Stable Predictor-Corrector Methods for Ordinary Differential Equations*

R. W. Hamming

Bell Telephone Laboratories, Murray Hill, New Jersey

Introduction

Milne's method is the classic "predictor-corrector method" for solving ordinary differential equations. In spite of its known instability property, Milne's method has a number of virtues not possessed by its principal rival, the Runge-Kutta method, which are especially important when the order of the system of equations is fairly high ($N = 10$ to 30 or more). Hence it is worth examining predictor-corrector methods that do not have this instability property and at the same time are well adapted to machine computation. This paper gives a general technique for finding such stable methods, discusses one specific case which seems "on the average" to be a good compromise between conflicting interests, and sketches a second example.

Milne's Method

Since Milne's method is the standard predictor-corrector method, it is worth going carefully over a slightly modified form of it that is adapted to large scale digital computation. Let the equation to be solved be

$$y' \equiv \frac{dy}{dx} = f(x, y), \tag{1}$$

and suppose that the solution has been started (probably by the Runge-Kutta method). The next value is *predicted* by

$$p_{n+1} = y_{n-3} + \frac{4h}{3}(2y_n' - y'_{n-1} + 2y'_{n-2}), \tag{2a}$$

which has an error term $\frac{28}{90} h^5 y^{(5)}$. Milne's remark[1] that one can guess ahead can be adapted to machine computation as follows. Since p_{n+1} has an error of $\frac{28}{90} h^5 y^{(5)}$ and c_{n+1} (which will be defined in equation (2c)) has an error term $-\frac{1}{90} h^5 y^{(5)}$, then *if* $y^{(5)}$ were a constant $p_{n+1} - c_{n+1} = \frac{29}{90} h^5 y^{(5)}$, and subtracting $\frac{28}{29}(p_n - c_n)$ from p_{n+1} would exactly compensate for the error. In practice it is usually true that $p_n - c_n$ varies slowly from step to step so that using the *modified* value

$$m_{n+1} = p_{n+1} - \frac{28}{29}[p_n - c_n] \tag{2b}$$

will "mop up" most of the error in the predictor.

* Received July, 1958.

[1] *Numerical Solution of Differential Equations*, page 65.

This modified value is then used in the differential equation (1) to obtain an estimate m'_{n+1} of the derivative.

The *corrected* value is given by

$$c_{n+1} = y_{n-1} + \frac{h}{3}(m'_{n+1} + 4y_n' + y'_{n-1}), \tag{2c}$$

which has an error term, $-\frac{1}{90}h^5y^{(5)}$. This error can be partially compensated for by using the *final value*

$$y_{n+1} = c_{n+1} + \frac{1}{29}[p_{n+1} - c_{n+1}]. \tag{2d}$$

Thus, if the fifth derivative of y is a constant, then the method is exact; in general, the error depends on the sixth derivative.

Using this y_{n+1} in a second evaluation of the right-hand side of the differential equation (1) gives the final value of the derivative y'_{n+1}, and one step forward has then been completed.

Virtues of Milne's Method

The numerical value of $p_n - c_n$ is used as a control on the computation. If $p_n - c_n$ is very small, this indicates that the interval is too short and that computation time is being wasted. If $p_n - c_n$ suddenly becomes large, this is highly indicative of a machine error, while if it gradually grows large this indicates the need for shortening the interval of integration. Thus Milne's method supplies a running check that the method and interval size are suitable and that the computation is locally accurate enough to warrant going on.

The second asset of the method is that only two evaluations of the derivatives are made per step forward, while in the Runge-Kutta method four evaluations are needed. For high order systems of equations the evaluation time on the machine may be from 90 to 99 per cent of the computing cycle so that, since both methods use about the same interval size, this amounts to almost a factor of two in machine time saved (and at modern machine costs this can amount to a lot of money in a short time).

Instability of Milne's Method

Since instability is the main defect of Milne's method, it is desirable to examine its origin closely before showing how to eliminate it. It is easier to study the stability of the usual Milne's method, which merely predicts and then iterates the corrector (2c) until no more change is found in the corrected y_{n+1} value, than it is to study the more elaborate (and more efficient) method given above. The stability of the two methods is not essentially different.

Instead of using Milne's corrector, a lot of later algebra can be saved by introducing a "generalized corrector" formula at this point:

$$y_{n+1} = ay_n + by_{n-1} + cy_{n-2} + h(dy'_{n+1} + ey_n' + fy'_{n-1}). \tag{3}$$

This is the most general linear form which uses the data indicated. Other forms can be used, such as one using y'_{n-2}, which would lead to a more extensive theory, or a more accurate formula, depending on how the extra parameter was used.

Let z be the true solution of the differential equation (1), that is, let z satisfy

$$\frac{dz}{dx} = z' = f(x, z). \tag{4}$$

On the other hand the calculated solution y satisfies

$$y_n' = f(x_n, y_n) + E_1(n), \tag{5}$$

where $E_1(n)$ is the error at the nth value and is assumed to be small. Returning to z, we have the difference equation

$$z_{n+1} = az_n + bz_{n-1} + cz_{n-2} + h(dz'_{n+1} + ez_n' + fz'_{n-1}) + E_2(n), \tag{6}$$

where $E_2(n)$ is the corresponding error.

The error ϵ in the solution is defined by

$$\epsilon_n \equiv z_n - y_n, \tag{7}$$

and subtracting the two corrector equations (6) and (3) we get

$$\epsilon_{n+1} = a\epsilon_n + b\epsilon_{n-1} + c\epsilon_{n-2} + h(d\epsilon'_{n+1} + e\epsilon_n' + f\epsilon'_{n-1}) + E_2(n). \tag{8}$$

Next subtract the two differential equations (4) and (5) for z and y, and apply the mean value theorem,

$$\epsilon_n' = f(x_n, z_n) - f(x_n, y_n) - E_1(n) = \frac{(\partial f)}{(\partial y)} \epsilon_n - E_1(n). \tag{9}$$

For purposes of studying the growth of the error it is reasonable to assume that $E_1(n)$, $E_2(n)$, $\partial f/\partial y$ are all constants. The fact that they change slowly in practical cases indicates that this assumption is not severe. It is also convenient to rescale the problem into "natural units" by setting

$$x = (1/A)t, \quad \text{where} \quad A = \frac{\partial f}{\partial y} \tag{10}$$

(since $A \neq 0$; otherwise we would have no differential equation). Using this new variable the differential equation (9) for the error ϵ becomes

$$\frac{d\epsilon}{dt} = \epsilon - \frac{E_1}{A},$$

and putting this in the difference equation (8) we get[2]

$$\epsilon_{n+1} = a\epsilon_n + b\epsilon_{n-1} + c\epsilon_{n-2} + h(d\epsilon_{n+1} + e\epsilon_n + f\epsilon_{n-1}) + E_2 - \frac{h}{A}(d + e + f)E_1. \tag{11}$$

[2] It is convenient to use the same letter h, though in fact they are different in equations (6) and (11).

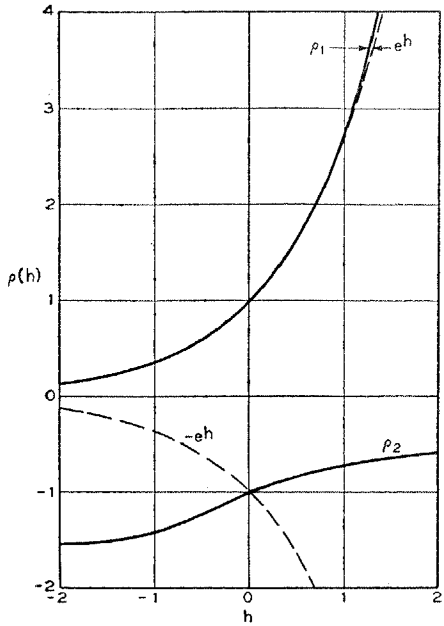

FIG. 1

This linear difference equation with constant coefficients may be solved by setting $\epsilon_n = \rho^n$, which gives the characteristic equation

$$\rho^3 = a\rho^2 + b\rho + c + h(d\rho^3 + e\rho^2 + f\rho)$$

or

$$h = \frac{\rho^3 - a\rho^2 - b\rho - c}{d\rho^3 + e\rho^2 + f\rho}. \tag{12}$$

Returning now to Milne's method where we have $a = c = 0, b = 1, d = f = \frac{1}{3}$, $e = \frac{4}{3}$, (12) becomes (see figure 1 for curve)

$$h = \frac{3(\rho^2 - 1)}{\rho^2 + 4\rho + 1}$$

which, for a given h, has two solutions, ρ_1 and ρ_2.

For any given h the general solution of (11) is, for Milne's method,

$$\epsilon_n = c_1(\rho_1)^n + c_2(\rho_2)^n + c_3. \tag{13}$$

In order to understand the implications of equation (13) it is necessary to examine the situation rather carefully. The first thing to note is that if $\partial f/\partial y$ is negative, then integration in the forward direction in x implies, because of (10), integration in the negative direction in t; i.e., we are then concerned with negative h values. In these cases $\rho_2(h) < -1$, and the term $c_2(\rho_2)^n$ ($c_2 \neq 0$) in (13) will oscillate and grow in size as n increases. Even if we start the numerical integration with $c_2 = 0$, "roundoff errors" are bound to introduce a c_2 which, while small, is not exactly zero, and ultimately this term will dominate the other two. Thus we see that Milne's method will not handle so simple an equation as

$$y' = -y, \qquad y(0) = 1 \tag{14}$$

because the method will inevitably introduce an error that grows in a geometric progression with a ratio more negative than -1, no matter how small the interval size h is chosen. As a result the true solution

$$z = e^{-x}$$

will gradually be lost in the numerical computations since, using (7) and (13),

$$y_n \equiv z_n - \epsilon_n = e^{-nh} - c_1(\rho_1)^n - c_2(\rho_2)^n - c_3 .$$

This is what is meant when Milne's method is said to be "unstable"; the method is unstable whenever $\partial f/\partial y < 0$.

A second point to note is that in practice it is often not the size of the error that matters, but rather the size in relation to the solution. Thus a method of integration which has a bounded error may not be satisfactory when, as in equation (14), the solution itself is decreasing to zero. On the other hand there is no point in trying hard to hold down the error in the equation

$$y' = y, \qquad y(0) = 1,$$

since the effect of a single error at one stage automatically grows exponentially with the solution. For such situations ρe^{-h} is a much better guide as to the seriousness of the error ϵ_n given in (13), and one may speak of "relative stability" as well as "stability". Curves of the relative error term are given in figure 2, while figure 3 gives an enlargement of the critical region near $\rho = 1$.

Finally, it should be noted that in integrating a set of equations whose solutions oscillate (for example $y_1' = y_2$, $y_2' = -y_1$, for which $y_1 = A \cos x + B \sin x$) the use of the relative error near a crossing of the x-axis is bound to be misleading. In such situations it is probably the absolute error rather than the relative error that matters.

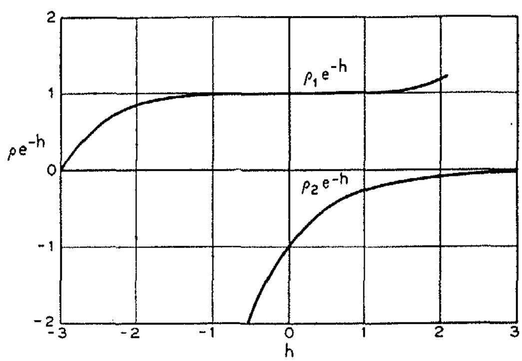

FIG. 2

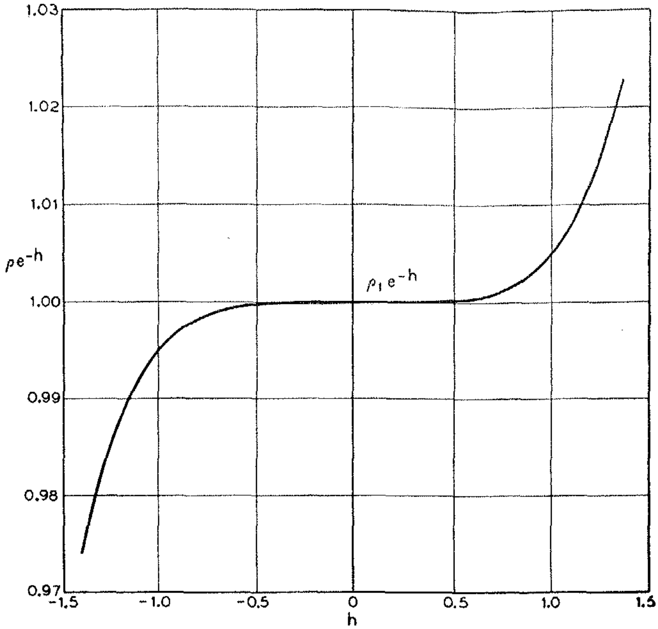

FIG. 3

In summary, then, the interpretation of the error ϵ_n as given in (13) requires careful examination of what is desired before asserting that the error is, or is not, serious.

Polynomial Approximation

We now return to the "generalized corrector" (3). Milne's method is exact in case the solution is a polynomial of degree 4 or less. Using this same criterion for the generalized corrector we obtain conditions on the coefficients $a, b, \cdots, f$. These conditions may be found either by expanding each of the y's in a Taylor series in h about x_n and equating the terms on both sides for $1, h, h^2, h^3, h^4$, or, equivalently, by requiring the equation to be exact for $y = 1, x, x^2, x^3, x^4$. In either case we get

$$\begin{aligned} a &= \frac{27(1-b)}{24} & d &= \frac{9-b}{24} \\ b &= b & e &= \frac{18+14b}{24} \\ c &= \frac{-3(1-b)}{24} & f &= \frac{-9+17b}{24}. \end{aligned} \tag{15}$$

The error term may be found either by considering the h^5 term in the Taylor series, or by using $x^5/5!$ in the formula. In either case the error term is found to be

$$kh^5y^{(5)} = \frac{-9 + 5b}{360} h^5y^{(5)}. \tag{16}$$

Roots When $h = 0$

The characteristic roots of the difference equation for $h = 0$ are given by

$$\rho^3 - a\rho^2 - b\rho - c = 0,$$

or, using (15), by

$$8\rho^3 - 9(1 - b)\rho^2 - 8b\rho + (1 - b) = 0,$$

which is shown in figure 4. The curves show that for absolute stability, $-6/10 < b < 1$. In this range there are a number of candidates for consideration, which have various special properties; they are given in the following table 1. (Note that the case $b = 1$ is Milne's method.)

A number of effects can be seen to happen as b goes from 1 to $-6/10$:

1) The error term k grows linearly from $-1/90$ to $-1/30$.

2) The sum of $|a| + |b| + |c|$, which roughly indicates the amount of roundoff trouble per step, goes from 1 at $b = 1$, to 1.25 at $b = 0$, to 2.6 at $b = -6/10$.

3) The cases $b = 1, 9/17, 0$, have one or more zero coefficients.

4) Every case has at least two coefficients of equal magnitude (which will save one multiplication).

5) The vertical asymptote just off the right-hand side of the paper in figures 1 and 2 moves from 3 to 2.5, making the integration error for positive $\partial f/\partial y$ relatively worse.

After examining and balancing these various effects the value $b = 0$ emerges

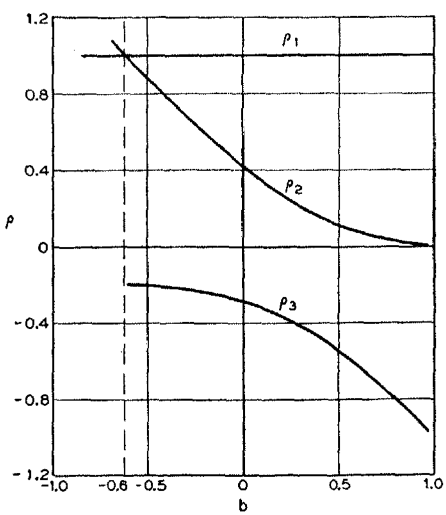

FIG. 4

TABLE 1

	$b = 1$	9/17	1/9	0	−1/7	−9/31	−6/10
a	0	9/17	1	9/8	9/7	45/31	9/5
b	1	9/17	1/9	0	−1/7	−9/31	−3/5
c	0	−1/17	−1/9	−1/8	−1/7	−5/31	−1/5
d	1/3	6/17	10/27	3/8	8/21	12/31	2/5
e	4/3	18/17	22/27	6/8	14/21	18/31	2/5
f	1/3	0	−8/27	−3/8	−10/21	−18/31	−4/5
k	−1/90	−3/170	−19/810	−1/40	−17/630	−9/310	−1/30
5!k	−4/3	−36/17	−76/27	−3	−68/21	−108/31	−4

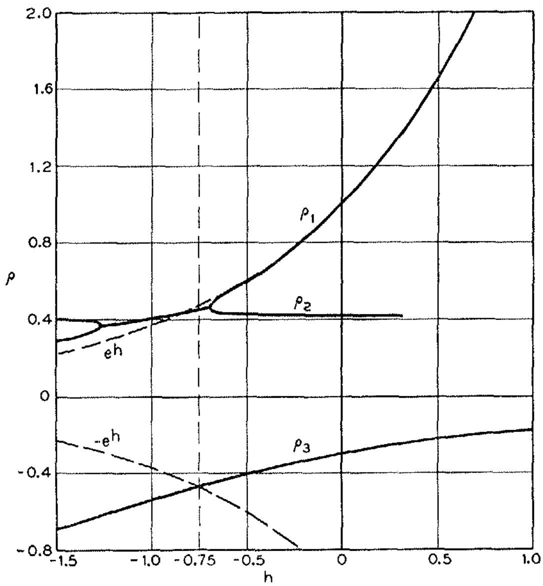

FIG. 5

as a good compromise for general use. A shortening of the interval of integration by 15 per cent will make the error term (16) about equal to that of Milne's method. For any specific situation, of course, a better choice of b can probably be made.

Figures 5–7 give plots of $\rho(h)$ in the case $b = 0$ corresponding to figures 1–3. These show that for $h > -.75$ the method is "relatively stable", as well as "stable".

The Case $b = 0$ In Practice

While the above discussion has been based on the repeated use of the corrector until the result converges, this is not the best way to proceed in practice.

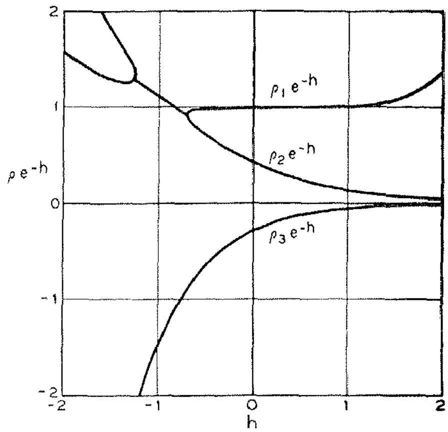

FIG. 6

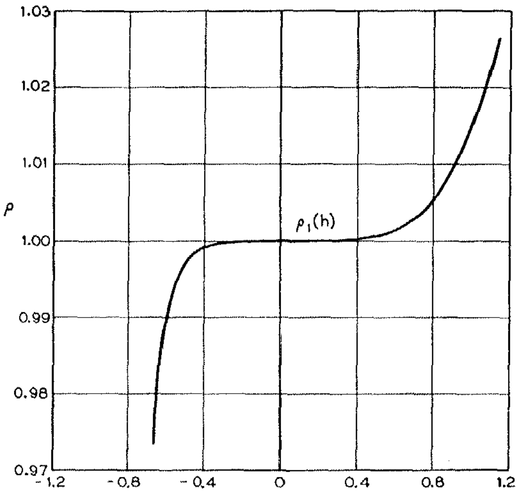

FIG. 7

It is usually better to shorten the interval than to repeatedly evaluate the derivatives. Thus we adopt the earlier pattern of operation for machine use. For the case $b = 0$ we have, corresponding to equations (2a), (2b), (2c), (2d):

$$\left.\begin{array}{ll}\text{predict} & p_{n+1} = y_{n-3} + \dfrac{4h}{3}(2y_n' - y_{n-1}' + 2y_{n-2}') \\[2ex] \text{modify} & m_{n+1} = p_{n+1} - \dfrac{112}{121}(p_n - c_n) \\[2ex] \text{correct} & c_{n+1} = \dfrac{1}{8}[9y_n - y_{n-2} + 3h(m_{n+1}' + 2y_n' - y_{n-1}')] \\[2ex] \text{final value} & y_{n+1} = c_{n+1} + \dfrac{9}{121}(p_{n+1} - c_{n+1})\end{array}\right\} \quad (17)$$

The Runge-Kutta method may be used for starting (the first three steps). While formulas can be given for computing the $p_n - c_n$ that is needed for the modification on the first step, it is probably better to set

$$p_3 - c_3 = 0. \tag{18}$$

This modified method is not as stable as the usual method, which repeatedly uses the corrector until no change occurs, and the bound on the stability moves from $h = .75$ to about $h = .65$. On the other hand the modified method is exact for solutions which are of degree 5 or lower, has an error term depending on $h^6 y^{(6)}$, and generally makes more efficient use of the computations done.

Miscellaneous Remarks

In the above equations the predictor is the same as in Milne's method. The main disadvantage of this formula is that it requires y_{n-3} as a starting value. It is natural to propose that the same methods as above be applied to the "generalized predictor",

$$y_{n+1} = ay_n + by_{n-1} + cy_{n-2} + h(dy_n' + ey'_{n-1} + fy'_{n-2}).$$

In the polynomial case, where we require exact prediction for $1, x, x^2, x^3, x^4$, the characteristic equation corresponding to (12) shows that for $h = 0$ such a formula would be highly unstable. Under the above method of not iterating the corrector some of this instability would filter through the equations and could cause trouble when $\partial f/\partial y \gg 0$.

Lower Accuracy Formulas

The same general methods can be applied to the problem of finding lower accuracy formulas. For a corrector of the form

$$y_{n+1} = ay_n + by_{n-1} + h(cy'_{n+1} + dy_n' + ey'_{n-1})$$

and requiring exact fit for $1, x, x^2, x^3$, we get

$$a = -4 + 12c \qquad d = 4 - 8c$$

$$b = 5 - 12c \qquad e = 2 - 5c$$

$$c = c \qquad \text{error term} = \frac{1 - 3c}{6} h^4 y^{(4)}.$$

The characteristic roots for $h = 0$ are

$$\rho_1 = 1$$

$$\rho_2 = -5 + 12c$$

so that stability requires

$$\tfrac{1}{2} > c > \tfrac{1}{3}.$$

The choice $c = 5/12$ is natural and leads to

$$y_{n+1} = y_n + \frac{h}{12}(5y'_{n+1} + 8y_n' - y'_{n-1}) - \frac{h^4}{4!}y^{(4)}, \tag{19}$$

which was given by Southard and Yowell.[3]

The corresponding predictor based on

$$y_{n+1} = ay_n + by_{n-1} + cy_{n-2} + h(dy_n' + ey'_{n-1})$$

leads to

$$\begin{aligned} a &= -4 - 5c & d &= 4 + 2c \\ b &= 5 + 4c & e &= 2 + 4c \\ c &= c & \text{error term} &= \frac{1-c}{6}h^4y^{(4)}. \end{aligned}$$

Southard and Yowell chose $c = 0$ which produces a nice short formula, but the product of the characteristic roots is 5, and this means instability in the prediction.

The only stable choice is $c = -1$ which leads to the very simple formula

$$y_{n+1} = y_n + y_{n-1} - y_{n-2} + 2h(y_n' - y'_{n-1}) + \frac{h^4}{3}y^{(4)}, \tag{20}$$

whose characteristic roots are 1, 1, −1.

If these two formulas, (19) and (20), are used in a method like the above, then the modifier is given by

$$m_{n+1} = p_{n+1} - \tfrac{8}{9}(p_n - c_n),$$

and the final value by

$$y_{n+1} = c_{n+1} + \tfrac{1}{9}(p_{n+1} - c_{n+1}),$$

while the error term is proportional to $h^5y^{(5)}$.

Conclusions

The main cost of gaining stability in a predictor-corrector method is the loss of some accuracy. In the principal case treated, $b = 0$, this loss in accuracy can be compensated for by shortening the interval of integration about 15 per cent.

The technique for finding stable methods has been illustrated by working examples comparable to Milne's method. Other generalized formulas can be used, and conditions other than being exact for polynomials can be imposed.

Acknowledgments

The author would like to thank Prof. J. W. Tukey, Miss R. A. Weiss, and D. C. Bomberger for advice, encouragement, and patience during the time this technique was first developed.

[3] *Math. Tables Aids Comp. 6* (1952), 253–4.