Closed rational integration formulas*

By Henry C. Thacher, Jr.†

Closed integration formulas are investigated which are exact when the integral is a rational function. The two-point formula is analogous to the trapezoidal rule, with the geometric replacing the arithmetic mean of the derivatives.

1. Introduction

In integrating the ordinary differential equation

$$y'(x) = f(x, y) \tag{1.1}$$

by predictor-corrector methods, the typical corrector formula is of the form

$$y(x_k) = \sum_{i=1}^{n} \alpha_i y(x_{k-i}) + \sum_{i=0}^{n} \beta_i y'(x_{k-i}) \tag{1.2}$$

and is thus a closed-type quadrature formula. The coefficients α_i and β_i are customarily determined partly by requiring that (1.2) be exact for polynomials of moderately high degree, and partly by stability considerations.

It is well known that many functions, particularly near a singularity in the complex plane, are far better represented by ratios of polynomials than they are by polynomials with the same number of parameters. Rational osculating interpolation algorithms have been investigated by Salzer (1962) and by Thacher (1961). Their algorithms, however, require that the value of the function be given at each point where a derivative is specified. They thus lead only to open formulas, valuable only as predictors.

It is the purpose of this paper to investigate a class of closed integration formulas, which are exact when the integral is a rational function. In Section 2, the problem is stated in its most general form, with arbitrary basis functions and spacing of base points. The only possible solution is shown to be one of two roots of a quadratic equation, the coefficients of which are rather complicated determinants. In Section 3, the problem is specialized to equally-spaced base points, and polynomial basis functions. In Section 4, the two-point case is investigated in detail, and is shown to lead to a formula analogous to the trapezoidal rule with the arithmetic mean of the derivatives replaced by the geometric mean. The formula is solved explicitly for the differential equations satisfied by several common functions, and is found to give considerably more accurate results than the three-term truncated Taylor series or the trapezoidal rule. Finally, in Section 5, the three-point formula is derived, and the conditions under which it fails are determined.

2. The general problem

Let $n + 1$ distinct base points, $x_i (i = 0, 1, \ldots, n)$, n function values, y_i, $(i = 1, \ldots, n)$, and $n + 1$ derivative values, $f_i (i = 0, 1, \ldots, n)$ be given. Let $R(x)$ be a specified generalized rational function with $m = \mu + \nu + 2$ parameters $\{p_k, q_k\}$:

$$R(x) = \frac{P(x)}{Q(x)} = \frac{\sum_{k=0}^{\mu} p_k \phi_k(x)}{\sum_{k=0}^{\nu} q_k \psi_k(x)}. \tag{2.1}$$

We wish to find an expression for $y_0 = R(x_0)$ which is valid whenever the y_i and f_i are, respectively, the values of a rational function of the form (2.1) at x_i, and of its derivative, i.e. when, for some set of $\{p_k, q_k\}$,

$$y_i = R(x_i) \tag{2.2}$$

$$f_i = R'(x_i). \tag{2.3}$$

Provided that $Q(x_i) \neq 0$, (2.2) leads to the condition

$$Q(x_i)\, y_i = P(x_i) \tag{2.4}$$

or

$$\Sigma y_i \psi_k(x_i) q_k - \Sigma \phi_k(x_i) p_k = 0. \tag{2.5}$$

Similarly, provided in addition that $P(x_i) \neq 0$,

$$P(x_i) R'(x_i) = R(x_i) P'(x_i) - R^2(x_i) Q'(x_i) \tag{2.6}$$

or, using (2.2) and (2.3),

$$\Sigma y_i^2 \psi_k'(x_i) q_k + \Sigma f_i \phi_k(x_i) p_k - \Sigma y_i \phi_k'(x_i) p_k = 0. \tag{2.7}$$

For each i such that y_i (and so $P(x_i)$ as well) does not vanish, adding f_i times (2.5) to (2.7) and dividing by y_i leads to:

$$\Sigma[y_i \psi_k'(x_i) + f_i \psi_k(x_i)] q_k - \Sigma \phi_k'(x_i) p_k = 0. \tag{2.8}$$

For the i for which $y_i = 0$, $P(x_i) = 0$, and

$$f_i = R'(x_i) = P'(x_i)/Q(x_i) \tag{2.9}$$

which is equivalent to (2.8).

Equations (2.5) and (2.8) form a system of $2n + 2$ homogeneous equations for the m quantities p_k and q_k. Since we are restricting ourselves to data consistent with

* Work supported by the U.S. Atomic Energy Commission.

† *Argonne National Laboratory*, 9700 *South Cass Avenue*, *Argonne*, *Illinois*, *U.S.A.*

a rational function of the form (2.1), a nontrivial set of p_k and q_k satisfying these equations exists. A necessary condition for this is that the rank of the matrix of the coefficients, Δ, be less than m, i.e. that all the determinants which may be formed by selecting m rows of Δ vanish. It is convenient to represent Δ as a compound matrix.

$$\Delta = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where

$$\left.\begin{aligned} A_{i+1,j+1} &= y_i\psi_j(x_i) \\ B_{i+1,j+1} &= -\phi_j(x_i) \\ C_{i+1,j+1} &= y_i\psi_j'(x_i) + f_i\psi_j(x_i) \\ D_{i+1,j+1} &= -\phi_j'(x_i). \end{aligned}\right\} \quad (2.11)$$

Each y_i appears in one equation of (2.5) and one of (2.8), and each f_i only in one equation of (2.8). The m-rowed determinants of Δ are therefore polynomials of total degree m in the y_i, and at most quadratic with respect to each of the y_i individually. In particular, the vanishing of each of the determinants which contains either the first row of A and B or the first row of C and D, or both, provides a set of quadratic equations for y_0 in terms of the f_i and the other y_i. Since we are assuming that the y_i and f_i are, in fact, derived from a rational function of the specified type, these equations will have a common solution, which will also satisfy any arbitrary linear combination of them. When y_0 satisfies more than one equation (i.e. when $m < 2n + 2$), a family of integration formulas exists corresponding to solutions of various linear combinations of the equations. By selecting the proper linear combination, the balance between stability and truncation error may be controlled. Extensions along these lines are, however, beyond the scope of the present study, in which we shall limit ourselves to the case where $m = 2n + 2$.

In this case, y_0 is subject to the single condition,

$$|\Delta| = 0 \qquad (2.12)$$

which is, as we have shown, an algebraic equation of degree 2 at most. The coefficients may be obtained in the usual fashion by expanding Δ in minors of the rows containing y_0.

Difficulties will arise if there are more than two distinct rational functions for which the data satisfy (2.2) and (2.3). In this case, there are more than two sets of independent solutions to (2.5) and (2.8), and the rank of Δ is less than $m - 2$. All the coefficients of the equation obtained by expanding Δ in minors of the first rows of A and B and of C and D will then vanish, and any value of y_0 will satisfy (2.12). A solution may still be obtained by deleting a sufficient number of rows and columns of Δ so that the rank is one less than the dimension. This amounts to assigning the value zero to those of the p_k and q_k which are not specified by the data.

3. Specialization

Further progress will be simplified by reducing the generality of our treatment. A particularly important case, especially for numerical solution of differential equations, is that with equally-spaced base points. By the simple linear transformation

$$x = (\bar{x}_0 - \bar{x})/h \qquad f = -h(dy/d\bar{x}) \qquad (3.1)$$

we may change a problem with arbitrary origin, $\bar{x}_0$, and step size, h, to one with origin zero, and base points at the negative integers,

$$x_i = -i. \qquad (3.2)$$

It is also convenient at this time to specify the form of the interpolating function (2.1) more precisely. We will restrict our treatment to the case where the numerator and denominator are both linear combinations of the same set of basis functions. Although other sets may be important in special problems, and can be treated by the same method, polynomials are undoubtedly the most generally useful set, and we will limit our explicit development to them, setting

$$\phi_k(x) = \psi_k(x) = x^k \qquad (k = 0, 1, 2, \ldots n) \quad (3.3)$$

$$\phi_k'(x) = \psi_k'(x) = \begin{cases} 0 & (k = 0) \\ kx^{k-1} & (k = 1, 2, \ldots, n). \end{cases} \quad (3.4)$$

With this selection of basis functions and base points, A, B, C, and D are all square $(n + 1) \times (n + 1)$ matrices with elements

$$\left.\begin{aligned} A_{i+1,j+1} &= (-i)^j y_i \\ B_{i+1,j+1} &= -(-i)^j \end{aligned}\right\} \quad (3.5)$$

$$(i, j = 0, \ldots, n)$$

$$\left.\begin{aligned} C_{i+1,j+1} &= j(-i)^{j-1}y_i + (-i)^j f_i \\ D_{i+1,j+1} &= -j(-i)^{j-1} \end{aligned}\right\} \quad (3.6)$$

$$(i, j = 0, \ldots, n)$$

where $j(0)^{j-1}$ and $(0)^{j-1}$ are taken to be 1 when $j = 1$, and 0 otherwise.

The first rows of these matrices are particularly sparse: $(y_0, 0, \ldots, 0)$, $(-1, 0, \ldots, 0)$, $(f_0, y_0, 0, \ldots, 0)$ and $(0, -1, 0, \ldots, 0)$ for A, B, C, and D, respectively. This characteristic suggests evaluating $|\Delta|$ by expanding in minors of the first and $(n+2)$th rows. Let $M^{(i, \ldots, j; k, \ldots, m)}$ denote the matrix obtained from M by deleting the $i, \ldots, j$th rows, and the $k, \ldots, m$th columns. Then, expanding in minors of the first row,

$$|\Delta| = y_0 \begin{vmatrix} A^{(1;1)}, & B^{(1;)} \\ C^{(;1)}, & D \end{vmatrix} + (-1)^n \begin{vmatrix} A^{(1;)}, & B^{(1;1)} \\ C \quad, & D^{(;1)} \end{vmatrix}. \quad (3.7)$$

Expanding these determinants in minors of the $(n + 1)$th row (the first rows of C and D),

$$0 = |\Delta| = (-1)^n y_0 \left\{ y_0 \begin{vmatrix} A^{(1;1,2)}, & B^{(1;)} \\ C^{(1;1,2)}, & D^{(1;)} \end{vmatrix} + (-1)^n \begin{vmatrix} A^{(1;1)}, & B^{(1;2)} \\ C^{(1;1)}, & D^{(1;2)} \end{vmatrix} \right\}$$
$$+ (-1)^n \left\{ (-1)^n f_0 \begin{vmatrix} A^{(1;1)}, & B^{(1;1)} \\ C^{(1;1)}, & D^{(1;1)} \end{vmatrix} - (-1)^n y_0 \begin{vmatrix} A^{(1;2)}, & B^{(1;1)} \\ C^{(1;2)}, & D^{(1;1)} \end{vmatrix} + \begin{vmatrix} A^{(1;)}, & B^{(1;1,2)} \\ C^{(1;)}, & D^{(1;1,2)} \end{vmatrix} \right\} \quad (3.8)$$

$$= (-1)^n \begin{vmatrix} A^{(1;1,2)}, & B^{(1;)} \\ C^{(1;1,2)}, & D^{(1;)} \end{vmatrix} y_0^2 + \left\{ \begin{vmatrix} A^{(1;1)}, & B^{(1;2)} \\ C^{(1;1)}, & D^{(1;2)} \end{vmatrix} - \begin{vmatrix} A^{(1;2)}, & B^{(1;1)} \\ C^{(1;2)}, & D^{(1;1)} \end{vmatrix} \right\} y_0$$
$$+ \begin{vmatrix} A^{(1;1,2)}, & B^{(1;1)} \\ C^{(1;1,2)}, & D^{(1;1)} \end{vmatrix} f_0 + (-1)^n \begin{vmatrix} A^{(1;)}; & B^{(1;1,2)} \\ C^{(1;)}, & D^{(1;1,2)} \end{vmatrix}. \quad (3.9)$$

Thus, $|\Delta| = 0$ may be written as:

$$0 = (-1)^n \alpha y_0^2 + (\beta - \gamma) y_0 + \delta f_0 + (-1)^n \epsilon \quad (3.10)$$

where α, β, γ, δ, and ϵ are independent of y_0 and f_0 and are given by

$$\alpha = \begin{vmatrix} A^{(1;1,2)}, & B^{(1;)} \\ C^{(1;1,2)}, & D^{(1;)} \end{vmatrix} \quad \beta = \begin{vmatrix} A^{(1;1)}, & B^{(1;2)} \\ C^{(1;1)}, & D^{(1;2)} \end{vmatrix}$$
$$\gamma = \begin{vmatrix} A^{(1;2)}, & B^{(1;1)} \\ C^{(1;2)}, & D^{(1;1)} \end{vmatrix} \quad \delta = \begin{vmatrix} A^{(1;1)}, & B^{(1;1)} \\ C^{(1;1)}, & D^{(1;1)} \end{vmatrix}$$
$$\epsilon = \begin{vmatrix} A^{(1;)}, & B^{(1;1,2)} \\ C^{(1;)}, & D^{(1;1,2)} \end{vmatrix} \quad (3.11)$$

4. The two-point formula

For $n = 1$,

$$A = \begin{pmatrix} y_0, 0 & \\ y_1, & -y_1 \end{pmatrix} \qquad B = \begin{pmatrix} -1, & 0 \\ -1, & 1 \end{pmatrix}$$
$$C = \begin{pmatrix} f_0, y_0 & \\ f_1, & y_1 - f_1 \end{pmatrix} \qquad D = \begin{pmatrix} 0, & -1 \\ 0, & -1 \end{pmatrix} \quad (4.1)$$

and the determinants (3.11) become

$$\alpha = \begin{vmatrix} -1, & 1 \\ 0, & -1 \end{vmatrix} = 1 \qquad \beta = \begin{vmatrix} -y_1, & -1 \\ y_1 - f_1, & 0 \end{vmatrix} = y_1 - f_1$$
$$\gamma = \begin{vmatrix} y_1, & 1 \\ f_1, & -1 \end{vmatrix} = -y_1 - f_1 \quad \delta = \begin{vmatrix} -y_1, & 1 \\ y_1 - f_1, & -1 \end{vmatrix} = f_1$$
$$\epsilon = \begin{vmatrix} y_1, & -y_1 \\ f_1, & y_1 - f_1 \end{vmatrix} = y_1^2. \quad (4.2)$$

Hence, we have for (3.10)

$$0 = -y_0^2 + 2y_1 y_0 + f_1 f_0 - y_1^2 \quad (4.3)$$

or

$$y_0 = y_1 \pm \sqrt{(f_1 f_0)}. \quad (4.4)$$

If (4.4) is to be correct for linear functions, the sign of the square root must be the same as the derivative f_1. Hence, returning at the same time to an arbitrary interval, $h = x_0 - x_1$, we arrive at the formula

$$y_0 = y_1 + h \operatorname{sgn}(f_1)\sqrt{(f_1 f_0)}. \quad (4.5)$$

It is of interest that this formula is formally equivalent to the trapezoidal rule, with the arithmetic mean of the derivatives replaced by their geometric mean.

Since α is independent of the data, the difficulties referred to at the end of Section 2, where more than one independent rational is consistent with the data, cannot occur. On the other hand, it is impossible for the derivative of the ratio of two linear functions to change sign, so that this formula is not suitable for functions which are not strictly monotone in the interval $[x_0, x_1]$.

If $f(x, y)$ is a polynomial of degree 2 or less in y (i.e. if the differential equation is a generalized Riccati equation) (4.3) becomes a quadratic equation for y_0 in terms of y_1, f_1, x_1, and x_0. We may thus obtain an explicit expression for y_0. Specifically, writing the differential equation as

$$y' = f = -(\alpha y^2 + \beta y + \gamma) \quad (4.6)$$

with α, β, and γ functions of x, we find

$$(y_0 - y_1)^2 = (x_0 - x_1)^2(\alpha_0 y_0^2 + \beta_0 y_0 + \gamma_0)$$
$$(\alpha_1 y_1^2 + \beta_1 y_1 + \gamma_1)$$
$$\equiv -h^2(\alpha_0 y_0^2 + \beta_0 y_0 + \gamma_0) f_1 \quad (4.7)$$

or

$$(1 + h^2 f_1 \alpha_0) y_0^2 - (2y_1 - h^2 f_1 \beta_0) y_0$$
$$+ (y_1^2 + h^2 f_1 \gamma_0) = 0 \quad (4.8)$$

so that

$$y_0 = \frac{y_1 \pm h\sqrt{[-f_1(\gamma_0 + \beta_0 y_1 + \alpha_0 y_1^2) - h^2 f_1^2(\alpha_0 \gamma_0 - \beta_0^2/4)]} - h^2 f_1 \beta_0/2}{1 + h^2 f_1 \alpha_0}. \quad (4.9)$$

Since many common functions obey a generalized Riccati equation, (4.9) is of interest as a source of approximations.* Some typical examples include:

(a) *The exponential function.* Setting $\alpha = \gamma = 0$, $\beta = -1$ in (4.6), we obtain the equation for the exponential function,

$$y' = y. \tag{4.10}$$

If we set $x_1 = 0$, $y_1 = 1$, $x_0 = x$ and $f_1 = 1$,

$$y_0 \equiv y \cong 1 \pm x\sqrt{(1 + \tfrac{1}{4}x^2)} + x^2/2. \tag{4.11}$$

Inspection reveals that the positive sign should be taken. Expanding the square root by the binomial theorem, and comparing with the power series for e^x, we find that the error is

$$e^x - y = \frac{1}{24}x^3 + \frac{1}{24}x^4 + \frac{31}{1920}x^5 + \dots \tag{4.12}$$

approximately half that of the trapezoidal rule approximation

$$y = \frac{1 + x/2}{1 - x/2} = 1 + x + \frac{x^2}{2} + \frac{x^3}{4} + \frac{x^4}{8} + \frac{x^5}{16} + \dots \tag{4.13}$$

and one-fourth the error of the truncated power series. It may be noted that (4.13) is the first diagonal Padé approximant.

(b) *The logarithm function.* Setting $\alpha = \beta = 0$, $\gamma = -1/(1 + x)$ leads to the equation

$$y' = 1/(1 + x) \tag{4.14}$$

with the solution $y = \ln(1 + x)$ if $y(0) = 0$. In this case (4.9) reduces to

$$y = x\sqrt{[1/(1 + x)]} \tag{4.15}$$

and the error becomes

$$\ln(1 + x) - y = -\frac{1}{24}x^3 + \frac{1}{16}x^4 - \frac{47}{640}x^5 + \dots \tag{4.16}$$

again appreciably better than the trapezoidal approximation

$$y = \frac{x}{2}\left(1 + \frac{1}{1 + x}\right) = x - \frac{x^2}{2} + \frac{x^3}{2} - \frac{x^4}{2} + \dots. \tag{4.17}$$

Substituting x for $1 + x$ in (4.15) gives the approximation

$$\ln x \approx x^{1/2} - x^{-1/2}. \tag{4.18}$$

(c) *The tangent.* An approximation may be derived for the tangent based on the differential equation

$$y' = 1 + y^2. \tag{4.19}$$

However, this approximation neglects the fact that the tangent is an odd function, and has an error proportional to x^3. A more efficient approximation may be obtained by considering the function

$$y(x) = \tan(\sqrt{x})/\sqrt{x}. \tag{4.20}$$

This function satisfies the differential equation

$$y'(x) = \frac{1}{2}y^2 - \frac{1}{2x}y + \frac{1}{2x} \quad y(0) = 1 \quad y'(0) = \frac{1}{3} \tag{4.21}$$

and thus may be approximated by

$$y(x) = \left[1 + \frac{5}{12}x\sqrt{\left(1 - \frac{4}{25}x\right)} - \frac{x}{12}\right] \Big/ \left(1 - \frac{x^2}{6}\right). \tag{4.22}$$

We thus obtain the approximation

$$\tan x \cong x\left[1 + \frac{5}{12}x^2\sqrt{(1 - 0{\cdot}16x^2)} - \frac{x^2}{12}\right] \Big/ \left(1 - \frac{x^4}{6}\right) \tag{4.23}$$

with an error $(2/7875)x^7 + O(x^9)$.

(d) *The error function.* Although the error function does not satisfy a differential equation with algebraic coefficients, the product $\exp(x^2)\,\mathrm{erf}(x)$ does. Since the latter function is an odd function of x, we choose to approximate the function

$$y(x) = e^x \int_0^{x^{1/2}} e^{-t^2}\,dt/x^{1/2}. \tag{4.24}$$

This function obeys the differential equation

$$y'(x) = \frac{1}{x\sqrt{\pi}} + \left(1 - \frac{1}{2x}\right)y \tag{4.25}$$

with boundary conditions

$$y(0) = \frac{2}{\sqrt{\pi}}\,y'(0) = \frac{4}{3\sqrt{\pi}}. \tag{4.26}$$

Using (4.7)

$$\begin{aligned} y^2 - \frac{4}{\sqrt{\pi}}y + \frac{4}{\pi} &= \frac{4x^2}{3\sqrt{\pi}}\left(\frac{1}{x\sqrt{\pi}} + \frac{2x - 1}{2x}y\right) \\ &= \frac{4x}{3\pi} + \frac{2}{3\sqrt{\pi}}(2x^2 - x)y. \end{aligned} \tag{4.27}$$

Hence

$$y^2 - \frac{2}{3\sqrt{\pi}}(6 - x + 2x^2)y + \frac{4}{3\pi}(3 - x) = 0 \tag{4.28}$$

and

$$y = \frac{6 - x + 2x^2 + 5x\sqrt{\left[1 - \frac{4}{25}(x - x^2)\right]}}{3\sqrt{\pi}} \tag{4.29}$$

with an error of $(64/2625\sqrt{\pi})x^3$.

* Investigation of the use of (4.4) to obtain explicit approximations to functions obeying a generalized Riccati equation was stimulated by discussions at the Conference on Approximations held at Gatlinburg, Tenn. on 21–26 October 1963. This conference was sponsored by the Society for Industrial and Applied Mathematics and supported by the National Science Foundation and the U.S. Atomic Energy Commission.

The corresponding approximation for the error function itself is:

$$\frac{2}{\pi}\int_0^z e^{-t^2}\,dt = \frac{z\,e^{-z^2}}{3\sqrt{\pi}}$$

$$\left(6+\left\{5\sqrt{\left[1-\frac{4}{25}(z^2-z^4)\right]}-1\right\}z^2+2z^4\right). \tag{4.30}$$

5. The three-point formula

For $n = 2$, the algebra becomes rather formidable. We have, for this case,

$$A = \begin{pmatrix} y_0, & 0, & 0 \\ y_1, & -\,y_1, & y_1 \\ y_2, & -2y_2, & 4y_2 \end{pmatrix} \qquad B = \begin{pmatrix} -1, & 0, & 0 \\ -1, & 1, & -1 \\ -1, & 1, & -4 \end{pmatrix}$$

$$C = \begin{pmatrix} f_0, & y_0, & 0 \\ f_1, & y_1 - f_1, & -2y_1 + f_1 \\ f_2, & y_2 - 2f_2, & -4y_2 + 4f_2 \end{pmatrix} \quad D = \begin{pmatrix} 0, & -1, & 0 \\ 0, & -1, & 2 \\ 0, & -1, & 4 \end{pmatrix} \tag{5.1}$$

After considerable manipulation, we find for the coefficients of the quadratic (3.10),

$$\alpha = 4y_1 - 4y_2 - f_1 - 4f_2 \tag{5.2}$$

$$\beta - \gamma = -4y_1^2 + 4y_2^2 + 2y_2 f_1 + 8y_1 f_2 \tag{5.3}$$

$$\delta = -4y_1^2 + 8y_1 y_2 - 4y_2^2 + 4f_1 f_2 \tag{5.4}$$

$$\epsilon = 4y_1^2 y_2 - 4y_1 y_2^2 - y_2^2 f_1 - 4y_1^2 f_2. \tag{5.5}$$

The equation simplifies slightly if we consider the increments in the dependent variable,

$$\Delta_0 = y_0 - y_1 \qquad \Delta_1 = y_1 - y_2. \tag{5.6}$$

Then, (3.10) reduces to

$$(4\Delta_1 - f_1 - 4f_2)\Delta_0^2 + 2\Delta_1(2\Delta_1 - f_1)\Delta_0 - \Delta_1^2(4f_0 + f_1) + 4f_0 f_1 f_2 = 0 \tag{5.7}$$

which we may abbreviate as

$$a_2\Delta_0^2 + a_1\Delta_0 + a_0 = 0. \tag{5.8}$$

When $a_2 \neq 0$, we may solve (5.8) by the quadratic formula, obtaining, after simplification

$$\Delta_0 = \frac{\Delta_1 f_1 - 2\Delta_1^2 + 2\surd\{(\Delta_1^2 - f_1 f_2)[\Delta_1^2 + f_0(4\Delta_1 - f_1 - 4f_2)]\}}{4\Delta_1 - f_1 - 4f_2}. \tag{5.9}$$

We have chosen the larger root of the quadratic to insure that our formula holds for $y(x)$ a polynomial. Provided $a_2 \neq 0$, this formula gives acceptable results unless the discriminant is negative. As with the two-point formula, a negative discriminant occurs only for data which could not have been derived from the ratio of two quadratic polynomials.

Vanishing of a_2 is not at all serious unless a_1 also vanishes, since (5.8) reduces to a linear equation with the solution

$$\Delta_0 = \frac{a_0}{a_1} = \frac{\Delta_1^2(4f_0 + f_1) - 4f_0 f_1 f_2}{2\Delta_1(2\Delta_1 - f_1)}$$

$$= \frac{4f_0(4f_2 - f_1)^2 + f_1(4f_2 + f_1)^2}{4(4f_2 - f_1)(4f_2 + f_1)} \tag{5.10}$$

after eliminating Δ_1 by the condition $a_2 = 0$.

When both a_2 and a_1 vanish, but $a_0 \neq 0$, the data are again inconsistent with the form of rational function. If all three coefficients vanish, it is impossible to select Δ_0 so that the rational function is uniquely determined. We thus have the possibility of the type of indeterminacy mentioned at the end of Section 2. We will investigate this somewhat unusual case in more detail.

The vanishing of a_2 implies that

$$\Delta_1 = \frac{1}{4}(4f_2 + f_1) \tag{5.11}$$

while the vanishing of a_1, taken with (5.11) leads to:

$$\frac{1}{4}(4f_2 + f_1)(4f_2 - f_1) = 0 \tag{5.12}$$

or

$$f_1 = \pm\, 4f_2. \tag{5.13}$$

Finally, the vanishing of a_0 implies that

$$\frac{f_0}{4}(4f_2 - f_1)^2 - \frac{f_1}{16}(4f_2 + f_1)^2 = 0. \tag{5.14}$$

If $f_1 = 4f_2$, (5.14) reduces to $-f_1^3/4 = 0$, and so $f_1 = f_2 = \Delta_1 = 0$. With these data, (2.5) and (2.8) can be satisfied only if $p_j = y_1 q_j (j = 0, 1, 2)$ so that

$$R(x) = \frac{y_1 q_0 + y_1 q_1 x + y_1 q_2 x^2}{q_0 + q_1 x + q_2 x^2} = y_1. \tag{5.15}$$

The q_j may indeed be chosen arbitrarily and still satisfy the data at x_1 and x_2. The value at x_0 is, however, uniquely determined, and, indeed, the data are only consistent if $f_0 = 0$.

The case $f_1 = -4f_2 \neq 0$ is more interesting. Then the vanishing of a_2 requires that $\Delta_1 = 0$, while for consistency, a_0 must vanish so that (5.14) implies that $f_0 = 0$. Using these results to eliminate f_0, f_1, and y_1 from (2.5) and (2.8), we may solve the four equations which are independent of y_0 to give

$$p_0 = y_2 q_0 - 8f_2(q_0 - q_1 + q_2) \tag{5.16}$$

$$q_1 = y_2 q_1 - 12f_2(q_0 - q_1 + q_2) \tag{5.17}$$

$$p_2 = y_2 q_2 - 4f_2(q_0 - q_1 + q_2) \tag{5.18}$$

$$q_1 = 3q_0/2. \tag{5.19}$$

Eliminating q_1 with (5.19), and introducing the parameter $p = 2q_2/q_0$, we find that the rational function

$$R(x) = y_2 + 4f_2(1 - p)\frac{2 + 3x + x^2}{2 + 3x + px^2} \tag{5.20}$$

is consistent with the data for any p except $p = 1$. Thus, when $y_1 = y_2$, $f_1 = -4f_2 \neq 0$, the data are inconsistent when $f_0 \neq 0$, and when $f_0 = 0$ neither the rational approximating function nor the value of y_0 is determined by the assumption that $y(x)$ is the ratio of two quadratics. In many cases, however, the requirement that $f_0 = f(x_0, y_0) = 0$ will suffice to determine, or at least restrict, the possible values of y_0.

It should be emphasized that these difficulties, which have been discussed at some length because of their pertinence to the general rational integration problem, are of little importance for the three-point case. They can only occur when both a_2 and a_1 vanish, i.e. when $y_1 = y_2$ and $f_1 = \pm 4f_2$. These circumstances will rarely be encountered in practical computation, where the major obstacle to successful application of (5.9) or (5.10) will be the failure of the assumption that the data can be derived from the ratio of two quadratic polynomials.

Note added in proof

While this paper was in process of publication, the contribution of Lambert and Shaw (1965) appeared. This paper mentions the geometric-mean formula (4.4), but devotes most attention, to a class of open two- and three-point formulas requiring one or more higher derivatives. Numerical results of applying these formulas to the equation (4.19), with $y(0) = 1$ are given, and confirm the superiority of even open rational formulas over comparable polynomial formulas.

References

LAMBERT, J. D. and SHAW, B. (1965). "On the Numerical Solution of $y' = f(x, y)$ by a Class of Formulas Based on Rational Approximation," *Math. of Comp.*, Vol. 19, p. 456.

SALZER, H. E. (1962). "Note on Osculatory Rational Interpolation," *Math. of Comp.*, Vol. 16, p. 486.

THACHER, H. C., Jr. (1961). "A Recursive Algorithm for Rational Osculatory Interpolation," *SIAM Rev.* Vol. 3, p. 359 (Abstract).

Book Reviews

Error in Digital Computation, Volume 1, edited by L. B. Rall, 1965; 324 pages. (London and New York: *John Wiley & Sons Ltd.*, 51s.).

This volume results from an advanced seminar conducted by the *Mathematics Research Center* of the *United States Army* at the *University of Wisconsin* during October, 1964. The subjects considered belong to numerical analysis, hence they exclude mistakes of programming and the malfunctioning of computers. At the seminar there were five sessions, each on a separate subject, and as a result this book contains expository papers by five different authors. It is convenient to discuss them separately.

The Problem of Error in Digital Computation, by John Todd.

Starting with the assertion that it does not appear feasible to perform rigorous error analyses of the traditional kind for all algorithms, or to accompany all computations by automatic error analyses, Professor Todd argues that we need to change our standards of acceptability. He surveys interpolation, square roots, eigenvalues of symmetric matrices, Monte Carlo methods, good error estimates and controlled numerical experiments. He prefers "cheap" error estimates, i.e. those which can be had for a fraction of the work needed to produce the result itself. The controlled numerical experiments are to be carried out on new algorithms using specially prepared sets of test data. Only algorithms which perform well in the experiments or are otherwise interesting need be subjected to more complete error analyses.

Techniques for Automatic Error Monitoring and Control, by Robert L. Ashenhurst.

Notwithstanding Professor Todd's opinion, Professor Ashenhurst discusses what can be done in computer arithmetic in order that one may have at the end of a computation not only an alleged result but also some idea of its accuracy. He investigates fixed and floating-point arithmetic, introduces the coefficient error amplification factor associated with an arithmetic operation, and studies significance adjustment rules which attempt to keep this amplification factor near unity. The rest of the paper discusses unnormalized computer arithmetic, which has the required property and has been implemented, and some particular applications.

The Automatic Analysis and Control of Error in Digital Computation Based on the Use of Interval Numbers, by R. E. Moore.

Dr. Moore's paper starts with alternative ways of specifying an approximately known number, either as an exact number with a bound to the error, or as an interval within which the exact number is known to lie. The approximate number could be a piece of data or a computed number, and so it is seen that interval arithmetic and interval-valued functions are of great relevance to computer work. Interval arithmetic can be programmed and can take account of uncertainty in the data, finite precision of the computer arithmetic and analytic errors. Dr. Moore elaborates this idea and its consequences and shows its application to numerical integration including Gaussian quadrature and to the initial-value problem for ordinary differential equations. The paper culminates with an explanation of a computer program *DIFEQ* which accepts as data a specification of a system of ordinary differential equations and the initial values, and produces solution values with guaranteed bounds on the overall error. The method used requires the numerical evaluation of Taylor-series coefficients, and a part of the program generates from the system specification, a subroutine for doing this by recursion formulae.

Error in Digital Solution of Linear Problems, by E. L. Albasiny.

The paper by Mr. Albasiny is an account of some of the work of Mr. J. H. Wilkinson and its applications to the

(*Continued on p.* 371)