Evaluation of Errors at Numerical Integration of Ordinary Differential Equations

JUN-ICHI BABA* AND SHIGEO HAYASHI*

Summary

In this paper, the authors present a new theory of error evaluation at numerical integration. The classical theories of error evaluation based on the Taylor expansion inform us only of the order of magnitude of errors, and are not capable of clarifying the exact value and the nature of errors.

It is known the numerical integration of the set of linear differential equations can be reduced into the solution of the set of the corresponding difference equations, and the difference equations can be solved by means of the matrix technique and the z-transform. With these facts in mind, authors developed a new theory, which clarifies the nature errors and gives criteria for selecting adequate time intervals that keep the errors for computation within the allowable limits.

Though the round-off errors are not discussed in this paper, it is shown by the example that the round off errors are much smaller than the truncation errors discussed in the paper.

1. *Introduction*

In order to analyze transient phenomena in a physical system, we have to solve a set of linear differential equations describing its performance. The solutions of the set of linear differential equations in explicit forms are not usually feasible and therefore numerical calculation becomes necessary.

For numerical calculation, care must be taken for the selection of the sampling time interval. Excessively long time interval causes appreciable errors in the numerical solution, while shorter time interval requires unduly long computation time to obtain the solution. Therefore, adequate sampling time interval must be chosen for the analysis of the transients, and a certain theory of error evalution of numerical calculation becomes necessary.

The classical theories[1] of error evaluation based on the Taylor expansion inform us only of the order of the magnitude of the errors, and are not capable of clarifying the exact value and the nature of the errors.

In order to overcome the difficulty of the classical theories, the authors have developed a new theory of the error evaluation. In the theory the numerical calculation of the set of linear differential equations can be turned into the solution of the set of the corresponding linear difference euations, which can be solved by means of the matrix technique[2)3)] and the z-transform.[4)]

* Research Inst., Mitsubishi Electric Mfg., Co., Ltd. Itami, Hyogo.

The new theory, discussed in this paper, can be used to choose the adequate sampling time interval for the numerical calculations.

2. *The General Form of the Simultaneous Linear Differential Equations with Constant Coefficients*

It is possible to change any set of differential equations into such a set of equations involving only first-order differential coefficients by writing the higher-order coefficients as the differential coefficients of new independent variables. Therefore, we have the general forms of simultaneous linear differential equations as

$$\frac{dx_i}{dt}+\sum_{i=1}^{n} a_{ij}x_i=b_i \quad (i=1,2,\cdots,n) \tag{2.1}$$

where $$x_i=x_{i_0} \quad \text{at} \quad t=0.$$

The set of equations (2.1) may be written in matrix form as

$$\frac{d[x]}{dt}+[A][x]=[B] \tag{2.2}$$

where the initial value of $[x]$ is $[x_0]$.

Since the equation (2.2) is linear, and the principle of superposition is applicable in this case, we may assume that $[B]$ is a constant matrix. Applying to eq. (2.2) the Laplace transformation yields

$$s[X(s)]+[A][X(s)]=[B]/s+[x_0] \tag{2.3}$$

where elements of the matrix $[A]$ are constant.

3. *Numerical Calculation of Ordinary Differential Equations*

Although there are many methods of numerical calculation of ordinary differential equations, here we discuss the following three representative ones:

3.1 The Method of Euler

If x is expressed as a function of t by the equation $\frac{dx}{\alpha t}=f(t,x)$, the increment Δx in x corresponding to an increment Δt in t is given approximately by the equation $\Delta x=f(t,x)\Delta t$ the value of $f(t,x)$ being that at the begining of the interval Δt. Thus if (t_0, x_0) are the corresponding initial values of the argument t and the function x, the first increment is given by

$$\Delta x_1=f(t_0,x_0)\Delta t_1 \tag{3.1}$$

where Δt_1 is the first increment in t. Proceeding in this manner, the value of x corresponding to any value of t, say t_n can be obtained by dividing the range t_n-t_0 into n suitable intervals.

3.2 Modified Method of Euler

For an improvement in accuracy, the foregoing method is modified in the following manner. The above method provides an approximate value of x at the end of the first interval, say x; the average of $(t_0\, x_0)$ and (t, x) gives the middle point of the first interval.

The coefficient $\frac{dx}{dt}$ at the middle point is then multiplied by the increment Δt, thus giving a more accurate value of x_1 than before. This value of x_1 is then used for the second interval, and the process is repeated.

$$\Delta x_1 = f\left(t_0+\frac{\Delta t_1}{2},\ x_0+\frac{1}{2}f(t_0, x_0)\Delta t_1\right)\Delta t_1 \tag{3.2}$$

3.3 The Method of Runge-Kutta

The Method of integrating the differential equations due to Runge was later developed by Kutta. The methods used by Kutta in obtaining his approximations are the third or higher order approximations. Only for the forth-order approximations, the results will be illustrated here.

In these approximations, the error is of the order of $(\Delta t)^5$. The formula with this degree of accuracy is given by

$$\Delta x_1 = \frac{1}{6}(\Delta^{\mathrm{I}}+2\Delta^{\mathrm{II}}+2\Delta^{\mathrm{III}}+\Delta^{\mathrm{IV}}) \tag{3.3}$$

where

$$\begin{aligned}
&\Delta^{\mathrm{I}}\ f(t_0, x_0)\Delta t_1\\
&\Delta^{\mathrm{II}} = f\left(t_0+\frac{\Delta_1 t}{2},\ x_0+\frac{\Delta^{\mathrm{I}}}{2}\right)\Delta t_1\\
&\Delta^{\mathrm{III}} = f\left(t_0+\frac{\Delta t_1}{2},\ x_0+\frac{\Delta^{\mathrm{II}}}{2}\right)\Delta t_1\\
&\Delta^{\mathrm{IV}} = f(t_0+\Delta t_1,\ x_0+\Delta^{\mathrm{III}})\Delta t_1 .
\end{aligned}$$

4. *The Formal Solutions of the Simultaneous Linear Differential Equations with Constant Coefficients*

If the set of equations (2.1) has constant coefficients, the elements of matrix $[A]$ are constant. In this case the solution of eq. (2.2) is

$$[x] = [A]^{-1}([I]-e^{[A]t})([B]-[A][x_0])+[x_0] \tag{4.1}$$

where

$[I]$: unit matrix

$$e^{-[A]t} = [I]-[A]t+\frac{[A]^2}{2!}t^2-\frac{[A]^3}{3!}t^3+\cdots \tag{4.2}$$

If then n characteristic roots $(r=1,2,\cdots, n)$ of the n-th order square matrix $[A]$ are distinct, then, using Sylvester's theorem the expression

(4.2) can be written in the form:

$$e^{-[A]t}=\sum_{r=1}^{n} e^{-\lambda_r t}[K(\lambda_r)] \tag{4.3}$$

where

$$[K(\lambda_r)]=\prod_{s\neq r}^{s=1,\cdots,n}\frac{\lambda_s[I]-[A]}{\lambda_s-\lambda_r} \quad (r=1,2,\cdots,n) \tag{4.4}$$

and the characteristic roots λ_r's satisfy the equation

$$\det\{\lambda_r[I]-[A]\}=0 \quad (r=1,2,\cdots,n)$$
$$\lambda_r=\mu_r+j\nu_r$$
$$\mu_r,\ \nu_r;\ \text{real number} \qquad j^2=-1.$$

Substituting eq. (4.3) into eq. (4.1) yields

$$[x]=[A]^{-1}[B]-\sum_{r=1}^{n}[A]^{-1}\left(\prod_{s\neq r}^{s=1,\cdots,n}\frac{\lambda_s[I]-[A]}{\lambda_s-\lambda_r}\right)([B]-[A][x_0])\,e^{-\lambda_r t} \tag{4.5}$$

Applying the Laplace transformation to eq. (4.5) gives

$$[X(s)]=\frac{[\alpha]}{s}+\sum_{r=1}^{n}\frac{[\beta_r]}{s+\mu_r+j\nu_r} \tag{4.6}$$

where

$$[\alpha]=[A]^{-1}[B]$$
$$[\beta_r]=-[A]^{-1}\prod_{s\neq r}^{s=1,\cdots,n}\frac{\lambda_s[I]-[A]}{\lambda_s-\lambda_r}([B]-[A][x_0]) \quad (r=1,\cdots,n).$$

5. *The Evaluation of Errors of the Solutions obtained by the above Methods*

5.1 The Method of Euler

Applying the method of Euler to the set of equations (2.1), we get the corresponding set of difference equations:

$$\frac{\{x_i(t+\Delta t)-x_i(t)\}}{\Delta t}+\sum_{j=1}^{n}a_{ij}x_j(t)=b_i \quad (i=1,\cdots,n) \tag{5.1}$$

The set (5.1) may be written in matrix form as

$$\frac{[x(t+\Delta t)-x(t)]}{\Delta t}+[A][x(t)]=[B] \tag{5.2}$$

Applying to eq. (5.2) the z-transformation yields

$$\frac{(z-1)[X(z)]}{\Delta t}+[A][X(z)]=[B]\frac{z}{z-1}+\frac{z}{\Delta t}[x_0] \tag{5.3}$$

where $[x_0]$ is the initial value of the column matrix $[x(t)]$ and $[X(z)]$ denotes the z-transform of $[x(t)]$.

Putting $p=\dfrac{z-1}{\Delta t}$, eq. (5.3) may be written in the form:

$$p[X(z)]+[A][X(z)]=\frac{z}{\Delta t}\left(\frac{[B]}{p}+[x_0]\right) \tag{5.4}$$

Remembering that the solution of eq. (2.3) is given by (4.6), we get the solution of eq. (5.4) in the form:

$$[X(z)]=\frac{z}{\Delta t}\frac{[\alpha]}{p}+\frac{z}{\Delta t}\sum_{r=1}^{n}\frac{[\beta_r]}{p+\mu_r+j\nu_r} \tag{5.5}.$$

Substituting $p=\frac{z-1}{\Delta t}$ into eq. (5.5) yields

$$[X(z)]=[\alpha]\frac{z}{z-1}+\sum_{r=1}^{n}\frac{[\beta_r]z}{z-(1-\mu_r\Delta t-j\nu_r\Delta t)} \tag{5.6}.$$

Then, referring to appendix 1.1, we now have the inverse z-transform of eq. (5.6)

$$[x(t)]=[\alpha]+\sum_{r=1}^{n}[\beta_r]\,e^{-\mu'_r t}(\cos\nu'_r t-j\sin\nu'_r t) \tag{5.7}.$$

Comparing eq. (5.7) with eq. (4.5), it can be said that the solution of the original equation is distorted at the application of the method of Euler, in such a manner that μ'_r and ν'_r are in the solution instead of μ_r and ν_r.

5.2 Modified Method of Euler

Applying the modified method of Euler to eq. (2.1), we get the corresponding set of difference equations:

$$\frac{\{x_i(t+\Delta t)-x_i(t)\}}{\Delta t}+\frac{1}{2}\sum_{j=1}^{n}a_{ij}\{x_i(t+\Delta t)+x_i(t)\}=b_i \quad (i=1,2,\cdots,n) \tag{5.8}.$$

We may write eq. (5.8) in matrix form as

$$\frac{[x(t+\Delta t)-x(t)]}{\Delta t}+\frac{1}{2}[A][x(t+\Delta t)+x(t)]=[B] \tag{5.9}.$$

Applying to eq. (5.9) the z-transformation yields

$$\frac{z-1}{\Delta t}[X(z)]+\frac{z+1}{2}[A][X(z)]=[B]\frac{z}{z-1}+\frac{z}{\Delta t}[x_0]+\frac{1}{2}z[A][x_0] \tag{5.10},$$

where $[x_0]$ is the initial value of $[x(t)]$ and $[X(z)]$ denotes the z-tranform of $[x(t)]$. Multiplying by $\frac{2}{z+1}$, we obtain

$$\begin{aligned}&\frac{2(z-1)}{\Delta t(z+1)}[X(z)]+[A][X(z)]\\&=\frac{2(z-1)}{\Delta t(z+1)}\left\{\frac{[B]}{\frac{2(z-1)}{\Delta t(z+1)}}+[x_0]\right\}-\frac{z}{z+1}\{[B]-[A][x_0]\}\end{aligned} \tag{5.11}.$$

Putting $p=\frac{2(z-1)}{\Delta t(z+1)}$, the eq. (5.11) may be written in the form:

$$p[X(z)]+[A][X(z)]=\frac{2z}{\Delta t(z+1)}\left\{\frac{[B]}{p}+[x_0]\right\}-\frac{z}{z+1}\{[B]-[A][x_0]\} \tag{5.12}$$

Since the set (5.12) is linear, the solution is equal to the sum of the solutions of the following two equations:

$$p[X_1(z)]+[A][X_1(z)]=\frac{2z}{\Delta t(z+1)}\left(\frac{[B]}{p}+[x_0]\right) \tag{5.13}$$

and

$$p[X_2(z)]+[A][X_2(z)]=-\frac{z}{z+1}([B]-[A][x_0]) \tag{5.14}$$

Remembering that the solution of eq. (2.3) is given by eq. (5.6) the solution of eq. (5.13) may be written in the form:

$$[X_1(z)]=\frac{2z}{\Delta t(z+1)}\left\{\frac{[\alpha]}{p}+\sum_{r=1}^{n}\frac{[\beta_r]}{p+\mu_r+j\nu_r}\right\} \tag{5.15}$$

Substituting $p=\frac{2(z-1)}{\Delta t(z+1)}$ into eq. (5.15) yields

$$[X_1(z)]=[\alpha]\frac{z}{z-1}+\sum_{r=1}^{n}\frac{[\beta_r]z}{\left(1+\frac{\mu_r+j\nu_r}{2}\Delta t\right)z-\left(1-\frac{\mu_r+j\nu_r}{2}\Delta t\right)} \tag{5.16}$$

According to the same consideration, the solution of eq. (5.14) may be written in the form:

$$[X_2(z)]=-\frac{z-1}{z+1}\frac{2z}{\Delta t(z+1)}\left\{\frac{[\alpha]-[x_0]}{p}+\sum_{r=1}^{n}\frac{[\beta_r]}{p+\mu_r+j\nu_r}\right\} \tag{5.17}$$

Substituting $p=\frac{2(z-1)}{\Delta t(z+1)}$ into eq. (5.17) yields

$$[X_2(z)]=-\frac{z}{z+1}\left\{[\alpha]-[x_0]+\sum_{r=1}^{n}\frac{[\beta_r][z-1]}{\left(1+\frac{\mu_r+j\nu_r}{2}\Delta t\right)z-\left(1-\frac{\mu_r+j\nu_r}{2}\Delta t\right)}\right\} \tag{5.18}$$

Therefore, using eq. (5.16) and eq. (5.18), we obtain

$$\begin{aligned}[X(z)]&=[X_1(z)]+[X_2(z)]\\&=[\alpha]\frac{z}{z-1}+\sum_{r=1}^{n}\frac{[\beta_r]z}{z-\left(\frac{1-\frac{\mu_r+j\nu_r}{2}\Delta t}{1+\frac{\mu_r+j\nu_r}{2}\Delta t}\right)}-\frac{z}{z+1}\left\{[\alpha]-[x_0]+\sum_{r=1}^{n}[\beta_r]\right\}\end{aligned} \tag{5.19}$$

then, remembering that $[x_0]=[\alpha]+\sum_{r=1}^{n}[\beta_r]$, eq. (5.19) may be written in the following form:

$$[X(z)]=[\alpha]\frac{z}{z-1}+\sum_{r=1}^{n}\frac{[\beta_r]z}{\left(\dfrac{1-\dfrac{\mu_r+j\nu_r}{2}\Delta t}{1+\dfrac{\mu_r+j\nu_r}{2}\Delta t}\right)} \tag{5.20}$$

Referring to appendix 1.2, we obtain to solution of eq. (5.9) in time domain, that is,

$$[x(t)]=[\alpha]+\sum_{r=1}^{n}[\beta_r]e^{-\mu_r' t}(\cos\nu_r' t-j\sin\nu_r' t) \tag{5.21}$$

Comparing eq. (5.21) with eq. (4.5), it can be said that the solution of original equation is distorted at the application of the modified method of Euler, in such a manner that μ_r' and ν_r' are in the solution instead of μ_r and ν_r.

5.3 The method of Runge-Kutta

Applying the method of Runge-Kutta to the set of equations (2.1), we get the corresponding set of difference equations:

$$\Delta^{\mathrm{I}}x_i=(b_i-\sum_j a_{ij}x_j)\Delta t$$

$$\Delta^{\mathrm{II}}x_i=\left[b_i-\sum_j a_{ij}\left(x_j+\frac{\Delta^{\mathrm{I}}x_j}{2}\right)\right]\Delta t$$

$$=\Delta^{\mathrm{I}}x_i-\frac{1}{2}(\sum_j a_{ij}\Delta^{\mathrm{I}}x_j)\Delta t$$

$$\Delta^{\mathrm{III}}x_i=\left[b_i-\sum_j a_{ij}\left(x_j+\frac{\Delta^{\mathrm{II}}x_j}{2}\right)\right]\Delta t$$

$$=\Delta^{\mathrm{I}}x_i-\frac{1}{2}(\sum_j a_{ij}\Delta^{\mathrm{I}}x_j)\Delta t+\frac{1}{4}(\sum_j\sum_k a_{ij}a_{jk}\Delta^{\mathrm{I}}x_k)\Delta t^2$$

$$\Delta^{\mathrm{IV}}x_i=[b_i-\sum_j a_{ij}(x_j+\Delta^{\mathrm{III}}x_j)]\Delta t$$

$$=\Delta^{\mathrm{I}}x_i-(\sum_j a_{ij}\Delta^{\mathrm{I}}x_j)\Delta t+\frac{1}{2}\left(\sum_j\sum_k a_{ij}a_{jk}\Delta^{\mathrm{I}}x_k\right)\Delta t^2-\frac{1}{4}(\sum_j\sum_k\sum_l a_{ij}a_{jk}a_{kl}\Delta^{\mathrm{I}}x_l)\Delta t^3,$$

therefore,

$$x_i(t+\Delta t)-x_i(t)=\frac{1}{6}(\Delta^{\mathrm{I}}x_i+2\Delta^{\mathrm{II}}x_i+2\Delta^{\mathrm{III}}x_i+\Delta^{\mathrm{IV}}x_i)$$

$$=b_i\Delta t-\frac{1}{2}(\sum_j a_{ij}b_j)(\Delta t)^2+\frac{1}{6}(\sum_j\sum_k a_{ij}a_{jk}b_k)(\Delta t)^3-\frac{1}{24}(\sum_j\sum_k\sum_l a_{ij}a_{jk}a_{kl}b_l)(\Delta t)^4$$

$$-\left\{(\sum_j a_{ij}x_j)\Delta t-\frac{1}{2}(\sum_j\sum_k a_{ij}a_{jk}x_k)(\Delta t)^2+\frac{1}{6}(\sum_j\sum_k\sum_l a_{ij}a_{jk}a_{kl}x_l)(\Delta t)^3\right.$$

$$\left.-\frac{1}{24}(\sum_j\sum_k\sum_l\sum_m a_{ij}a_{jk}a_{kl}a_{lm}x_l)(\Delta t)^4\right\}\quad(i=1,2,\cdots,n) \tag{5.22}$$

The set (5.22) may be written in matrix form as

$$[x(t+\Delta t)-x(t)]=[K]\{[B]-[A][x(t)]\} \tag{5.23}$$

where

$$[K]=[I]\Delta t-\frac{1}{2}[A](\Delta t)^2+\frac{1}{6}[A]^2(\Delta t)^3-\frac{1}{24}[A]^3(\Delta t)^4 \qquad (5.24)$$

[I] : Unit matrix.

Applying to eq. (5.23) the z-transformation yields

$$(z-1)[X(z)]=[K][B]\frac{z}{z-1}-[K][A][X(z)]+z[x_0]$$

where $[x_0]$ is the initial value of $[x(t)]$ and $[X(z)]$ denotes the z-transform. Putting $p=z-1$, eq. (5.25) may be written in the form

$$p[X(z)]+[K][A][X(z)]=z\left\{\frac{[K][B]}{p}+[x_0]\right\} \qquad 5(.26).$$

Remembering that eq. (4.5) is the solution of eq. (2.3), wc get the solution of eq. (5.26) in the form:

$$[X(z)]=[KA]^{-1}[K][B]\frac{z}{p} \\ -z\sum_{r=1}^{n}[KA]^{-1}\left\{\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\lambda_s'[I]-[KA]}{\lambda_r'-\lambda_s'}\right\}\{[KB]-[KA][x_0]\}\frac{1}{p+\lambda_r'} \qquad (5.27),$$

where λ_r' is the characteristic root of the matrix $[KA]$, that is

$$\det\{\lambda_r'[I]-[KA]\}=0 \quad (r=1,2,\cdots,n) \qquad (5.28).$$

Referring to appendix 2. the relation between λ_r' and λ_r, the characteristic root of the matrix $[A]$ is

$$\lambda_r'=(\Delta t)\lambda_r-\frac{(\Delta t)^2}{2}\lambda_r^2+\frac{(\Delta t)^3}{6}\lambda_r^3-\frac{(\Delta t)^4}{24}\lambda_r^4 \quad (r=1,2,\cdots,n) \qquad (5.29),$$

and referring to appendix 3,

$$\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\lambda_s'[I]-[KA]}{\lambda_s'-\lambda_r'}=\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\lambda_s[I]-[A]}{\lambda_s-\lambda_r} \qquad (5.30).$$

Since the matrix $[K]$ is the polynomials of the matrix $[A]$, $[K]$ and $[A]$ are commutative. Then we may write eq. (5.27) in the form:

$$[X(z)]=[A]^{-1}[B]\frac{z}{p}-\sum_{r=1}^{n}\left\{[A]^{-1}\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\lambda_s[I]-[A]}{\lambda_s-\lambda_r}\right. \\ \left.([B]-[A][x_0])\frac{1}{p+\lambda_r}\right\} \qquad (5.31).$$

Substituting $p=z-1$ into the above equation yields

$$[X(z)]=[\alpha]\frac{z}{z-1}+\sum_{r=1}^{n}\frac{[\beta_r]z}{z-(1-\lambda_r')} \qquad (5.32).$$

Referring to appendix 1.3, we have

$$[x(t)]=[\alpha]+\sum_{r=1}^{n}[\beta_r]\, e^{-\mu_r' t}(\cos \nu_r' t - j \sin \nu_r' t) \qquad (5.33).$$

Comparing eq. (5.33) with eq. (4.5), it can be said that the solution of the original equation is distorted at the application of the method of Runge-Kutta, in such a manner that μ_r' and ν_r' are in the solution instead of μ_r and ν_r.

6. *The Suitable Time Intervals when the Allowable Error is Given*

It is difficult to discuss on the problem of suitable time intervals in general. Here, we study this problem for some simple cases, from which general conclusion will be derived to some extent.

6.1 The Application on the Method of Euler to the Differential Equation of the First Order:

$$\frac{dx}{dt}+\frac{x}{T}=0\,.$$

Referring to appendix 1.1 (a) the error in time constant is given in the form:

$$\varepsilon=\frac{\Delta t}{2T}.$$

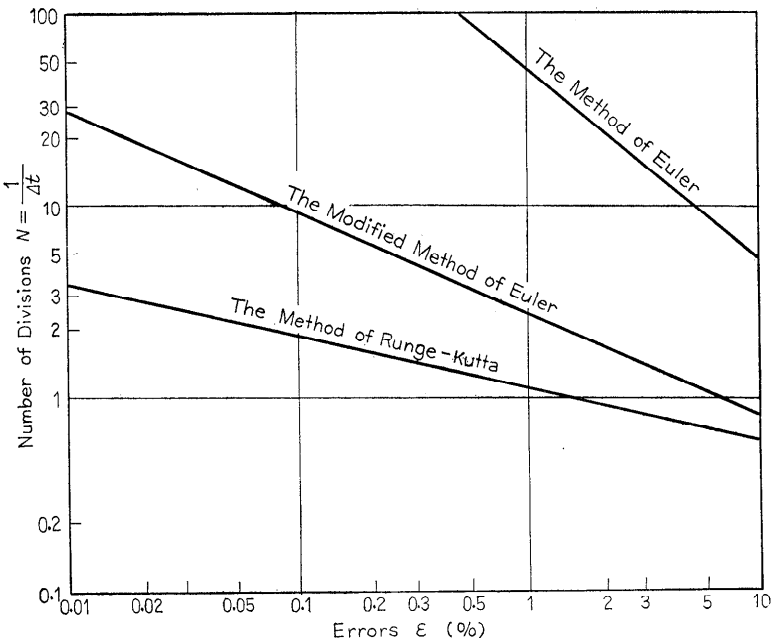

Fig. 1. The relation between the errors and the number of divisions, calculating the equation $\frac{dx}{dt}+x=0$.

In order to keep the error within the allowable error, it is required to adopt the time interval as

$$\Delta t \leq 2T\varepsilon_0 \times \frac{1}{100},$$

where ε_0 is the allowable error. The number of divisions corresponding to this time interval is given in the expression:

$$N=\frac{1}{\Delta t} \geq \frac{100}{2\varepsilon T}.$$

Fig. 1 shows the relation between the number of divisions N and the allowable error ε_0 in the case $T=1$.

6.2 The Application of the Method of Euler to the Equation:

$$\frac{d^2x}{dt^2}+\omega^2 x=0$$

Referring to appendix 1.1 (b) the rate of divergence and the error in frequency are given by the expressions:

$$\alpha=e^{\pi\omega\Delta t}-1, \quad \varepsilon_f=\frac{(\omega\Delta t)^2}{3}.$$

In order to keep the frequency error within ε_0 it is required to select the time interval as

$$\Delta t \leq \sqrt{\frac{\omega\varepsilon_0}{100}} \times \frac{1}{100}.$$

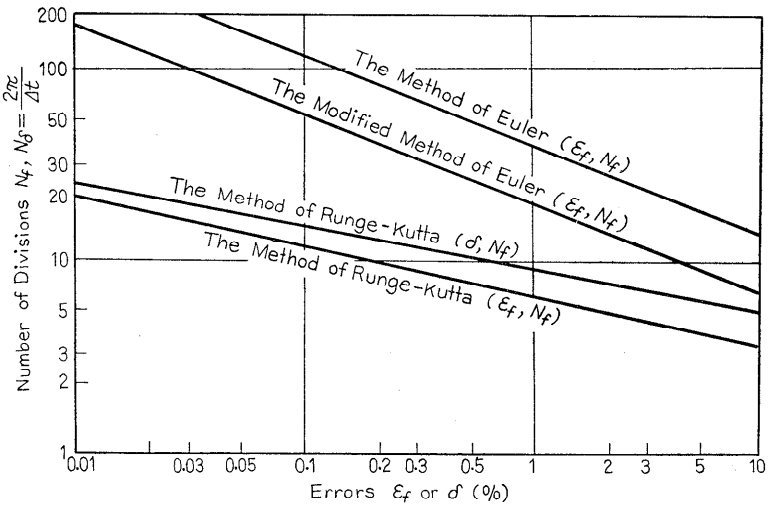

Fig. 2. The relation between the errors and the number of divisions, calculating the equation $\frac{d^2x}{dt}+x=0$.

Then, the number of divisions corresponding to the above time interval is

$$N_f=\frac{1}{\Delta t}\geq\frac{\omega}{\sqrt{3\varepsilon_0}}\times 10=\frac{2\pi f}{\sqrt{3\varepsilon_0}}\times 10=\frac{36.3f}{\sqrt{\varepsilon_0}}$$

Fig. 2 shows the relation between N and ε_f in the case $\omega=1$.

6.3 The Application of the Modified Method of Euler to the Equation:

$$\frac{dx}{dt}+\frac{x}{T}=0$$

Referring to appendix 1.2. (a), the error in time constant is given in the form:

$$\varepsilon=\frac{1}{12}\left(\frac{\Delta t}{T}\right)^2.$$

The length of time interval which keeps the error within the allowable error is given in the relation:

$$\Delta t\leq\sqrt{12\varepsilon_0}\,T\times 100\,.$$

Then, the number of divisions corresponding to the time interval is

$$N=\frac{1}{\Delta t}\geq\frac{100}{\sqrt{12\varepsilon_0}\,T}=28.8\,\frac{1}{\sqrt{\varepsilon_0}\,T}\,.$$

Fig. 1 shows the relation between N and ε in the case $T=1$.

6.4 The Application of the Modified Method of Euler to the Equation:

$$\frac{d^2x}{dt^2}+\omega^2x=0$$

Referring to appendix 1.2 (b), the rate of divergence and the frequency error are given in the form:

$$\alpha=0,\quad \varepsilon_f=\frac{1}{12}(\omega\Delta t)^2.$$

The number of divisions which keeps the frequency error within ε_0 is given in the relation:

$$N=\frac{1}{\Delta t}\geq\sqrt{\frac{100}{12\varepsilon_0}}\,\omega=\frac{2\pi f\times 10}{\sqrt{12\varepsilon}}=\frac{18.2f}{\sqrt{\varepsilon}}\,.$$

Fig. 2 shows the above relation in the case $\omega=1$.

6.5 The Application of the Method of Runge-Kutta to the Equation:

$$\frac{dx}{dt}+\frac{x}{T}=0$$

Referring to appendix 1.3 (a), the error of time constant is written in the form:

$$\varepsilon=\frac{1}{120}\left(\frac{\Delta t}{T}\right)^4\exp\left(\frac{\Delta t}{T}\right).$$

The number of division which keeps the error within ε_0 is given in the relation:

$$N\exp\left(-\frac{1}{4N}\right)\geq\frac{1}{4\sqrt{1.2\varepsilon_0 T}}=\frac{0.96}{4\sqrt{\varepsilon_0 T}}\quad\left(N=\frac{1}{\Delta t}\right).$$

Fig. 1 shows the above relation in the case $T=1$.

6.6 The Application of the Method of Runge-Kutta to the Equation:

$$\frac{d^2x}{dt^2}+\omega^2x=0$$

Referring to appendix 1.3 (b), the rate of attenuation and the frequency error are given in the forms:

$$\delta=\frac{1}{144}(\omega\Delta t)^5,\quad \varepsilon_f=\frac{(\omega\Delta t)^4}{120}.$$

The length of time interval which keeps δ within δ_0 is given in the relation:

$$\Delta t\leq\sqrt{\frac{144\delta_0}{100\omega}}\times\frac{5}{\omega},$$

then, the number of divisions corresponding to the above time interval is

$$N_\delta=\frac{1}{\Delta t}>\sqrt{\frac{100\omega}{144\delta_0}}\times\omega=8.4\sqrt{\frac{f}{\delta_0}}f$$

The length of time interval which keeps the frequency error within ε_0 is given in the relation:

$$\Delta t\leq\sqrt{\frac{120\varepsilon_0}{100}}\times\frac{1}{\omega}$$

The number divisions corresponding to the above the intervals is

$$N_f=\frac{1}{\Delta t}\geq\frac{\omega}{4\sqrt{1.2\varepsilon_0}}=\frac{6}{4\sqrt{\varepsilon_0}}f.$$

The relation between N_δ and δ and that between N_f and ε_f are shown in Fig. 2.

7. *Conclusions*

The characteristic of the solutions of the simultaneous linear ordinary differential equations with constant coefficients can be described by such characteristic constants as the time-constants and the frequencies. With special regard to these characteristic constants, the authors studied the problem of errors due to numerical calculation. The results of this study give the suitable time intervals when the allowable errors are given.

In many engineering problems, it is adequate to keep the errors within 1 percent. The suitable time intervals for above requirements are shown in Table 1. In the section 1, we assumed that B is the constant matrix, however, the results of this study may be extended to the general cases

in which elements of B are given in the explicit functions of time t.

Table 1. Suitable time intervals required to keep the errors of time constants and that of Frequencies within 1 percent.

The method of numerical calculation	The suitable time interval Δt
The modified method of Euler	$\min\left(\frac{T\min}{5}, \frac{1}{20f\min}\right)$
The method of Runge-Kutta	$\min\left(\frac{T\min}{2}, \frac{1}{10f\min}\right)$

T min: The Minimum time constant in the system.
f min: The Minimum frequency in the system.

The paper discusses only about the truncation errors; the round-off errors are not discussed. However, when the time intervals for computation are determined by the above criteria, the round-off errors are much smaller than the truncation errors, and it is sufficient to consider only about the latter errors.

References

1. Levy, H. and Baggott, E. A.: Numerical Solutions of Differential Equations, pp. 91–108, New York, Dover Publications (1950).
2. Hayashi, S.: Surges on Transmission Systems, pp. 23–40, Kyoto, Denkishoin (1955).
3. Beckenback, E. F.: Modern Mathematics for the Engineer, pp. 325–328, New York, Mc-Graw-Hill, J. R. and Franklin, G. F. (1956).
4. Ragazzini, J. R. and Franklin, G. F.: Sampled Data Control Systems, New York, Mc-Graw-Hill (1958).

Appendix 1 The inverse z-transformation of $X(z)$

1.1 The Method of Euler

——— the inverse-transformation of the equation $X(z)=\dfrac{z}{z-(1-\mu\Delta t-\nu\Delta t)}$

Putting

$$e^{-(\mu'+j\nu')\Delta t}=1-\mu\Delta t-j\nu\Delta t,$$

that is,

$$\mu'+j\nu'=-\frac{1}{\Delta t}\ln(1-\mu\Delta t-j\nu\Delta t)$$

we may write the $X(z)$ in the following form:

$$X(z)=\frac{z}{z-e^{-(\mu'+j\nu')\Delta t}}$$

Then, applying the inverse z-transformation to the above equation yields

$$x(t)=e^{-(\mu'+j\nu')t}$$

1.1 (a) In the case: $\nu=0$.

In this case, we have

$$\mu'+j\nu'=-\frac{1}{\Delta t}\ln(1-\mu\Delta t)$$

that is,

$$\mu' = -\frac{1}{\Delta t}\ln(1-\mu\Delta t) \simeq \mu\left(1+\frac{\mu}{2}\Delta t\right),$$
$$\nu' = 0$$

The time constant T is given by the inverse of μ then

$$T' = \frac{1}{\mu'} = \frac{1}{\mu\left(1+\frac{\mu}{2}\Delta t\right)} \simeq T\left(1-\frac{\Delta t}{2T}\right)$$

In this case, T' is in the solution instead of the exact time constant T. Then the error in the time constant is

$$\varepsilon = \frac{\Delta t}{2T}$$

1.1 (b) In the case: $\mu=0$.

In this case, we have

$$\mu' + j\nu' = -\frac{1}{\Delta t}\ln(1-j\nu\Delta t) \simeq -\frac{1}{2}\nu^2\Delta t + j\nu\left[1-\frac{1}{3}(\nu\Delta t)^2\right].$$

Although the exact solution is $e^{j\nu t}$, the solution obtained by the method of Euler is

$$e^{-\frac{1}{2}[\nu\Delta t\cdot\nu t]}e^{j\nu\left[1-\frac{1}{3}(\nu\Delta t)^2\right]t}$$

In this case, the rate of divergence and the frequency error are written in the forms:

$$\alpha = e^{\pi(\nu\Delta t)} - 1,$$
$$\varepsilon_f = \frac{1}{3}(\nu\Delta t)^2.$$

1.2 The modified Method of Euler

——— the inverse z-transformation of the equation

$$X(z) = \frac{z}{z - \dfrac{\left(1-\dfrac{\mu+j\nu}{2}\Delta t\right)}{\left(1+\dfrac{\mu+j\nu}{2}\Delta t\right)}}$$

Putting

$$e^{-(\mu'+j\nu')\Delta t} = \frac{1-\dfrac{\mu+j\nu}{2}\Delta t}{1+\dfrac{\mu+j\nu}{2}\Delta t},$$

that this

$$\mu' + j\nu' = -\frac{1}{\Delta t}\ln\left(\frac{1-\dfrac{\mu+j\nu}{2}\Delta t}{1+\dfrac{\mu+j\nu}{2}\Delta t}\right),$$

we may write $X(z)$ in the form:

$$X(z) = \frac{z}{z - e^{-(\mu'+j\nu')\Delta t}}.$$

Then, the inverse z-transform of $X(z)$ is

$$x(t) = e^{-(\mu'+j\nu')t}.$$

1.2 (a) In the case: $\mu\neq 0$, $\nu=0$.

In this case, we have

$$\mu'+j\nu'=-\frac{1}{\Delta t}\ln\left(\frac{1-\frac{\mu}{2}\Delta t}{1+\frac{\mu}{2}\Delta t}\right)$$

that is

$$\mu'=-\frac{1}{\Delta t}\ln\left(\frac{1-\frac{\mu}{2}\Delta t}{1+\frac{\mu}{2}\Delta t}\right)\simeq\mu\left[1+\frac{1}{12}(\mu\Delta t)^2\right],\quad \nu'=0.$$

The time constant is given by the inverse of μ; then we get relation about time constant:

$$T'=\frac{1}{\mu'}=\frac{1}{\mu\left[1+\frac{1}{12}(\mu\Delta t)^2\right]}\simeq T\left[1-\frac{1}{12}\left(\frac{\Delta t}{T}\right)^2\right].$$

Then, the error in time constant is given in the form:

$$\varepsilon=\frac{1}{12}\left(\frac{\Delta t}{T}\right)^2.$$

1.2(b) In the case: $\mu=0$, $\nu\neq0$.

In this case, we have

$$\mu'+j\nu'=-\frac{1}{\Delta t}\ln\left(\frac{1-j\frac{\nu}{2}\Delta t}{1+j\frac{\nu}{2}\Delta t}\right)$$

that is,

$$\mu'=0,\ \nu'=\frac{2}{\Delta t}\tan^{-1}\frac{(\nu\Delta t)}{2}.$$

Then, the rate of divergence α and the frequency error ε_f are given in the forms:

$$\alpha=0,\ \varepsilon_f=\frac{1}{12}(\nu\Delta t)^2.$$

1.3 The Method of Runge-Kutta

——— the inverse z-transformation of the equation

$$X(z)=\frac{z}{z-(1-\lambda')}$$

Putting

$$e^{-(\mu'+j\nu')\Delta t}\Delta t=\sum_{n=0}^{4}(-1)^n\frac{(\mu+j\nu)^n(\Delta t)^n}{n!},$$

we may write $X(z)$ in the form:

$$X(z)=\frac{z}{z-e^{-(\mu'+j\nu')\Delta t}}.$$

Then, the inverse z-transform of $X(z)$ is

$$x(t)=e^{-(\mu'+j\nu')t}.$$

1.3(a) In the case: $\mu\neq0$, $\nu=0$.

In this case, we have

$$e^{-(\mu'+j\nu')\Delta t}=\sum_{n=0}^{4}(-1)^n\frac{(\mu\Delta t)^n}{n!}\simeq e^{-(\mu\Delta t)}+\frac{(\mu\Delta t^5)}{120},$$

that is,

$$\mu'=-\frac{1}{\Delta t}\ln\left(e^{-\mu\Delta t}+\frac{(\mu\Delta t)^5}{120}\right)\simeq\mu\left(1-\frac{(\mu\Delta t)^4}{120}\right)e^{\mu\Delta t},$$

$$\nu'=0.$$

The time constant T is given by the inverse of μ; then we have

$$T'=\frac{1}{\mu'}=\frac{1}{\mu\left\{1-\frac{(\mu\Delta t)^4}{120}e^{\mu\Delta t}\right\}}\simeq T\left[1+\frac{1}{120}\left(\frac{\Delta t}{T}\right)^4 e^{\frac{\Delta t}{T}}\right]$$

Then, the error in time constant is given in the form:

$$\varepsilon=\frac{1}{120}\left(\frac{\Delta t}{T}\right)^4 e^{\frac{\Delta t}{T}}$$

1.2(b) In the case: $\mu=0$, $\nu\neq 0$

$$\begin{aligned}
e^{-(\mu'+j\nu')\Delta t}&=\sum_{n=0}^{4}(-1)^n\frac{(j\nu)^n}{n!}(\Delta t)^n\\
&\simeq e^{-j\nu\Delta t}+\frac{(j\nu\Delta t)^5}{120}-\frac{(j\nu\Delta t)^6}{720}\\
&=e^{-j\nu\Delta t}\left[1+\left\{\frac{(j\nu\Delta t)^5}{120}-\frac{(j\nu\Delta t)^6}{720}\right\}e^{j\nu\Delta t}\right]\\
&\simeq e^{-j\nu\Delta t}\left[1+\frac{(j\nu\Delta t)^5}{129}+\frac{(j\nu\Delta t)^6}{144}\right]\\
&\simeq e^{-j\nu\Delta t}\exp\left[\frac{(j\nu\Delta t)^5}{120}+\frac{(\nu\Delta t)^6}{144}\right]\\
&\simeq\exp\left\{-j\nu\Delta t\left(1-\frac{(\nu\Delta t)^4}{120}-j\frac{(\nu\Delta t)^5}{144}\right)\right\},
\end{aligned}$$

that is

$$\mu'\simeq\frac{\nu}{144}(\nu\Delta t)^5,$$

$$\nu'\simeq\left[1+\frac{(\nu\Delta t)^4}{120}\right].$$

Although the exact solution is $e^{j\nu t}$, the solution obtained by the method of Runge-Kutta is given in the expression as

$$e^{-\frac{(\nu\Delta t)^5}{144}\nu t}e^{j\nu\left[1-\frac{(\nu\Delta t)^4}{120}\right]t}$$

Then, the rate of attenuation δ and the frequency error ε_f are given in the forms:

$$\delta=1-e^{-\frac{(\nu\Delta t)^5}{72}\pi},\quad \varepsilon_f=\frac{(\nu\Delta t)^4}{120}.$$

Appendix 2 Characteristic Roots of the Matrix $[KA]$

λ and λ' denote the characteristic roots of the matrix $[A]$ and $[KA]$ respectively. The matrix $[KA]$ is given in the foam as

$$[KA]=[\Delta t][A]-\frac{(\Delta t)^2}{2!}[A]^2+\frac{(\Delta t)^3}{3!}[A]^3-\frac{(\Delta t)^4}{4!}[A]^4$$

Applying Frobenius's theorem, which states that if λ is the characteristic root of the square matrix $[A]$, the characteristic root of $P(A)$, a polynomial of $[A]$, is given by $P(\lambda)$, λ_r' can be written in the form:

$$\lambda_r'=(\Delta t)\lambda_r-\frac{(\Delta t)^2}{2!}\lambda_r^2+\frac{(\Delta t)^3}{3!}\lambda_r^3-\frac{(\Delta t)^4}{4!}\lambda_r^4$$

Appendix 3 The Proof of the Equation

$$\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\lambda_s'[I]-[KA]}{\lambda_s'-\lambda_r'}=\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\lambda_s[I]-[A]}{\lambda_s-\lambda_r}$$

Using the result of appendix 2, we have the equation as

$$\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\lambda_s'[I]-[KA]}{\lambda_s'-\lambda_r'}$$

$$=\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\left\{(\Delta t)\lambda_s-\frac{(\Delta t)^2\lambda_s^2}{2!}+\frac{(\Delta t)^3\lambda_s^3}{3!}-\frac{(\Delta t)^4\lambda_s^4}{4!}\right\}[I]-\left\{(\Delta t)[A]-\frac{(\Delta t)^2[A]^2}{2!}-\frac{(\Delta t)^3[A]^3}{3!}-\frac{(\Delta t)^4[A]^4}{4!}\right\}}{(\Delta t)\lambda_s-\frac{(\Delta t)^2\lambda_s^2}{2!}+\frac{(\Delta t)^3\lambda_s^3}{3!}-\frac{(\Delta t)^4\lambda_s^4}{4!}-\left\{(\Delta t)\lambda_r-\frac{(\Delta t)^2\lambda_r^2}{2!}-\frac{(\Delta t)^3\lambda_r^3}{3!}-\frac{(\Delta t)^4\lambda_r^4}{4!}\right\}}$$

$$=\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\lambda_s[I]-[A]}{\lambda_s-\lambda_r}[\delta_s],$$

where

$$[\delta_s]=\frac{[I]-\frac{1}{2}(\lambda_s[I]+[A])\Delta t+\frac{1}{6}(\lambda_s[I]+\lambda_s[A]+[A]^2)\Delta t^2}{1+\frac{1}{2}(\lambda_s+\lambda_r)\Delta t+\frac{1}{6}(\lambda_s+\lambda_s\lambda_r+\lambda_r^2)\Delta t^2}$$

$$\frac{-\frac{1}{24}(\lambda_s^3[I]+\lambda_s^2[A]+\lambda_s[A]^2+[A]^3)\Delta t^3}{-\frac{1}{24}(\lambda_s^3+\lambda_s^2\lambda_r+\lambda_s\lambda_r^2+\lambda_r^3)\Delta t}.$$

Using the identity as

$$\frac{[A]}{a}\equiv\frac{[A]-a[I]}{a}+[I]$$

$[\delta_s]$ may be written in the form:

$$[\delta_s]=\frac{\frac{1}{2}(\lambda_r[I]-[A])\Delta t-\frac{1}{6}(\lambda_r^2[I]+\lambda_s\lambda_r[I]-\lambda_s[A]-[A]^2)}{1-\frac{1}{2}(\lambda_s+\lambda_r)\Delta t+\frac{1}{6}(\lambda_s^2+\lambda_s\lambda_r+\lambda_r^2)\Delta t^2}$$

$$\frac{+\frac{1}{24}(\lambda_r^3[I]+\lambda_r^2\lambda_s[I]+\lambda_r\lambda_s^2[I]-\lambda_s^2[A]-\lambda_s[A]^2-[A]^3)}{\frac{1}{24}(\lambda_s^3+\lambda_s^2\lambda_r+\lambda_s\lambda_r^2+\lambda_r^3)\Delta t^3}+[I]$$

that is

$$[\delta_s]=(\lambda_r[I]-[A])F_s(A)+[I],$$

where $F_s(A)$ is a polynominal of finite degree in the matrix $[A]$. Then, we have

$$\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\lambda_s'[I]-[KA]}{\lambda_s'-\lambda_r'}=\prod_{s\neq r}^{s=1,2,\cdots,n}\left[\frac{\lambda_s[I]-[A]}{\lambda_s-\lambda_r}\{\lambda_r[I]-[A])F_s(A)+[I]\}\right]$$

$$=\left[\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{1}{\lambda_s-\lambda_r}\right]\left[\prod_{s=1}^{n}\{\lambda_s[I]-[A]\}\right]F(A)+\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\lambda_s[I]-[A]}{\lambda_s-\lambda_r}\Bigg],$$

where $F(A)$ is a polynominal of finite degree in $[A]$.

In the above equation, λ_r' are the characteristic roots of the matrix $[A]$. Then, applying Cayley-Hamilton's theorem yields

$$(\lambda_1[I]-[A])(\lambda_2[I]-[A])\cdots(\lambda_n[I]-[A])=0,$$

therefore, we have the equation as

$$\prod_{t\neq s}^{s=1,2,\cdots,n}\frac{\lambda_s'[I]-[KA]}{\lambda_s'-\lambda_r'}=\prod_{s\neq r}^{s=1,2,\cdots,n}\frac{\lambda_s[I]-[A]}{\lambda_s-\lambda_r}.$$

Appendix 4. Truncation errors vs round-off errors

An ordinary differential equation of the harmonic oscillation type $\frac{dx^2}{dt^2}+x=0$ under the initial condition $\left\{x=0,\ \frac{dx}{dt}=0\right\}$ is solved by means of a digital computor (5 digits. floating decimal) using the method of modified Euler and the result of the computation is shown in Table.

Table II. Numerical Errors $\left(\text{Modified Euler method } N=20,\ \Delta t=\frac{2\pi}{20}\right)$

	$\sqrt{x^2+y^2}$ amplitude			$\tan^{-1} x/y$ (deg) (Phase)			Phase error (deg)		
	True value	Computed value	Theoretical value	True value	Computed value	Theoretical value	Computed value (A)	Theoretical value (B)	$\frac{B-A}{B}$ (%)
0.31416	1.0000	1.0000	1.0000	18	17.854	17.852	0.146	0.148	2.7
0.62832	〃	〃	〃	36	35.708	35.704	0.292	0.296	1.3
0.94248	〃	〃	〃	54	53.562	53.556	0.438	0.446	1.8
1.2566	〃	〃	〃	72	71.417	71.408	0.583	0.592	1.5
1.5708	〃	〃	〃	90	89.271	89.260	0.729	0.740	〃
1.8850	〃	〃	〃	108	107.13	107.11	0.87	0.89	2.2
2·1991	〃	〃	〃	126	124.98	121.96	1.02	1.04	1.9
2.5133	〃	〃	〃	144	142.83	142.82	1.17	1.18	0.9
2.8274	〃	〃	〃	162	160.69	160.67	1.31	1.33	1.5
3.1416	〃	〃	〃	180	178.54	178.52	1.46	1.48	1.4
3.4558	〃	〃	〃	198	196.40	196.37	1.60	1.63	1.8
3.7699	〃	〃	〃	216	214.25	214.22	1.75	1.78	1.6
4.0841	〃	〃	〃	234	232.10	232.08	1.90	1.92	1.0
4.3982	〃	〃	〃	252	249.96	249.93	2.04	2.07	1.5
4.7124	〃	〃	〃	270	267.81	267.78	2.19	2.22	1.4
5.0266	〃	〃	〃	288	285.67	285.63	2.33	2.37	1.7
5.3407	〃	〃	〃	306	303.52	303.48	2.48	2.52	1.6
5.6549	〃	〃	〃	324	321.38	321.34	2.62	2.66	1.5
5.9691	〃	〃	〃	342	339.23	339.19	2.77	2.81	1.4
6.2832	〃	〃	〃	360	357.08	357.04	2.92	2.96	〃

As shown in Table II, the computed value is nearly equal to the theoretical value and this means that the round-off errors are much smaller than the truncation errors.