James W. Daniel†

A Partial Error Analysis for the Solution of Differential Equations in Simulation: A Look at Fowler's z-Transform Root-Locus Method*

Abstract: A partial analysis is made of the types of error of Fowler's method, which uses the z-transform procedure for digital simulation of complex systems.

I. Introduction

In a series of recent papers[1-5] M. E. Fowler and others at IBM have developed a new technique for digital simulation of complex systems, based on root-locus and z-transform analysis; the present paper presents a partial analysis of the nature of the error inherent in this new approach and in others.

The complex simulation problems which often arise, say, in the aerospace industry have some distinctive characteristics. For one thing, the continuous system is often stable and, if unstable, the only important requirement on the numerical solution is that it indicate this instability; for another, accuracy requirements may vary widely, ranging from one percent during transients for angular motion to 0.001% for velocity and position, while the requirement of short computation time remains constant. Fowler's method was designed primarily to allow the use of a large time step in those parts of a simulation problem which require only moderate accuracy during transients, but somewhat greater precision in the steady-state values. Although practical experience indicates the success of the method,[1] no proofs have been given for even simple cases.

One reason for the lack of a theoretical analysis of the error for Fowler's method is that the method is to some extent an "engineering" technique rather than a uniquely defined "mathematical" one, in the sense that for large systems of any interest the user makes many qualitative judgments, such as whether or not the relevant root-loci will "match," a judgment which can not be expressed by formulae. Also, since the method is designed for use with a large time step, it is not clear that an analysis of what happens when that time step converges to zero is at all relevant. Nonetheless, for stable systems and in cases in which the root matching technique can be well defined, it is possible to analyze the resulting error; because the ideas involved are not restricted to Fowler's method alone, we first present the analysis in a general form and proceed with the application to Fowler's method in Section IV.

Many standard methods for solving differential equations use iteration formulae based on the classical theories of interpolation, numerical differentiation and quadrature, or Taylor series; for the purpose of analysis, such methods are convenient because of the availability of information, e.g., error estimates, concerning the underlying approach. Fowler's method and some other simulation methods apparently have no such convenient existing theoretical basis; it seems necessary and appropriate to analyze them primarily from the viewpoint from which they often arise, i.e., root-locus matching. To this end we choose a "backward" error analysis in the sense that the numerical solution x_n is considered to equal $x_n(nT)$ where $x_n(t)$ is the exact solution to a differential equation that is a slight perturbation of the one we seek to solve. The numerical error $|x_n - y(nT)|$ can then be measured in terms of these perturbations, which are in turn measured by certain parameters of the solution method, in particular by the accuracy of the root-locus approximation.

II. Linear equations

We consider the system of N equations in N unknowns:

$$\dot{y} + a(t)y = c(t), \qquad y(0) = y_0,\ a(t) \text{ and } c(t) \text{ continuous.} \tag{1}$$

We further assume that $a_\infty \equiv \lim_{t\to\infty} a(t)$ exists and is finite.

Perhaps the simplest method of numerical solution of (1) is Euler's method, which takes the form $x_{n+1} = x_n + T(c_n - a_n x_n)$, $x_0 = y_0$, where x_n approximates $y(nT)$, and $a_n = a(nT)$, $c_n = c(nT)$. If we define $\bar{a}_n$ via $e^{-\bar{a}_n T} = 1 - a_n T$, ρ_n via $\bar{a}_n = (1 + \rho_n)a_n$ and $\bar{c}_n$ via $\bar{c}_n = (1 + \rho_n)c_n$, then we see that $x_n = x_n(nT)$, where $\dot{x}_n(t) + \bar{a}_n x_n(t) = \bar{c}_n$, $x_n(nT) = x_{n-1}(nT)$, $x_0(0) = y_0$, for $t \geq nT$. Thus the numerical solution is a discretized continuous solution to perturbed equations.

* This research was performed while the author was a member of the IBM Systems Research and Development Center, Palo Alto, California.

† Now at the U. S. Army Mathematics Research Center, Madison, Wisconsin.

Returning to the general problem, we wish to compare the solution $y(t)$ of (1) with the solutions to the equations:

$$\dot{x}_n(t) + \bar{a}_n x_n(t) = \bar{c}_n, \qquad x_n(nT) = x_{n-1}(nT), \qquad x_0(0) = y_0, \tag{2}$$

where $\bar{a}_n = (1 + \rho_n)a_n$ and $\bar{c}_n = (1 + \rho_n)c_n$, with a_n and c_n some appropriate representation of $a(t)$ and $c(t)$ on $nT \leq t \leq nT + T$; the particular form of a_n, c_n, and ρ_n is determined as in the above example by the specific method in use.

Clearly we may write

$$\dot{E}_n + \bar{a}_n E_n = f_n(t) \equiv c(t) - \bar{c}_n + (\bar{a}_n - a(t))y(t), \qquad E_n(nT) = E_{n-1}(nT), \qquad E_0(0) = 0,$$

where $E_n(t) = y(t) - x_n(t)$, $nT \leq t$.

Writing $d_n \equiv E_n(nT)$, this yields

$$d_{n+1} = e^{-\bar{a}_n T} d_n + \gamma_n, \qquad d_0 = 0, \tag{3}$$

$$\gamma_n \equiv \int_{nT}^{nT+T} e^{-\bar{a}_n(nT+T-t)} f_n(t)\, dt.$$

For convenience, we write $A_n \equiv e^{-\bar{a}_n T}$. Thus (3) becomes

$$d_{n+1} = A_n d_n + \gamma_n, \qquad d_0 = 0. \tag{4}$$

Define $A \equiv e^{-\bar{a}_\infty T}$, $\varepsilon_n = A_n - A$, where $\bar{a}_\infty = (1 + \rho_\infty)a_\infty = \operatorname{limit}_{nT\to\infty} \bar{a}_n$ is assumed to exist. Then $d_{n+1} = Ad_n + \varepsilon_n d_n + \gamma_n$, $d_0 = 0$ which implies

$$d_{n+p} = A^p d_n + \sum_{i=0}^{p-1} A^{p-i-1}\gamma_{n+i} + \sum_{i=0}^{p-1} A^{p-i-1}\varepsilon_{n+i} d_{n+i}. \tag{5}$$

Next we assume that $||\varepsilon_i|| \leq \eta_n$, $i \geq n$, that $||A^i|| \leq \alpha r^i$, $i \geq 0$, and that $||\gamma_i|| \leq \Gamma_n$, $i \geq n$.

From (5) it follows then that

$$||d_{n+p}|| \leqq \alpha r^p ||d_n|| + \alpha\Gamma_n \frac{1 - r^p}{1 - r} + \alpha\eta_n \sum_{i=0}^{p-1} r^{p-i-1} ||d_{n+i}||. \tag{6}$$

Defining

$$D_{n,p} \equiv \alpha\eta_n \sum_{i=0}^{p-1} r^{p-i-1} ||d_{n+i}|| + \alpha r^p ||d_n||$$

we have

$$D_{n,p+1} - rD_{n,p} = \alpha\eta_n ||d_{n+p}|| \leqq \alpha\eta_n \left[D_{n,p} + \alpha\Gamma_n \frac{1 - r^p}{1 - r} \right],$$

which yields

$$D_{n,p+1} \leqq (r + \alpha\eta_n)^p D_{n,1} + \alpha^2 \sum_{i=1}^{p} (r + \alpha\eta_n)^{p-i} \eta_n \Gamma_n \frac{1 - r^i}{1 - r}.$$

Finally,

$$\begin{aligned} ||d_{n+p}|| &\leqq \alpha\Gamma_n \frac{1 - r^p}{1 - r} + (r + \alpha\eta_n)^{p-1}(\alpha\eta_n + \alpha r)\, ||d_n|| \\ &\quad + \alpha^2\Gamma_n\eta_n \sum_{i=1}^{p-1} (r + \alpha\eta_n)^{p-i-1} \frac{1 - r^i}{1 - r} \\ &\leqq \alpha\Gamma_n \frac{1 - r^p}{1 - r} + (r + \alpha\eta_n)^{p-1}(\alpha\eta_n + \alpha r)\, ||d_n|| \\ &\quad + \frac{\alpha^2\Gamma_n\eta_n}{1 - r} \left[\frac{1 - (r + \alpha\eta_n)^{p-1}}{1 - (r + \alpha\eta_n)} \right. \\ &\quad \left. - \frac{r}{\alpha\eta_n} ((r + \alpha\eta_n)^{p-1} - r^{p-1}) \right]. \end{aligned} \tag{7}$$

Under certain additional assumptions (7) will give us the error analysis we seek.

First, we wish to determine what happens as nT tends to infinity in the case for which all eigenvalues of a_∞ have positive real parts and $c_\infty \equiv \lim_{t\to\infty} c(t)$ exists, i.e., for the case in which the continuous solution, $y(t)$, tends to $y_\infty = a_\infty^{-1} c_\infty$ as t tends to infinity. We can draw valid conclusions here if $\bar{a}_\infty$ satisfies the same assumption as a_∞; we assume this. Then clearly we have $r < 1$ in (7), and, if we take n so large that $r + \alpha\eta_n < 1$, sending p to infinity in (7) yields

$$\overline{\lim_{i\to\infty}} ||d_i|| \leqq \alpha\Gamma_n \left(\frac{1}{1 - r} + \frac{\alpha\eta_n}{1 - r - \alpha\eta_n} \right) \tag{8}$$

for all large n.

To go further we must analyze Γ_n. Recall that we assume $||\gamma_i|| \leq \Gamma_n$, $i \geq n$. Also from (3) and the definition of f_j, we have

$$||\gamma_i|| \leq \delta_j + \epsilon_j + ||\rho_j||\, g_j,$$

where

$$\begin{aligned} \delta_j &= \left|\left| \int_{jT}^{jT+T} e^{-\bar{a}_j(jT+T-t)}(c(t) - c_j)\, dt \right|\right| \\ \epsilon_j &= \left|\left| \int_{jT}^{jT+T} e^{-\bar{a}_j(jT+T-t)}(a_j - a(t))y(t)\, dt \right|\right| \\ g_j &= \left|\left| \int_{jT}^{jT+T} e^{-\bar{a}_j(jT+T-t)}(a_j y(t) - c_j)\, dt \right|\right|. \end{aligned} \tag{9}$$

Further assuming that a_n, c_n tend to a_∞, c_∞ as n nears infinity, we know that δ_j, ε_j, and g_j approach zero as j approaches infinity. Thus

Theorem 1: *If a_∞, c_∞ exist, if a_∞ and $\bar{a}_\infty$ have all their eigenvalues with positive real parts, if $\lim_{nT\to\infty} a_n = a_\infty$, $\lim_{nT\to\infty} c_n = c_\infty$, then for fixed T, the solutions $y(t)$, $x_n(t)$ of (1), (2) satisfy $\lim_{n\to\infty} ||y(nT) - x_n(nT)|| = 0$.*

Next we wish to analyze the order of the error at $t = nT$ as T approaches zero. Because of the boundedness of

$a(t)$, if ρ_n (and $\bar{a}_n$) is bounded, we have $||e^{-\bar{a}_n T}|| \leq 1 + \beta T$ for some β. Thus, from (3), we have

$$||d_n|| \leq \frac{\Gamma_0}{\beta T}[(1+\beta T)^n - 1] \leq \frac{\Gamma_0}{\beta T}(e^{\beta t} - 1), \qquad \Gamma_0 \geq \sup_{i \geq 0} ||\gamma_i||. \tag{10}$$

Let us assume that there exist constants α_1, α_2, α_3, α_4, p_1, p_2, p_3 such that $\delta_n \leq \alpha_1 T^{p_1}$, $\varepsilon_n \leq \alpha_2 T^{p_2}$, $||\rho_n|| \leq \alpha_3 T^{p_3}$, $g_n \leq \alpha_4 T$, for all n. Then (10) yields

$$||x_n(nT) - y(nT)|| \leq \frac{e^{\beta t} - 1}{\beta}(\alpha_1 T^{p_1 - 1} + \alpha_2 T^{p_2 - 1} + \alpha_3 \alpha_4 T^{p_3}). \tag{11}$$

Theorem 2: *If there exist constants α_1, α_2, α_3, α_4, p_1, p_2, p_3, such that $\delta_n \leq \alpha_1 T^{p_1}$, $\varepsilon_n \leq \alpha_2 T^{p_2}$, $||\rho_n|| \leq \alpha_3 T^{p_3}$, $g_n \leq \alpha_4 T$ for all n, then, defining $p = \min(p_1 - 1, p_2 - 1, p_3)$, $||x_n(nT) - y(nT)||$ tends to zero like T^p as T tends to zero with $nT = t$ fixed; that is, there exists a constant α such that $||x_n(nT) - y(nT)|| \leq \alpha T^p$ (α depends on t).*

Intuitively, (11) says that our total error comes from three distinct types of error: (i) the error in representing the input $c(t)$; (ii) the error in representing the time varying parameter $a(t)$; (iii) the error in matching the roots of the difference and the differential equations. Errors of type (i) are measured by δ, errors of type (ii) by ε, and errors of type (iii) by ρ.

III. Nonlinear equations

For systems of nonlinear differential equations we can arrive at roughly the same results, using essentially the same standard analysis given in Section II; of course, rather strong assumptions must be made in some cases. For completeness we demonstrate the techniques involved.

Consider the nonlinear system analogous to (1), i.e.,

$$\dot{y} + a(t, y(t))y(t) = c(t), \qquad y(0) = y_0. \tag{12}$$

Suppose there exists a continuous solution $y(t)$, and denote $a(t, y(t))$ by $a(t)$. Then the solution $y(t)$ to (12) also solves

$$\dot{z} + a(t)z = c(t), \qquad z(0) = y_0 \tag{12'}$$

Under the assumptions of Theorem 2 for this equation, we conclude that the numbers $w_n(nT)$ defined via

$$\dot{w}_n + \bar{a}_n w_n = \bar{c}_n, \qquad w_n(nT) = w_{n-1}(nT), \qquad w_0(0) = y_0 \tag{13}$$

satisfy $||w_n(nT) - y(nT)|| \leq \alpha T^p$, where α depends on T and p depends upon the method of approximation, as in Theorem 2. Introducing slightly new notation we have $\bar{a}_n(y_n) \equiv [1 + \rho(a_n(y_n))]a_n(y_n)$, $\bar{c}_n = [1 + \rho(a_n(y_n))]c_n$, where 1 denotes the identity matrix and $\rho(u)$ a matrix.

From (13) we deduce that $\dot{w}_n(t) + \bar{a}_n(w_n)w_n(t) = \bar{c}_n + [\bar{a}_n(w_n) - \bar{a}_n(y_n)]w_n(t)$. Thus

$$\dot{w}_n(t) + \bar{a}_n(w_n)w_n(t) = c_n + \rho(a_n(y_n))c_n + [\bar{a}_n(w_n) - \bar{a}_n(y_n)]w_n(t). \tag{14}$$

However, the numbers we actually compute satisfy not (14) but

$$\dot{x}_n(t) + \bar{a}_n(x_n)x_n(t) = c_n', \qquad x_n(nT) = x_{n-1}(nT), \qquad x_0(0) = y_0, \tag{15}$$

where

$$\bar{a}_n(x_n) = [1 + \rho(a_n(x_n))]a_n(x_n), \qquad \bar{c}_n = [1 + \rho(a_n(x_n))]c_n.$$

Solving (14) and (15) we deduce

$$w_{n+1} = e^{-\bar{a}_n(w_n)T}w_n + \bar{a}_n(w_n) - (1 - e^{-\bar{a}_n(w_n)T})(c_n + \rho(a_n(w_n))c_n) + (\bar{a}_n(w_n) - \bar{a}_n(y_n))\int_0^T e^{-\bar{a}_n(w_n)(T-t)}w_n(nT + t)\,dt, \tag{16}$$

$$x_{n+1} = e^{-\bar{a}_n(x_n)T}x_n + \bar{a}_n(x_n)^{-1} \times (1 - e^{-\bar{a}_n(x_n)T})(c_n + \rho(a_n(x_n))c_n).$$

Subtracting yields

$$w_{n+1} - x_{n+1} = e^{-\bar{a}_n(w_n)T}w_n - e^{-\bar{a}_n(x_n)T}x_n + \mu_n, \tag{17}$$

where the definition of μ_n is obvious.

Assuming that $a(t, u)$ is uniformly bounded and satisfies a Lipschitz condition in u, and recalling our assumption that the function $\rho(\cdot)$ is of order T^p as T approaches zero, it is simple to deduce that there exist constants β_1, β_2 such that $||\mu_n|| \leq \beta_1 T^{p+1} + \beta_2 T\,||w_n - x_n||$. These conditions further imply that there exists a constant β_3 such that $||e^{-\bar{a}_n(w_n)T}w_n - e^{-\bar{a}_n(x_n)T}x_n|| \leq (1 + \beta_3 T)\,||w_n - x_n||$ as T nears zero. Thus

$$||w_{n+1} - x_{n+1}|| \leq (1 + \beta_2 T + \beta_3 T)\,||w_n - x_n|| + \beta_1 T^{p+1}. \tag{18}$$

Thus, just as in the steps leading up to Theorem 2, we can easily show that there exists a constant $\beta_4 = \beta_4(t)$, $t = nT$, such that $||w_n - x_n|| \leq \beta_4 T^p$ as T approaches zero. Since a similar relation has already been demonstrated between w_n and $y(nT)$, we have

Theorem 3: *Suppose that the assumptions of Theorem 2 for the linear equation (12') are valid; this then defines the order p. If $a(t, u)$ is uniformly bounded and satisfies a Lipschitz condition in u, then there exists a $\beta = \beta(t)$ such that for $t = nT$ fixed as T approaches zero, $||x_n - y(nT)|| \leq \beta T^p$.*

As usual, to prove anything about the final values we must make stronger assumptions.

Theorem 4: *If, to the assumptions of Theorem 3, we add the assumption that* $\lim_{t\to\infty} a(t, u) = a_\infty$ *uniformly in u, with all the eigenvalues of* a_∞ *and* $\bar{a}_\infty$ *having positive real parts, and if the final value of the continuous solution* $y(t)$ *is* y_∞, *then* $\lim_{n\to\infty} x_n$ *exists and equals* y_∞.

Sketch of Proof: Defining $A_\infty = e^{-\bar{a}_\infty T}$, we replace the expressions $e^{-\bar{a}_n(w_n)T}$ and $e^{-\bar{a}_n(x_n)T}$ in (17) by A_∞ plus error terms. Since $||A_\infty^n|| \leq \alpha r^n$, $r < 1$, we can proceed exactly as we did prior to Theorem 1, deducing that $\lim_{j\to\infty} ||x_j - w_j|| \leq M_n \times$ constant, where $\mu_i \leq M_n$, $i \geq n$. But our assumptions imply that μ_i tends to zero. Thus x_∞ exists and equals $w_\infty = y_\infty$. Q.E.D.

Thus we are able to prove convergence to the correct final value in those cases in which the nonlinearity "disappears" for a large time, leaving an otherwise stable system.

IV. Examples; Fowler's method

Before we examine the error in approximation by Fowler's method, let us show how the previous results can be used to derive error bounds for a simpler method for the sake of clarity.

A. *Euler's method*

Euler's method was stated at the start of Section II; consider its application to $\dot{y} + a(t, y)y = c(t)$, $y(0) = y_0$. We define $\rho(a)$ via $e^{-(1+\rho(u))uT} = 1 - uT$. Thus $\bar{a}_n(y_n) = 1/T \ln (1 - a_n(y_n)T) = a_n(y_n)[1 + \frac{1}{2}a_n(y_n)T + \cdots]$, yielding $p_3 = 1$. Clearly $p_1 = p_2 = 2$, so that the error is of order T, as is well known. If $a(t, u)$ approaches a_∞, clearly we require T to be small for stability (if a_∞ is stable).

B. *Fowler's method*

We first present a trivial example solely to illustrate the ideas involved; consider the first-order scalar equation $\dot{y} + a(t, y)y = c(t)$, $y(0) = y_0$, where we assume $0 < m \leq a(t, y) \leq M$, $\lim_{t\to\infty} a(t, y) = a_\infty$, $\lim_{t\to\infty} c(t) = c_\infty$. The z-transform based simulation for this gives the iteration $x_{n+1} = (1 - \kappa a_n T)x_n + \kappa T c_n$, where a_n and c_n approximate $a(t, y)$, and $c(t)$ and κ is a suitable factor. Most simply we let $a_n = a(nT, x_n)$, $c_n = c(nT)$, and $\kappa = (1 - e^{-MT})/MT$, this choice of κ guaranteeing that $\bar{a}_n$ defined by $e^{-\bar{a}_n T} = 1 - \kappa a_n T$ is bounded away from zero. To detect the order of the error we write $\bar{a}_n = -1/T \ln (1 - \kappa a_n T) = -1/T \ln [1 - a_n T(1 - AT/2 \cdots)] = a_n[1 - \{(A - a_n)/2\} T + \cdots]$ which states that ρ is of order T. By our choice of a_n and c_n this implies that the error is of order T, while our choice of κ guarantees the attainment of the correct final value and the stability of the iteration defining x_n.

We next present a somewhat more complex example for which Fowler's method can be precisely stated for all T, yielding an error estimate; we do not, however, describe how the simulation is determined since the application of the method has been described thoroughly elsewhere.[3,4] Consider the nonlinear system

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}^{\cdot} + \begin{bmatrix} 0 & A \\ -1 & 2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \tag{19}$$

where A is a nonlinear function of t, y_1, and y_2 satisfying $0 < m \leq A \leq M$, $\lim_{t\to\infty} A = A_\infty$; for clarity, let $A(t, y_1, y_2) = 4 + 5.12te^{-t} - 3.25e^{-t-y_1^2-y_2^2}$, so $0.75 \leq A \leq 5$, $A_\infty = 4$. As A varies, the "instantaneous eigenvalues" of the system are $s_A = 1 \pm \sqrt{1 - A}$.

Fowler's method, applied to simulate the continuous system (19), gives the difference equations (in vector notation):

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^{(n+1)} = \begin{bmatrix} 1 & -\kappa AT \\ \dfrac{1 - e^{-2T}}{2} & (1 - e^{-2T})1 + \dfrac{\kappa AT}{2} \end{bmatrix}^{(n)} \times \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^{(n)} + T \begin{bmatrix} \kappa c_1 \\ \dfrac{1 - e^{-2T}}{2T}(c_2 + Tc_1) \end{bmatrix}^{(n)}. \tag{20}$$

The behaviour of such a system is influenced by the "instantaneous roots" z_A satisfying $(z - 1)(z - e^{-2T}) + \frac{1}{2}\kappa AT(1 - e^{-2T})z = 0$; the crucial part of Fowler's method is to match the two loci z_A and $e^{-s_A T}$ as closely as possible over the entire range of A, keeping $|z_A| < 1$ if possible. This is easily accomplished by picking κ such that $z_A = e^{-s_A T}$ at the extreme value $A = 5$. If we write

$$a(t, y_1, y_2) = \begin{bmatrix} 0 & A \\ -1 & 2 \end{bmatrix}$$

and

$$a_n = \begin{bmatrix} 0 & A(nT, x_1^{(n)}, x_2^{(n)}) \\ -1 & 2 \end{bmatrix}$$

and note that $\kappa = 1 - 2T + 0(T^2)$, then defining $\bar{a}_n$ as usual from (20) we have $\bar{a}_n = a_n(1 + 0(T))$; choosing $c_i^{(n)} = c_i(nT)$ for $i = 1, 2$ then yields a total error of order T. We also note that the difference equation (20) is now stable and that we obtain the correct final value. The close matching of the loci (in fact for all A in $0.75 \leq A \leq 5$ there exists A' in $0.75 \leq A' \leq 5$ with $z_{A'} = e^{-s_A T}$) implies that the dynamics of the solution to (20) will closely match those of the solution to (19).

In the last example the proof of the order of convergence was made possible by the existence of a fixed rule for setting up the difference equations, i.e. force $z_A = e^{-s_A T}$ at $A = 5$; any time that such a rule exists we can estimate the error via Sections II and III. Unfortunately it seems impossible to give a more precise analysis of the errors

resulting from Fowler's method without restricting the equations considered to entirely unrealistic special cases;* the method as it stands is an excellent approach to practical problems but its very flexibility limits our ability to analyze it completely.

V. Concluding remarks

Although, as we mentioned in Section I, the order of convergence is sometimes of little practical interest, we regret that a completely general analysis of Fowler's method was not possible; perhaps a more important limitation, however, lies in the strong assumptions used. The appropriate stability condition for the continuous system is that $a(t, x)x$ be strongly monotone in x, a much weaker condition than we imposed. One might hypothesize that if a continuous system is stable and is simulated by a discrete system with close enough root-loci, then the discrete system is also stable and converges to the correct final value; this is true, for example, in the Fowler simulation of the (only apparently) nonlinear system

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}^{\cdot} + \begin{bmatrix} 1 - y_2 & y_1 \\ y_2 & 1 - y_1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix},$$

$$y_1(0) = 10, \qquad y_2(0) = 0.9,$$

whose matrix has one positive and one negative root. In complete generality this hypothesis is probably untrue; we wonder under what assumptions it is valid.

* For example, the system (12) sometimes leads to the Fowler's simulation $x_{n+1} = (1 - T\kappa a_n)x_n + T\kappa c_n$, where $a_n = a(nT, x_n)$, $c_n = c(nT)$, and κ is a constant matrix chosen such that $\bar{a}_n$ is uniformly positive definite and satisfies $\bar{a}_n = a_n(1 + 0(T))$, thus giving a total error of order T; the example (19), (20) shows however that this is not a widely applicable special case.

References

1. IBM Publication (E20-8186), *Flight Simulation Experience Using a New IBM Numerical Technique.*
2. IBM Publication (E20-0029-1), *Numerical Techniques for Real-Time Digital Flight Simulation.*
3. M. E. Fowler, "Numerical Methods for the Synthesis of Linear Control Systems," *Automatica*, 1, 207–225 (1963).
4. M. E. Fowler, "A New Numerical Method for Simulation," *Simulation*, 6, 324–330 (1965).
5. M. E. Fowler, "An Example Showing the Use of Root Locus Techniques to Study Nonlinear Systems," Technical Report, IBM Systems Research and Development Center, Palo Alto, California, August 12, 1964.

Received April 26, 1966