, which says that the behavior of A-stable methods can be seen via the construction of a Liapunov function, which is a quadratic with a positive matrix, G, constructed algebraically by a sort of continued-fraction expansion from the method (ρ, σ) [7]. Consider now the error equation

$$\rho x_n + h\sigma F_n = hp_{n+k}. \tag{28}$$

Operate on (28) by σ^{-1} and scalar-multiply by $\pi\nu^{-1}x_n$, where the multiplier μ, being rational, may be written as $\pi\nu^{-1}$ where π, ν are polynomials of degree ℓ. Then one gets the energy equality

$$\langle \nu^{-1}\pi x_n, \sigma^{-1}\rho x_n\rangle + h\langle \mu * x_n, F_n\rangle = h\langle v * x_n, q_n\rangle. \tag{29}$$

When the variable $y_n = \sigma^{-1}\nu^{-1}x_{n+1}$ is introduced, the first term in (29) becomes $\langle \pi\sigma y_{n-1}, \nu\rho y_{n-1}\rangle$. Noting that, by (18), the "method" $(\nu\rho, \pi\sigma)$ is A-stable, there exists [7] a definite quadratic form G such that

$$G(Y_n) - G(Y_{n-1}) \le 2\mathrm{Re}\,\langle \pi\sigma y_{n-1}, \nu\rho y_{n-1}\rangle, \tag{30}$$

where $Y_n = (y_n, \cdots, y_{n+k+\ell-1})$. Substituting in (29), summing, and using the positivity of the second term, one obtains

$$G(Y_n) \le 2h\sum\langle \mu * x_n, q_n\rangle. \tag{31}$$

Further manipulation of (31) then yields the following.

- *Theorem 6*

If μ is a multiplier for (ρ, σ) and f is such that $\Sigma\langle \mu * \nu_n, F_n\rangle$ is positive, then, for some C,

$$|x_n| < C\{\text{(initial data)} + h\sum|p_j|\}. \tag{32}$$

If the above positivity still holds with f replaced by $f - \alpha$ for some $\alpha > 0$, i.e., there is some dissipation in the problem, one obtains

$$\sup_{n\le M}|(\text{error})_n| \le \frac{C}{\alpha}\sup_{n\le M}|\text{local errors}|. \tag{33}$$

The above, as well as other estimation and convergence results given in [6], indicates that it is possible to develop, for typical stiff systems, a stability theory which is free from Lipschitz constant restrictions and is quite analogous to the classical one.

4. Variability and contractions

There are a variety of reasons for considering the effects of time variation on the stability of LMF. Among them is that the efficient numerical integration of a differential system dictates the use of variable time-steps. The strategy of such a change is usually controlled by local error considerations which leave the question of the stability of the numerical scheme in doubt. Two approaches to ensure stability are briefly described. One is a rather general method for *constructing*, when it exists, a (nonsmooth) Liapunov function for dynamical systems. The other is to *design* methods which are contractive, in a fixed norm, under arbitrary time variations of the model problem $\dot{x} = \lambda(t)x$ and of the time-step size.

- *Liapunov approach [8]*

Consider the variable LMF

$$L_{h,n}x \equiv \sum^{k}\alpha_{j,n}x_{n+j} - h_n\sum\beta_{j,n}\dot{x}_{n+j} = 0. \tag{34}$$

Applying (34) to $\dot{x} = \lambda x$, one obtains a recurrence relation of the form

$$z_{n+1} = M_n z_n, \tag{35}$$

where $z_n = (x_n, \cdots, x_{n+k-1})$.
To show the stability of (35), it is sufficient to construct a Liapunov function w, e.g., a quadratic or some other norm, such that

$$w(M_n z) \le w(z). \tag{36}$$

To construct w from M_n, it is useful to give a more generous definition of stability. Consider a set $\mathcal{A} = \{A, B, \cdots\}$ of square stable matrices and let $\mathcal{A}'$ be the semi-group generated by $\mathcal{A}$ (i.e., all finite products); then one defines $\mathcal{A}$ to be stable (at the origin 0) if for every neighborhood U of 0 there is a neighborhood V so that $MV \subset U$ for all $M \in \mathcal{A}'$. Then one can show that the stability of $\mathcal{A}$ is equivalent to the boundedness of $\mathcal{A}'$ or to the existence of a bounded balanced convex set W which is *invariant* under $\mathcal{A}'$. The norm sought after in (36) above then has W as the unit ball and the stability question reduces to that of constructing the invariant set. Before describing this constructing it may be noted that

1. The stability of every finite product of $\{M_i\}$ is *not* sufficient for the stability of $\mathcal{A}$. For example, suppose $\mathcal{A} = \{A, B\}$, where

   $$A = \begin{pmatrix} e^{i\theta} & a_1 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},$$

   where θ/π is irrational; then it is easy to see that every $M \in \mathcal{A}'$ has the form

   $$\begin{pmatrix} e^{i\psi} & \alpha \\ 0 & 1 \end{pmatrix},$$

   where ψ/π is irrational and hence M is stable. However, for the sequence

   $$A_k \equiv \begin{pmatrix} e^{i\theta_k} & a_k \\ 0 & 1 \end{pmatrix}$$

   defined by $A_1 = A$,

$$\begin{aligned} A_{k+1} &= A_k^2 \qquad \text{if } \operatorname{Re} e^{2i\theta_k} \ge 0, \\ &= A_k^2 B \qquad \text{if not,} \end{aligned}$$

one has $\operatorname{Re}(e^{i\theta_{k+1}}) \ge 0$. Then

$$\begin{aligned} |a_{k+1}| &= |1 + e^{i\theta_k}||a_k| \\ &\ge \sqrt{2}|a_k|, \end{aligned}$$

which shows the instability of $\mathcal{A}$.

2. The set W must have corners. For it is easy to see that if, for some $M \in \mathcal{A}$, there exists one eigenvalue λ on the unit circle and ξ is its eigenvector (normalized to belong to ∂W), then the plane π passing through ξ and parallel to the complementary subspace is a support plane of W at ξ. Hence, if there are two matrices M_i as above, with the same right eigenvector ξ but with two different left eigenvectors η_i, W must have at ξ two different support planes; i.e., it has corners. Candidates for the norm w are generalized max. norm

$$w(z) = \max_i |\textstyle\sum C_{ij} z_j|.$$

An iterative algorithm to determine the stability of a set of matrices, in virtually all cases, is based on the following construction.

- *Theorem 7*

Given a finite set $\mathcal{A} = \{M_0, M_1, \cdots, M_{m-1}\}$ of m distinct matrices, let W_0 be a bounded neighborhood of the origin and define

$$W_k = \mathcal{H}\left[\bigcup_i M^i_{k'} W_{k-1}\right], \qquad k' = (k-1) \bmod m, \tag{37}$$

where $\mathcal{H}$ denotes the *convex hull*. Then $\mathcal{A}$ is stable if $W = \cup\, W_k$ is bounded.

The above result is made constructive by choosing W_0 to be a *polyhedral* region; hence, by tracking extreme points, all the subsequent W_k are also polyhedral, and one has to generate extreme points of W_k and add these to the extreme points of W_{k-1}. The main computational step is, therefore, to check whether all the extreme points have been generated by applying $M = M_{k'}$ to the extremes of W_{k-1}, or, equivalently, to find whether at any stage of the construction, one more application of $M_{k'}$ takes one into the convex hull of the previous extreme points. This can be accomplished by using the (first phase of) the linear program: Maximize 0 (any constant) such that

$$x \equiv M_{k'} x_i = \textstyle\sum \lambda_j x_j,$$

$$\textstyle\sum \lambda_j = 1, \qquad \lambda_j \ge 0.$$

The construction was proved to be finite in many interesting cases [8].

In applying the above method to numerical methods one encounters an infinite set $\mathcal{A}$ of matrices, but it usually has a *finite* set of extreme points, and it is clear that the stability of the convex hull of $\mathcal{A}$ is equivalent to that of its extreme set. This has been used to show the stability of the backward differentiation method of order two if the step ratio does not exceed 1.2, and was accomplished by constructing an invariant set W with 76 vertices!

- *Constructive methods [9]*

Since the solutions to the model problem $\dot{x} = \lambda x$, $\operatorname{Re}\lambda \le 0$, decay in time, it seems reasonable to *design* LMF (1) so that the discrete solution enjoys a similar property. Formally, one says that a formula is contractive at $q = h\lambda$ if the solutions to

$$\rho(E)x_n - (h\lambda)\sigma(E)x_n = 0 \tag{38}$$

satisfy $|X_{n+1}| < |X_n|$, where $X_n = (x_n, \cdots, x_{n+k-1})$ and $|\;|$ denotes a chosen norm; the standard max. norm turns out most useful. Clearly contractivity at q implies stability there, and thus the contractivity region K, i.e., the set of q at which a formula is contractive, is contained in S. Because the concept is local, one obtains stability for contractive LMF when it is applied to variable step-size and variable model equation $\dot{x} = \lambda(t)x$, even though one tests with constant λ only. This holds only when the formula is implemented in a one-leg fashion, i.e., when σ and f in the algebraic system (6) are permuted so that one has to consider, for each n, only *one* $q = h_n\lambda(\Sigma\beta_{j,n}t_{n+j})$ which is assumed to belong to K. Various concepts of stability have their contractivity twins, but it is generally much easier to devise algebraic conditions sufficient for the latter. For example, (ρ, σ) is i) contractive at the origin if $\alpha_k > 0$ and $\alpha_j \le 0$ $j < k$; ii) contractive at infinity if $\beta_k > \Sigma|\beta_j|$. It is also easy to show that ∂K is smooth (except possibly at its intersection with the real axis) and that K is closed and, in contrast to S, is connected, in fact by arcs of circles. However, one should note that characterizing A-contractive methods (methods which are contractive for all $q \in C^-$) is quite delicate. For example, for two-step second-order methods there is only a one-parameter family of methods which connect the two extremes of the one- and two-step trapezoidal rule. They may be derived roughly as follows: The contractivity condition may be written as

$$F_1(q) \equiv \sum^{k-1} |a_j - q\beta_j| - |a_k - q\beta_k| \le 0. \tag{39}$$

The critical case occurs on the imaginary axis $q = iy$. Let $\eta = y^2$, substitute in (39) and expand near $\eta = 0$ to obtain

$$F(\eta) = -\frac{1}{2}\,\eta\left(\sum^{k}\frac{\beta_j^2}{\alpha_j}\right) + 0(\eta^2) \equiv -C + 0(\eta^2).$$

One can show that $C > 0$ contradicts accuracy, hence $C = 0$ is a necessary condition which is actually *achievable* by maximizing C under accuracy constraints. This local argument may be extended to show that $F_1(iy) \le 0$ by using the geometric-arithmetic means inequality. Hence, the three accuracy constraints and $C = 0$ restrict the class to a one-parameter set which may be given different convenient parameterizations.

Contractive methods enjoy a lot of robustness when applied to fast-varying systems, again because their stability is derived from local properties. For example, it is easy to show the stability of $A(0)$ contractive methods when applied to diagonal-like dissipative nonlinear systems

$$\dot{x} = A(t, x)x + b \tag{40}$$

by modification of arguments valid for scalar equations. Also, if the nonlinearity f is maximal accretive in some Barach space, and the method satisfies mild contractivity conditions together with

$$\frac{\alpha_j}{\alpha_k} \le \frac{\beta_j}{\beta_k} \le 0,$$

then one can bound global errors in terms of sums of local ones [9]. A-contractive methods have even more remarkable stability properties. Recall, from the previous section, that every A-stable method has a positive definite quadratic function exhibiting its stability; i.e., for (ρ, σ), there is a $G > 0$—see (30)—such that

$$G(Y_{n+1}) - G(Y_n) \le 2\mathrm{Re}\langle \sigma y_n, \rho y_n \rangle. \tag{41}$$

Since applying the one-leg LMF to $\dot{x} = f(x)$ results in an error equation,

$$\rho y_n = h[f(\sigma x_n + \sigma y_n) - f(\sigma x_n)]. \tag{42}$$

Then, for dissipative f,

$$\mathrm{Re}\langle \sigma y_n, \rho y_n \rangle \le 0,$$

and (41) shows the decay of Y in the G norm. However, for variable time-steps, the coefficients of (ρ, σ) vary with n, and so does G. It was shown in [10] via a constructive procedure that the only A-stable, $p = 2$, $k = 2$ methods which have a *fixed* G—thus immediately guaranteeing stability—consist of the A-contractive class. The constancy of G in fact defines a unique extension of the methods from the case of uniform to that of variable time-steps. An implementation of a specially selected A-contractive method has recently been incorporated into a code for the robust simulation of electromechanical systems.

5. Waveform relaxations for circuits

The stable *implicit* methods described in the last two sections reduce the simulation of a differential system to the solution of a nonlinear algebraic system (6). This solution is obtained in standard codes by combining Newton-like methods and sparse matrix "technology." When the system (6) is very large, as for example in digital circuits, the approach becomes quite expensive, requiring ~0.2 minute per device on a 3081, and is thus limited to $O(10^2)$ devices, a small number in a VLSI era. Decomposition of such large systems together with relaxation techniques offers obvious advantages in both storage and speed requirements. There are basically two approaches for carrying out such relaxations. One, at the "algebraic" level of solving (6) or some linearized version, is by blocking it into subcircuits and relaxing in some fashion. This is referred to as the *incremental* approach; see [11, 12]. The other, introduced in [13] and called waveform relaxation, WR, is simply to lift the procedure to a "function-space" level and solve for the values of the unknowns in a subcircuit for *all* (or most of) the relevant time before feeding these values as inputs for the other subcircuits. This procedure proved to be very efficient for *special* classes of metal-on-oxide (MOS) digital circuits. Intuitively this happens because of the loose coupling of MOS devices and because WR allows each subcircuit to be integrated at its own optimum speed, i.e., with time-steps dictated by that possibly quiescent circuit and not by distant active ones which for large portions of time have little effect on that particular circuit. To describe the WR iteration, assume that a large system was decomposed into m blocks. Then the governing equations have the form

$$\dot{y}_i = F_i(y_i, y_j, \dot{y}_j), \qquad i \le m. \tag{43}$$

Combining (43) with some chosen relaxation procedure leads to the iteration

$$\dot{x}^k = f(x^k, x^{k-1}, \dot{x}^{k-1}), \tag{44}$$

where x is the whole vector of unknown functions and f is obtained from F_i and the relaxation procedure. For example, for MOS circuits, (43) is given by

$$C(v)\dot{v} + f(v) = 0. \tag{45}$$

Node-by-node decomposition of (45), together with Gauss-Seidel relaxation, leads to the iteration

$$\begin{aligned} &\sum_{j=1}^{i} C_{ij}(\cdots, v_i^k, v_{i+1}^{k-1}, \cdots)\dot{v}_j^k + \sum_{j=i+1} C_{ij}(\cdots)\dot{v}_j \\ &\qquad + f_i(\cdots, v_i^k, v_{i+1}^{k-1}, \cdots) = 0. \end{aligned} \tag{46}$$

The discretized version of (44), at least for constant time-steps, formally reads

$$\frac{1}{h}\frac{\rho}{\sigma}\, x^k = f(x^k, x^{k-1}, \frac{1}{h}\frac{\rho}{\sigma}\, x^{k-1}). \tag{47}$$

Let J_1, J_2, J_3 be bounds on the Jacobians of f with respect to its arguments. Then the behavior of the discretized WR for Lipschitz systems and small h may be described by the following.

- *Proposition 8 [14]*

Assume that i) (ρ, σ) is consistent with the roots of σ and $\rho(\zeta - 1)^{-1}$ inside the unit disk and ii) J_1, J_2 are bounded and $J_3 < 1$. Then the iteration (47) converges, for small enough h, on finite time intervals.

An easy proof follows from considering (47) as an iteration on

$$y = \frac{1}{h}\frac{\rho}{\sigma}x.$$

Since x is basically an "integral" of y, it can be thought of, in a standard manner, as a small operator on y in an appropriate exponential norm. Then (47), because $J_3 < 1$, defines a convergent contraction in that norm.

There are a fair number of other results which show the robustness of WR under a variety of more practical conditions. For example, the stability of WR under small errors, or when combined with function-space Newton as well as a detailed proof of the convergence of the important case (46), were given in [15]. The effect of truncating a (MOS) large circuit, to minimize storage, and relaxing between a few neighboring circuits only was discussed in [14], where a bound was given for the error caused by this *truncation in space*. In recent, to-be-published results, [16], the convergence of Gauss-Seidel WR in the *uniform* norm for special monotone systems $\dot{x} = f(x)$ was shown; also, since the implementation of WR generally involves interpolation, the above convergence proposition was modified to take account of linear interpolation. Convergence results, in ℓ_2-norms, for WR which do not assume Lipschitz continuity of the system were given in [17]. Considered there is the model iteration

$$\dot{x}^{n+1} + \partial\varphi(x^{n+1}) = g(\dot{x}^n), \tag{48}$$

where $\partial\varphi$ is the subdifferential of a convex function in some real Hilbert space H. The function g was assumed Lipschitz with a small constant <1 (this corresponds to small capacitive feedback in MOS); then boundedness estimates for (48), as well as its one-leg discretization with, e.g., backward differentiation methods of order up to 5, were given. These estimates are independent of the time-steps and rely heavily on the work of Section 3 above. For the iteration

$$\frac{1}{h}\rho x^{n+1} + \partial\varphi(\sigma x^{n+1}) = g\left(\frac{1}{h}\rho x^n\right), \tag{49}$$

the strong convergence for fixed time and the weak convergence in $\ell_2(H)$ of the iterates x^n to the unique fixed point of (49) were proved. When the nonlinearity f is not a gradient but is a Lipschitz perturbation on a linear operator, uniform convergence was shown on "time-windows" whose size depends only on the Lipschitz part.

Finally we remark that WR is quite suitable for being implemented on *parallel* machines, especially for digital circuits. This is because large digital circuits tend to be *wide*; that is, rather than being like one loop chain which has to be simulated serially, they are like many parallel chains with some interaction between them. The amount of parallelism can be improved by *timepoint-pipelining*, where, once a first timepoint is generated by the first processor, a second processor could begin computing the first timepoint for a second subcircuit, while the first processor computes the second timepoint for the first subcircuit. An implementation of WR on a nine-processor configuration with shared memory, and on a variety of increasingly larger problems, was carried out [18], with the moral that *parallelism scales with size*, so that one can, in WR, *effectively* use more and more processors as the problem size increases.

References

1. P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley & Sons, Inc., New York, 1962.
2. H. Stetter, *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag, New York, 1973.
3. G. Dahlquist, "A Special Stability Problem for Linear Multistep Methods," *BIT* **3**, 27–43 (1963).
4. G. Wanner, E. Hairer, and S. Nörsett, "Order Stars and Stability Theorems," *BIT* **18**, 475–489 (1978).
5. W. Liniger and F. Odeh, "*A*-Stable Accurate Averaging of Multistep Methods for Stiff Differential Equations," *IBM J. Res. Develop.* **16**, 335–348 (1972).
6. O. Nevanlinna and F. Odeh, "Multiplier Techniques for Linear Multistep Methods," *Numer. Fund. Anal. & Opt.* **3**, 377–423 (1981).
7. G. Dahlquist, "*G*-Stability Is Equivalent to *A*-Stability," *BIT* **18**, 384–401 (1978).
8. R. Brayton and C. Tong, "Stability of Dynamic Systems: A Constructive Approach," *IEEE Trans. Circ. & Syst.* **26**, 224–234 (1979).
9. O. Nevanlinna and W. Liniger, "Contractive Methods for Stiff Ordinary Differential Equations, Parts I, II," *BIT* **18, 19,** 457–474 and 53–72 (1978).
10. G. Dahlquist, O. Nevanlinna, and W. Liniger, "Stability of Two-Step Methods for Variable Integration Steps," *SIAM J. Numer. Anal.* **20**, 1071–1085 (1983).
11. F. Odeh and D. Zein, "A Semi-Direct Method for Modular Circuits," *Proceedings, IEEE International Symposium on Circuits and Systems*, 1983, pp. 226–229.
12. G. de Micheli, H. Y. Hsieh, and I. Hajj, "Decomposition Techniques for Large Scale Circuit Simulations," Ch. 7 in *Circuit Analysis, Simulation and Design*, A. Ruehli, Ed., North-Holland Publishing Co. (to be published).
13. E. Lelarasmee, A. E. Ruehli, and A. L. Sangiovanni-Vincentelli, "The Wave-Form Relaxation Method for Time-Domain Analysis of Large Scale Integrated Circuits," *IEEE Trans. Computer-Aided Design* **CAD-1**, 131–145 (1982).
14. F. Odeh, A. Ruehli, and C. Carlin, "Robustness Aspects of an Adaptive Waveform Relaxation Scheme," *IEEE International Conference on Computer Design: VLSI in Computers, ICCD 83*, 1983, pp. 396–399.
15. J. White, A. L. Sangiovanni-Vincentelli, F. Odeh, and A. Ruehli, "Waveform Relaxation: Theory and Practice," *Trans. Soc. Comp. Sim.* **2**, 95–133 (1985).
16. J. White and F. Odeh, "Some Uniform Convergence Results for Waveform Relaxations," *Int. J. Num. Meth. in Eng.*, to be published.
17. O. Nevanlinna and F. Odeh, "Remarks on the Convergence of the Waveform Relaxation Method," *Num. Funct. Anal. Appl.*, to be published.
18. J. White, "MOS Digital Circuit Simulation Using Waveform Relaxation and Its Application to Parallel Processors," Ph.D. thesis, Department of Electrical Engineering and Computer Science, University of California, Berkeley, 1986.

Received September 3, 1986; accepted for publication November 6, 1986

F. M. Odeh *IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.* Dr. Odeh received a degree from the University of California, Berkeley, in 1961, joining IBM Research to work on mathematical aspects of superconductivity. Since then, he has worked in the areas of foundations of scattering and band theory, bifurcation theory, and numerical stability for stiff equations; he is currently interested in circuit and device analysis. Dr. Odeh was an associate professor at the American University of Beirut, Lebanon, in 1968 and a visiting member of the Courant Institute, New York, New York, in 1963 and 1976. He is currently manager of the differential equations group in the Mathematical Sciences Department at the IBM Thomas J. Watson Research Center.