A Stabilizing Device for Unstable Numerical Solutions of Ordinary Differential Equations—Design Principle and Application of a "Filter"

MASAO IRI*

Introduction

Numerical instability has been counted as a great disadvantage of those numerical methods of solution of ordinary differential equations which are of high precision in spite of their simplicity [1]. Among them we find the midpoint rule and Milne's method. A number of methods have been proposed which are free from numerical instability [2], and for Milne's method, in particular, a device to suppress the instability was designed by W. E. Milne himself and R. R. Reynolds [3]. Meanwhile, detailed analysis of errors of various kinds included in numerical solutions by linear multistep methods was published in the form of a monograph by P. Henrici [4]. Professor S. Moriguti of the University of Tokyo suggested the idea of removing the extraneous error components causing numerical instability by the help of a kind of averaging operation.

In this paper, calling such an averaging operator a "filter", we shall show the existence of a filter of arbitrary precision for any multistep method, formulate the design procedure and illustrate the efficacy of a filter by examples.

1. *Linear Multistep Methods*

The results known about linear multistep methods for ordinary differential equations may be summarized as follows [4].

In solving the initial value problem of the system of equations

$$\frac{dy^i(x)}{dx}=f^i(x, y^j(x)) \qquad (i, j=1, \cdots, m), \tag{1}$$

$$y^i(a)=\eta^i \qquad (i=1, \cdots, m) \tag{2}$$

by means of a linear k-step method defined by the formula

$$\alpha_k {y_{n+k}}^i+\alpha_{k-1}{y_{n+k-1}}^i+\cdots+\alpha_1{y_{n+1}}^i+\alpha_0{y_n}^i$$
$$=h(\beta_k {f_{n+k}}^i+\beta_{k-1}{f_{n+k-1}}^i+\cdots+\beta_1{f_{n+1}}^i+\beta_0{f_n}^i$$
$$(i=1, \cdots, m;\ n=0, 1, 2, \cdots) \tag{3}$$

This paper first appeared in Japanese in Joho Shori (the Journal of the Information Processing Society of Japan), Vol. 4, No. 5 (1963), pp. 249–260.

* Faculty of Engineering, University of Tokyo.

where

$$x_n=a+nh \qquad (n=0, 1, 2, \cdots), \tag{4}$$

$y^i(x)$ is the exact solution, $y_n{}^i$ is the value at $x=x_n$ of the numerical solution and

$$f_n{}^i=f^i(x_n, y_n{}^j). \tag{5}$$

We denote the error included in $y_n{}^i$ by

$$e_n{}^i=y_n{}^i-y^i(x_n) \tag{6}$$

and put

$$\left.\begin{aligned}\rho(\zeta)&=\alpha_k\zeta^k+\alpha_{k-1}\zeta^{k-1}+\cdots+\alpha_1\zeta+\alpha_0,\\ \sigma(\zeta)&=\beta_k\zeta^k+\beta_{k-1}\zeta^{k-1}+\cdots+\beta_1\zeta+\beta_0.\end{aligned}\right\} \tag{7}$$

Then, for (3) to be an approximation to (1), ρ and σ in (7) must satisfy

$$\rho(1)=0, \qquad \rho'(1)=\sigma(1)\neq 0. \tag{8}$$

If we regard ζ as the operator of increasing the value of the suffix n by one, then (3) can be rewritten as

$$\rho(\zeta)y_n{}^i=h\sigma(\zeta)f_n{}^i. \tag{9}$$

The general solution of (3) in the vicinity of $(x_{n_0}, y_{n_0}{}^i)$ may be written as follows under appropriate conditions.

$$\begin{aligned}y_{n_0+n}{}^i\sim&[a_0{}^i+a_1{}^i(nh)+a_2{}^i(nh)^2+\cdots]_{(1)}\\&+\sum_{\mu=2}^{s}\zeta_\mu{}^n[b_{\mu 0}{}^i+b_{\mu 1}{}^i(nh)+b_{\mu 2}{}^i(nh)^2+\cdots]_{(\mu)}\\&(i=1, \cdots, m;\ n=\cdots, -2, -1, 0, 1, 2, \cdots),\end{aligned} \tag{10}$$

where ζ_μ $(\mu=1, 2, \cdots, s)$ is the root of multiplicity p_μ $(\mu\geqq 1)$ of $\rho(\zeta)=0$ and we put, in particular,

$$\zeta_1=1 \qquad (p_\mu=1) \tag{11}$$

in view of (8). Hence

$$\sum_{\mu=2}^{s}p_\mu=k-1. \tag{12}$$

$[\cdots\cdots]_{(1)}$ is the principal component approximating the solution of the problem, while $\zeta_\mu{}^n[\cdots\cdots]_{(\mu)}$'s $(\mu\neq 1)$ are extraneous error components. Furthermore, the respective components of (10) have the following asymptotic expressions.

$$[\cdots\cdots]_{(1)}-z^i(x_{n_0+n})\sim h^p e_1{}^i(x_{n_0+n}), \tag{13}$$

$$\left.\begin{aligned}\frac{dz^i(x)}{dx}&=f^i(x, z^j(x)), \qquad z^i(x_{n_0})=y_{n_0}{}^i,\\ \frac{de_1{}^i(x)}{dx}&=\sum_{j=1}^{m}g_j{}^i(x)e_1{}^j(x)-C\left(\frac{d}{dx}\right)^{p+1}z^i(x),\end{aligned}\right\} \tag{14}$$

where $g_j{}^i(x)$ (function of x), C (real constant) and p (positive integer) are defined by

$$\left.\begin{aligned} g_j{}^i(x)&=\left[\frac{\partial f^i(x, z^l)}{\partial z^j}\right]_{z^l=z^l(x)},\\ \{\rho(e^h)-h\sigma(e^h)\}/\sigma(1)&=Ch^{p+1}+O(h^{p+2}). \end{aligned}\right\} \tag{15}$$

$$\left.\begin{aligned} &[\cdots\cdots]_{(\mu)}\sim e_\mu{}^i(x_{n_0+n}),\\ &\left(\frac{d}{dx}\right)^{p_\mu}e_\mu{}^i(x)=0 \qquad (p_\mu\leqq q_\mu),\\ &\left(\frac{d}{dx}\right)^{p_\mu}e_\mu{}^i(x)=\lambda_\mu\left(\frac{d}{dx}\right)^{q_\mu}\left(\sum_{j=1}^{m}g_j{}^i(x)e_\mu{}^j(x)\right) \quad (p_\mu=q_\mu+1), \end{aligned}\right\} \tag{16}$$

where ζ_μ is assumed to be a root of multiplicity $p_\mu\ (\geqq 1)$ of $\rho(\zeta)=0$ and, at the same time, a root of multiplicity $q_\mu\ (\geqq 0)$ of $\sigma(\zeta)=0$ $(q_\mu\geqq p_\mu-1)$, and λ_μ is the "growth parameter" associated with ζ_μ:

$$\lambda_\mu=\frac{\sigma^{(q_\mu)}(\zeta_\mu)/q_\mu!}{\zeta_\mu{}^{p_\mu-q_\mu}\rho^{(p_\mu)}(\zeta_\mu)/p_\mu!}. \tag{17}$$

As the origin of the extraneous error components corresponding to the roots $\zeta_\mu\ (\mu\neq 1)$ the following facts are to be considered.

1° the starting values of numerical solution $y_n{}^i\ (n=0, 1, 2, \cdots, k-1)$ have already such components.

2° Round-off errors or the effect of interaction among different components which are neglected in the approximation of (13)~(16) arouse them.

2. *Instability*

The extraneous error component corresponding to $\zeta_\mu\ (\mu\neq 1)$ causes the instability phenomenon when

(a) $|\zeta_\mu|>1$ (in this case the term $\zeta_\mu{}^n[\cdots\cdots]_{(\mu)}$ on the right-hand side of (10) grows as fast as $\zeta_\mu{}^n$);

(b) $|\zeta_\mu|=1$ and $p_\mu\geqq 2$ (in this case $|\zeta_\mu{}^n[\cdots\cdots]_{(\mu)}|\sim|e_\mu{}^i(x)|$ may contain the term growing as fast as $x^{p_\mu-1}$);

or

(c) $|\zeta_\mu|=1$, $p_\mu=1$, $q_\mu=0$ and the matrix $[\lambda_\mu g_j{}^i(x)]$ has an eigenvalue with positive real part (in this case $|\zeta_\mu{}^n[\cdots\cdots]_\mu|\sim|e_\mu{}^i(x)|$ grows so as to satisfy

$$\frac{de_\mu{}^i(x)}{dx}=\lambda_\mu\left(\sum_{j=1}^{m}g_j{}^i(x)e_\mu{}^j(x)\right). \tag{18}$$

The instability of the case (a) or (b) appears independently of the properties of the equation to be solved and is called "strong instability", whereas the instability of the case (c), dependent on the property of the functions $f^i(x, y^j)$, is called "numerical instability".

3. *Design Principle and Formulae of a Filter*

Let us consider how it is possible to apply a linear operator

$$Y: \quad \{y_n{}^i\} \longrightarrow \{y_n{}^{i*}\} \tag{19}$$

to the sequence $\{y_n{}^i\}$ expressed as in (10) in order to obtain the new sequence $\{y_n{}^{i*}\}$ in which the extraneous error components are as sufficiently suppressed as we want while the principal component, i.e. $[\cdots\cdots]_{(1)}$ in (10), remains as less affected. We shall call the operator Y a "filter". Y may be put in the form

$$Y(\zeta)=\zeta^{-K}P(\zeta), \tag{20}$$

where ζ is regarded as the operator of increasing n by one, K an integer and $P(\zeta)$ a polynomial. Next, we assume that,

〈1〉 for each $\mu(\neq 1)$, the extraneous component $\zeta_\mu{}^n[\cdots\cdots]_{(\mu)}$ in $\{y_n{}^i\}$ should be removed up to $O(h^{M_\mu-1})$, and in $\{y_n{}^{i*}\}$ only $O(h^{M_\mu})$ should remain;

〈2〉 the principal component $[\cdots\cdots]_{(1)}$ in $\{y_n{}^i\}$ should be preserved unaffected up to $O(h^N)$, and the affected quantity should be $O(h^{N+1})$;

〈3〉 integers K, N and M_μ's ($\mu=2, \cdots, s$) are preassigned.

Then, from 〈1〉 and the well-known theorems in calculus of difference, it follows that $Y(\zeta)$ should be factored as

$$\left.\begin{aligned} Y(\zeta)&=\zeta^{-K}\tau(\zeta)\omega(\zeta), \\ \tau(\zeta)&=\prod_{\mu=2}^{s}(\zeta-\zeta_\mu)^{M_\mu}, \\ \omega(\zeta)&=\text{a polynomial in } \zeta. \end{aligned}\right\} \tag{21}$$

Application of the Y of (21) to $\{y_n{}^i\}$ yields

$$\begin{aligned} Y(\zeta)y_{n_0+n}{}^i \sim\ & Y(\zeta)[\cdots\cdots]_{(1)} \\ & +\sum_{\mu=2}^{s}[b_{\mu M_\mu}{}^i\zeta_\mu{}^{n+M_\mu}Y^{(M_\mu)}(\zeta_\mu)h^{M_\mu}+O(h^{M_\mu+1})], \end{aligned} \tag{22}$$

where

$$b_{\mu M_\mu}{}^i \sim \frac{1}{M!}\left(\frac{d}{dx}\right)^{M_\mu} e_\mu{}^i(x). \tag{23}$$

Similarly, from 〈2〉 it follows that $Y(\zeta)-1$ should contain the factor $(\zeta-1)^{N+1}$. Hence $Y(\zeta)$ obtains another expression

$$Y(\zeta)=1+\zeta^{-K}(\zeta-1)^{N+1}\psi(\zeta) \tag{24}$$

with some rational function $\psi(\zeta)$ which is regular at $\zeta=1$.

It is obvious that the simplest of the polynomials $\omega(\zeta)$ that satisfy (21) and (24) can be obtained by first expanding

$$\frac{\zeta^K}{\tau(\zeta)}=\frac{\{1+(\zeta-1)\}^K}{\tau(1+(\zeta-1))} \tag{25}$$

in the power series in $(\zeta-1)$ and then truncating the terms beyond $(\zeta-1)^N$. If we adopt the simplest polynomial as $\omega(\zeta)$, denoting the coefficient of $(\zeta-1)^{N+1}$ (i.e. of the first truncated term) in the expansion of (25) by $\tau_{N+1}/(N+1)!$, we have

$$\begin{aligned} &Y(\zeta)y_{n_0+n}{}^i \\ &\quad\sim[a_0{}^i+a_1{}^i(nh)+a_2{}^i(nh)^2+\cdots+a_N{}^i(nh)^N \\ &\qquad+(1+\underline{\tau_{N+1}\tau(1))a_{N+1}{}^i(nh)^{N+1}}+O(h^{N+2})] \\ &\qquad+\sum_{\mu=2}^{s}[\underline{b_{\mu M_\mu}{}^i\zeta_\mu{}^{n+M_\mu}Y^{(M_\mu)}(\zeta_\mu)h^{M_\mu}}+O(h^{M_\mu+1})]\,. \end{aligned} \tag{26}$$

The underlines in (26) indicate the leading parts, respectively, of the "disturbance of the principal compnent" and the "residual extraneous error components" after application of $Y(\zeta)$.

$Y(\zeta)$ thus determined consists of the terms from ζ^{-K} to $\zeta^{N+\sum_{\mu=2}^{s}M_\mu-K}$, and, obviously, satisfies the conditions $\langle 1\rangle$ and $\langle 2\rangle$.

In the following several examples of filters are shown which are for the multistep methods with $\rho(\zeta)=\zeta^2-1$ ($\zeta_1=1$, $\zeta_2=-1$), where M stands for M_2.

○ $M=1$, $N=1$:—

$$\begin{cases}\tau(\zeta)=\zeta-\zeta_2=\zeta+1=2+\Delta=2\left(1+\dfrac{\Delta}{2}\right),\\ [\tau(\zeta)]^{-1}=\dfrac{1}{2}\left(1-\dfrac{\Delta}{2}+\dfrac{\Delta^2}{4}+\cdots\right);\end{cases}$$

$K=0$:

$$Y(\zeta)=\frac{1}{4}(3-\zeta)(\zeta+1)=\frac{1}{4}(-\zeta^2+2\zeta+3)=1-\frac{\Delta^2}{4}\,;$$

$K=1$:

$$Y(\zeta)=\frac{1}{4\zeta}(\zeta+1)^2=\frac{1}{4}(\zeta+2+\zeta^{-1})=1+\frac{1}{4}\delta^2\,;$$

$K=2$:

$$Y(\zeta)=\frac{1}{4\zeta^2}(3\zeta+1)(\zeta+1)=\frac{1}{4}(3+4\zeta^{-1}+\zeta^{-2})=1-\frac{\nabla^2}{4}\,.$$

○ $M=2$, $N=2$:—

$$\begin{cases}\tau(\zeta)=(\zeta-\zeta_2)^2=(\zeta+1)^2=4\left(1+\Delta+\dfrac{\Delta^2}{4}\right),\\ [\tau(\zeta)]^{-1}=\dfrac{1}{4}\left(1-\Delta+\dfrac{3}{4}\Delta^2-\dfrac{1}{2}\Delta^3+\dfrac{5}{16}\Delta^4+\cdots\right);\end{cases}$$

$K=0$:

$$Y(\zeta)=\frac{1}{16}(3\zeta^4-4\zeta^3-6\zeta^2+12\zeta+11)=1+\frac{1}{2}\Delta^3+\frac{3}{16}\Delta^4\,;$$

$K=1$:

$$Y(\zeta)=\frac{1}{16\zeta}(3-\zeta)(\zeta+1)^3=\frac{1}{16}(-\zeta^3+6\zeta+8+3\zeta^{-1})$$
$$=1-\frac{1}{\zeta}\left(\frac{1}{4}\Delta^3+\frac{1}{16}\Delta^4\right);$$

$K=2$:

$$Y(\zeta)=\frac{1}{16}(-\zeta^2+4\zeta+10+4\zeta^{-1}-\zeta^{-2})=1-\frac{1}{16}\delta^4;$$

$K=3$:

$$Y(\zeta)=\frac{1}{16}(3\zeta+8+6\zeta^{-1}-\zeta^{-3})=1+\frac{1}{16\zeta^3}(4\Delta^3+3\Delta^4);$$

$K=4$:

$$Y(\zeta)=\frac{1}{16}(11+12\zeta^{-1}-6\zeta^{-2}-4\zeta^{-3}+3\zeta^{-4})=1-\frac{1}{2}\nabla^3+\frac{3}{16}\nabla^4. \qquad (27)$$

○ $M=2$, $N=4$, $K=6$:

$$Y(\zeta)=\frac{1}{64}(57+30\zeta^{-1}-45\zeta^{-2}+20\zeta^{-3}+15\zeta^{-4}-18\zeta^{-5}+5\zeta^{-6}). \qquad (28)$$

$\Delta\equiv\zeta-1$, $\nabla\equiv1-\zeta^{-1}$, $\delta=\zeta^{\frac{1}{2}}-\zeta^{-\frac{1}{2}}$.

4. *Application of a Filter*

The practical way of using a filter will be as follows.

To begin with, we proceed the integration of the equations (1) with the initial values (2) by the method (3) with appropriate starting values until as many y_n^i's are obtained as are sufficient for the application of the filter. Then, applying the filter to the sequence $\{y_n^i\}$ to get successive $k-1$ sets of values of y_n^{i*}, we restart the integration by (3) with these newly obtained values of y_n^{i*} as the starting values. If the undesirable extraneous error components have grown up, we use the filter to remove them, restarting therefrom.

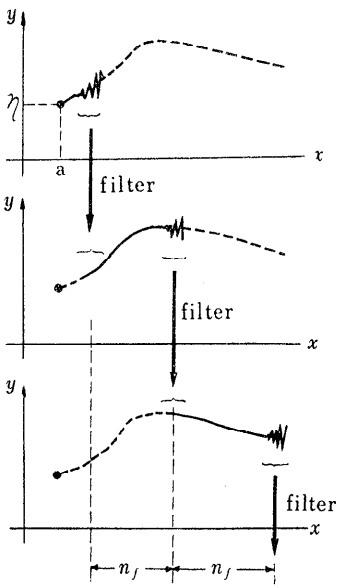

Fig. 1

The above process is illustrated figuratively in Fig. 1.

From our experience it is recommended to choose N equal to the order p (defined in (15)) of the method concerned, $M_\mu=0$ for $|\zeta_\mu|<1$ and $M_\mu=2$ for $|\zeta_\mu|\geqq1$. Furthermore, it is convenient to choose $K=N+\sum_{\mu=2}^{s}M_\mu$, because we then have $Y(\zeta)$ which is expressible in terms of the backward difference operators

only so that we can obtain y_n^{i*} using the values y_n^i, y_{n-1}^i, $\cdots$, y_{n-K}^i only. (27) and (28) are the filters determined in this manner for the midpoint rule and Milne's method, respectively.

It can also be shown that, if the step size h is small enough, the round-off error is small in comparison with h and the interval n_f of two successive filtering operations is chosen smaller than hg (where g is the maximum of the moduli of the eigenvalues of the matrix $[\lambda_\mu g_j^i(x)]$ in the relevant interval of x), then the growth of the extraneous error components in one interval is sufficiently suppressed by one operation of the filter.

5. *Examples*

Example 1:—Figs. 2 and 3 show the results of solving the initial value problem

$$\frac{dy}{dx}=f(x, y)=1-y^2, \qquad y(0)=0$$

by the midpoint rule

$$y_{n+1}=y_{n-1}+2hf_n$$

with $h=0.01$. The exact solution of the problem is evidently $y(x)=\tanh x$. y_A in Fig. 2 and e_A in Fig. 3 are the values of the numerical solution and the error which were obtained without using a filter, where the graphs of y_A and e_A oscillate rapidly up and down within the hatched region. y_B and e_B are the corresponding values obtained by using the filter of (27) every 150 steps, where the numerical instability seen in y_A and e_A is completely suppressed.

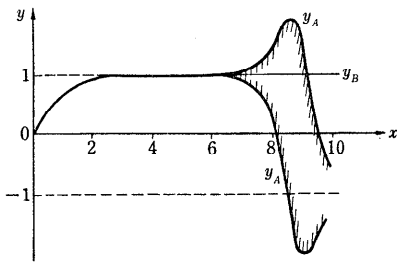

Fig. 2

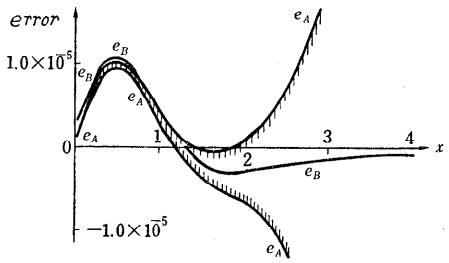

Fig. 3

Example 2:—Figs. 4 and 5 are the results of solving the problem

$$\frac{d^2y}{dx^2}+x\frac{dy}{dx}+y=0, \qquad y(0)=0, \qquad y'(0)=1$$

or

$$\left.\begin{aligned} \frac{dy^1}{dx}&=f^1(x, y^1, y^2)=y^2, \qquad y^1(0)=0, \\ \frac{dy^2}{dx}&=f^2(x, y^1, y^2)=-y^1-xy^2, \qquad y^2(0)=1, \\ y(x)&=y^1(x), \qquad y'(x)=y^2(x) \end{aligned}\right\}$$

by Milne's method

$$y_{n+1}{}^i=y_{n-1}{}^i+\frac{h}{3}(f_{n+1}{}^i+4f_n{}^i+f_{n-1}{}^i) \qquad (i=1, 2)$$

with $h=0.1$ (The predictor used is

$$y_{n+1}{}^i=y_{n-3}{}^i+\frac{4}{3}h(2f_n{}^i-f_{n-1}{}^i+2f_{n-2}{}^i).)$$

The exact solution is

$$y(x)=\exp\left(-\frac{1}{2}x^2\right)\int_0^x \exp\left(\frac{1}{2}t^2\right)dt.$$

y_A (and y'_A) and e_A are the values of the numerical solution and the error obtained without using a filter, while y_B (and y'_B) and e_B are those obtained by using the filter of (28) every 10 steps. The effect of the filter seems too obvious to add any further comment.

Besides the above two examples, we have made a number of experi-

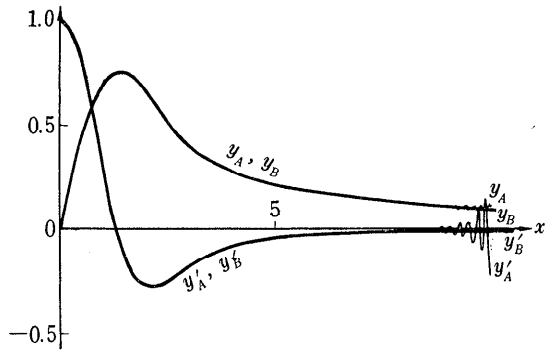

Fig. 4

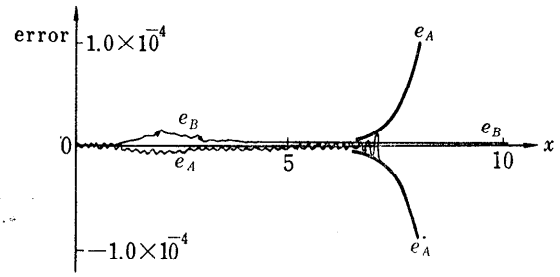

Fig. 5

mental computations for different problems by different methods. The filters designed according to the principle of § 3 and § 4 worked successfully in all those experiments.

Acknowledgement

The author withes to express his sincere gratitude to Professor Sigeiti Moriguti of the University of Tokyo for his constant guidance and many valuable suggestions regarding this research.

References

[1] MORIGUTI, S., AND M. TAKATA, *Suchi Keisan Ho, II* (in Japanese). "Modern Applied Mathematics" Series, B. 13. II., Iwanami, Tokyo, 1958.

[2] HAMMING, R. W., Stable predictor-corrector methods for ordinary differential equations. *Journal of the Association for Computing Machinery, 6* (1959), 37-47.

[3] MILNE, W. E., AND R. R. REYNOLDS, Stability of a numerical solution of differential equations. *Journal of the Association for Computing Machinery, 6* (1959), 196-203.

[4] HENRICI, P., *Discrete Variable Methods in Ordinary Differential Equations.* John Wiley & Sons, New York—London, 1962.