TECHNICAL COMMENT

Some critical remarks on a new numerical method for simulation of dynamical systems

by ELMER G. GILBERT

Information and Control Engineering Program
The University of Michigan

This work was supported by USAF under grant number AF-AFOSR-814-65 and was carried out while the author was at the Frank J. Seiler Research Laboratory at the USAF Academy.

Recently a new numerical method for simulation of complex dynamical systems has been proposed by M. E. Fowler.[1,2] There is no question that for some problems, e.g., single-loop servo-systems with simple nonlinearities or slowly varying gains, the method is a good procedure for developing difference equations which generate solutions of adequate accuracy at high speed on the digital computer. However, it is misleading to imply that the method is of general applicability and will replace, because of tremendous speed advantage, the well established classical methods (Runge-Kutta, Adams-Moulton, etc.), which have a firm basis, both in the manner of their mathematical development and their long history of successful use. This assertion will be supported by the rather general remarks of the next three paragraphs and by the more detailed analysis which follows in the subsequent paragraphs.

First of all the required root locus matching, even when computer assisted, involves a lengthy (z-transform) analysis of the system to be simulated by a skilled practitioner of the method. This is in opposition to classical methods in which one goes directly from the differential equations of the system to the computer program. Certainly, in some cases the resulting system analysis may be a useful by-product, but more generally it does not take a satisfactory form to answer the usual engineering questions.

Another difficulty in the method is its application to differential equations which are coupled in such a complex way (example—equations of motion for a space vehicle) that it is not obvious how to form the loops for root-locus analysis. The forming of the loops is certainly not unique and the solution accuracy obtained will depend in an obscure way on the choice made. In strongly coupled equations it is even doubtful that the method will prove workable.

Suppose that it is possible to form the loops required in the method. There still is no assurance that the method can be made to work in all instances. It may be impossible to construct root-locus diagrams which match over the required gain range. Even if they do match it is not certain that when time-varying gains (or nonlinearities, see reference 2) are inserted the simulation will be accurate. The method is based on the z-transform and Laplace transform which are valid in the context considered when the systems are *linear* and *time-invariant* (n.b., equation (11) in reference 1 is not valid unless $C(t)$ is constant). There are numerous examples of time-varying linear systems where such "quasistatic" analysis fails. For example it is possible to construct a system which is stable for any constant gain C such that $.5 \leq C \leq 1.$, but is unstable when C is made to vary in the same range. But even if the method is applied to linear time in variant systems it has serious limitations. Let us point these out by examining closely a first order system.

Suppose we try to solve

$$\frac{dx}{dt} = -x + r(t), \tag{1}$$

where $x(t)$ is the system response and $r(t)$ is the input function. Table I shows four different difference equation representations for (1) where T is the integration interval, n is an integer, $r_n = r(nT)$, and x_n approximates $x(nT)$.

Table I – Difference equation representations of first order differential equation

METHOD	DIFFERENCE EQUATION	TRANSFER FUNCTION
A Euler	$x_{n+1} = x_n(1-T) + Tr_n$	$\frac{T}{z-1+T}$
B Heun	$x_{n+1} = x_n\left(1 - T + \frac{1}{2}T^2\right) + \frac{1}{2}Tr_{n+1} + \left(\frac{1}{2}T - \frac{1}{2}T^2\right)r_n$	$\frac{\frac{1}{2}T\left(T(z+1-T)\right)}{z-1+T-\frac{1}{2}T^2}$
C Fowler	$x_{n+1} = x_n e^{-T} + (1-e^{-T})\, r_n$	$\frac{(1-e^{-T})}{z-e^{-T}}$
D Fowler	$x_{n+1} = x_n e^{-T} + [1 - T^{-1}(1-e^{-T})]r_{n+1} + [-e^{-T} + T^{-1}(1-e^{-T})]r_n$	$\frac{T^{-1}(1-e^{-T})(1-z) + z - e^{-T}}{z-e^{-T}}$

Representations A and B are obtained by direct application of the relatively crude Euler and Heun integration methods.[3] Representation C has been obtained by writing the solution of (1) in the form

$$x(t) = e^{-t}\left[x(0) + \int_0^t e^{\sigma} r(\sigma)\, d\sigma\right] \tag{2}$$

and letting $x_n = x(nT)$ with $r(t)$ replaced by a stepwise approximation (*i.e.*, $r(t) \leftarrow r_n$, $nT \leq t < T + nT$). Representation D also follows from (2) but with $r(t)$ replaced by a continuous piecewise linear approximation $[r(t) \leftarrow r_n + T^{-1}(r_{n+1} - r_n)\,(t - nT),\ nT \leq t \leq T + nT]$. These last two representations are of the type discussed by Fowler.[1] When generalized to vector-matrix notation this method of derivation gives a systematic, computationally oriented procedure for determining the difference equations (corresponding to C and D) for a linear time-invariant system of any order. Also shown in table I are the transfer functions $H^*(z)$ corresponding to the four representations. That is, if $X^*(z)$ and $R^*(z)$ are z-transforms[4] of the data sequences $\{x_n\}$ and $\{r_n\}$ and the systems are initially at rest $(x_o = 0)$, then

$$H^*(z) = H^*(z)R^*(z). \tag{3}$$

For the unforced case, $r_n = 0$, $n \geq 0$, the solutions of A, B, C, and D can be expressed $[x_o = x(0)]$ as

$$x_n = e^{\lambda(nT)}x(0). \tag{4}$$

Ideally, of course, $\lambda = -1$. As expected $\lambda = -1$ for C and D. For the other two cases it is easy to show

representation A:

$$\lambda = T^{-1}\ln(1-T) = -1 - \frac{1}{2}T + \cdots, \tag{5}$$

representation B:

$$\lambda = T^{-1}\ln\left(1 - T + \frac{1}{2}T^2\right) = -1 + \frac{1}{6}T^2 + \cdots. \tag{6}$$

Table II – Errors in step and ramp response

CASE		e_1	e_∞
Step Input A	$\left(T - \frac{1}{2}T^2 + \cdots\right)$	$\left(T - 2T^2 + \cdots\right)$	0
Step Input B	$\left(T - \frac{1}{2}T^2 + \cdots\right)$	$\left(T - \frac{3}{2}T^2\right)$	0
Step Input C	$\left(T - \frac{1}{2}T^2 + \cdots\right)$	$\left(T - \frac{3}{2}T^2 + \cdots\right)$	0
Step Input D	$\left(T - \frac{1}{2}T^2 + \cdots\right)$	$\left(T - \frac{3}{2}T^2 + \cdots\right)$	0
Ramp Input B	$\left(\frac{1}{2}T^2 - \frac{1}{6}T^3 + \cdots\right)$	$\left(\frac{1}{2}T^2 - \frac{5}{6}T^3 + \cdots\right)$	0
Ramp Input D	$\left(\frac{1}{2}T^2 - \frac{1}{6}T^3 + \cdots\right)$	$\left(\frac{1}{2}T^2 - \frac{2}{3}T^3 + \cdots\right)$	0

Table III – Fractional error in complex gain for $r(t) = e^{jt}$

METHOD	A (Euler)	B (Heun)	C (Fowler)	D (Fowler)
E	$T(.25 + j.25)$	$-T^2(.17 - j.58)$	$-T(.5)$	$-T^2(.5 + j.33)$

Thus there is *no error* for the Fowler representations, while the Euler and Heun methods give errors in the exponential constant which are of the order of T and T^2, respectively.

The apparent advantage of representations C and D disappears, however, when input forcing is considered. It is true that C(D) will give zero error if the input is a step (ramp) starting at $t = nT$, $n =$ integer. This is obvious from the way in which representations C and D were derived. However, if the step (ramp) input is not applied precisely at $t = nT$, $n =$ integer, a large error may be produced. This effect appears in figure 8 and table I of reference 1. If the step input were applied at $t = -.0199$ or $t = -.0001$ the same solution would be obtained from the difference equation representation. However, the actual solution of the differential equation (labeled "exact solution" in figure 8 and table I of reference 1) would then be advanced or delayed by .0099 seconds. On the leading edge of the response (for $C = 35$) where the slope is great this can result in a solution error of approximately .2 units, which is about 15 times the worst error observed in table I of reference 1.

Now consider a similar comparison for representations A, B, C, and D of (1). It is assumed hereafter that the system (1) and its representations A, B, C, and D are at rest before the inputs are applied. For a unit step input occurring at $t = -T^+$ (just after $t = -T$) it follows that $r_n = 0$, $n < 0$ and $r_n = 1$, $n \geq 0$. Substituting this in A, B, C, and D and using $e_n = x(nT) - x_n$ for the error, where $x(t) = 1 - e^{-(t+T)}$ is the solution of (1) for $t > -T$, the results in rows 1, 2, 3, and 4 of table II are obtained. Power series in T are used to express e_n so that the effect of changes in the interval size is more easily ascertained. It is clear that solution errors are comparable in all four representations! Rows 5 and 6 of table II show a similar comparison between representations B and D for a unit ramp input starting $t = -T^+$. Again the classical approach unit ramp input starting $t = -T^+$. Again the classical approach and Fowler's approach yield comparable performance (note that the fractional errors in the solution are large since

$$x(0) = \frac{1}{2}T^2 - \frac{1}{6}T^3 + \cdots \quad \text{and} \quad x(T) = 2T^2 - \frac{4}{3}T^3 + \cdots).$$

It is interesting to note that equally poor (errors same order in T) solution accuracy is obtained for the above inputs if a fourth order Runge-Kuta formula is used. This is because the full accuracy of the Runge-Kutta formula is not available unless $(d^4/dt^4)r(t)$ is continuous,[3] which is clearly not the case. With step inputs the Euler formula does about as well as any of the more involved integration formulas. This shows that one must be careful to choose a valid classical formula when comparisons of computing speed are made for discontinuous inputs.

To show that accuracy comparisons of the above type are not dependent on inputs which are discontinuous or have discontinuous derivatives of some order, consider a sinusoidal input $r(t) = e^{j\omega t}$ where $\omega = 1$, the break frequency of the first order system. The complex gain of the system (1) is $(j\omega + 1)^{-1}$. The corresponding gains for the discrete representations A, B, C, and D are found from $H^*(e^{j\omega T})$ (see reference 3). Table III gives

$$E = (j + 1)H^*(e^{jT}) - 1, \tag{7}$$

which can be interpreted as the fractional error in the complex gain. The entries shown in table III are the leading terms in power series in T. Thus the Euler method A gives an error comparable to C and the Heun method B gives an error comparable to D! The error associated with D is not nearly so small as the error produced by a fourth order Runge-Kutta formula for which it can be shown that E is the order of T^4.

REFERENCES

1 M E FOWLER
A new numerical method for simulation
Simulation vol 4 1965 pp 324-330

2 J M HURT
A new difference equation technique for solving nonlinear differential equations
AFIPS Conference Proceedings vol 25 1964 pp 169-179

3 P HENRICI
Discrete variable methods in ordinary differential equations
Wiley New York 1962

4 J R RAGAZZINI G F FRANKLIN
Sampled-data control systems
McGraw Hill New York 1958