JOURNAL OF COMPUTATIONAL PHYSICS 20, 259–267 (1976)

Conservation Properties of Numerical Integration Methods for Systems of Ordinary Differential Equations*

JERROLD S. ROSENBAUM

ICASE, National Aeronautics and Space Administration, Langley Research Center, Hampton, Virginia 23665

Received November 22, 1974; revised January 21, 1975

If a system of ordinary differential equations represents a property conserving system that can be expressed linearly (e.g., conservation of mass), it is then desirable that the numerical integration method used conserve the same quantity. It is shown that both linear multistep methods and Runge–Kutta methods are "conservative" and that Newton-type methods used to solve the implicit equations preserve the inherent conservation of the numerical method. It is further shown that a method used by several authors is not conservative.

I. PRELIMINARIES

Consider a system of differential equations of the form

$$dy/dx = f(x, y), \qquad y(x_0) = \eta_0 , \tag{1}$$

where $y = (y_1 ,..., y_s)^T$ and $f = (f_1 ,..., f_s)^T$. Suppose that the solution satisfies one or more conservation laws of the form $w^T y = M$, where $w = (w_1 ,..., w_s)^T$, is a vector of constant weights. (This is a linear invariant of the system.) Then, by differentiating, we get

$$w^T f(x, y) = 0. \tag{2}$$

For the remainder of the paper, we assume that (2) is valid not only for the solution, but for all (x, y).

DEFINITION. Let (1) be a system of differential equations satisfying (2), and let $\{(x_n , y^n)\}$, $n = 0, 1, 2,...$, be a discrete numerical solution to (1) produced by

* This paper was a result of work performed under NASA Grant NGR 47-102-001 while the author was in residence at ICASE, NASA Langley Research Center.

a given method. The method is said to be conservative with respect to a given weight vector $w = (w_1, ..., w_s)^T$, if

$$w^T y^n = M, \qquad n = 0, 1, 2, ..., \tag{3}$$

where M is a constant.

For example, if the y_i represent the number densities of s chemical species and (2) holds with the $w_i = 1$, then the chemical reaction would conserve the total number density, $\sum_{i=1}^{s} y_i$.

The fact that many methods conserve linear invariants has several ramifications: (i) One can choose to integrate all the equations, or to eliminate one equation and integrate the remaining equations. The choice depends on the size of the system and whether or not the elimination of one equation destroys the sparsity of the system (or the Jacobian). (ii) The use of invariants to check the accuracy of integration is often misleading because all it usually checks for is roundoff error (and not truncation error). (iii) If the physical system satisfies a linear conservation law, then a numerical method that preserves the same law should be used, to eliminate one source of numerical error.

In Section II, it is shown that any consistent linear multistep, hybrid, or Runge–Kutta method is conservative; one of the nonconservative methods is discussed in Section III. In Section IV, it is shown that the use of certain Newton-type methods used to solve implicit methods is also conservative if the method itself is conservative.

II. Conservation for Linear Multistep, Runge–Kutta, and Hybrid Methods

Consider a general linear multistep method [1], of degree k

$$\sum_{j=0}^{k} \alpha_{k-j} y^{n-j} = h \sum_{j=0}^{k} \beta_{k-j} f^{n-j}, \tag{4}$$

where $y^{n-j} = (y_1^{n-j}, ..., y_s^{n-j})^T$ and $f^{n-j} = f(x_{n-j}, y^{n-j})$ and h is the step size. We will show that if the conservation property is valid for preceding values of y, then the conservation property is also valid for y^n.

Theorem 1. *If $w^T y^l = M$ for $l = n - 1, n - 2, ..., n - k$, and if (2) holds for all (x, y), then $w^T y^n = M$, where y^n has been computed by a consistent linear multistep method of degree k.*

Proof. It follows from (4) that

$$\sum_{j=0}^{k} \alpha_{k-j} w^T y^{n-j} = h \sum_{j=0}^{k} \beta_{k-j} w^T f^{n-j}. \tag{5}$$

From (2), $w^T f^l = 0$ and by our hypothesis, $w^T y^l = M$ for $n - k \leqslant l < n$; hence,

$$\alpha_k w^T y^n + \sum_{j=0}^{k-1} \alpha_j M = 0. \tag{6}$$

Any consistent linear multistep method satisfies $\sum_{j=0}^{k} \alpha_j = 0$; hence, (6) implies $w^T y^n = M$.

Consider a general m-stage Runge–Kutta method [1]

$$y^n = y^{n-1} + \sum_{j=1}^{m} a_j K^j, \qquad n = 1, 2,..., \tag{7}$$

where the K^j satisfy the system of equations

$$K^j = hf(x_{n-1} + c_j h, y^{n-1} + \sum_{l=1}^{m} b_{jl} K^l), \qquad j = 1,..., m. \tag{8}$$

From (2) and (8), it follows that $w^T K^j = 0$, and thus,

$$\begin{aligned} w^T y^n &= w^T y^{n-1} + \sum_{j=1}^{m} a_j w^T K^j, \qquad n = 1, 2,..., \\ &= w^T y^{n-1}. \end{aligned} \tag{9}$$

Thus, we have shown

THEOREM 2. *If $w^T y^{n-1} = M$ and (2) holds for all (x, y), then $w^T y^n = M$ for $n \geqslant 1$, where y^n has been computed by an m-stage Runge–Kutta method.*

Theorems 1 and 2 are known [6], but are repeated here because the results are not widely known.

Consider the hybrid method [4] of degree k

$$\sum_{j=0}^{k} \alpha_{k-j} y^{n-j} = h \sum_{j=0}^{k} \beta_{k-j} f^{n-j} + \beta_{k-\nu} f^{k-\nu}, \qquad 0 < \nu < 1.$$

THEOREM 3. *If $w^T y^l = M$ for $l = n - 1,..., n - k$, and if (2) holds for all (x, y), then $w^T y^n = M$, where y^n has been computed by a consistent hybrid method.*

The proof follows the same lines as Theorem 1.

III. Example of a Nonconservative Method

In finite-rate chemical reaction calculations, the following system of s equations is encountered.

$$dy_i/dx = P_i(x, y) - y_i L_i(x, y), \qquad i = 1,..., s, \tag{10}$$

where, as before, $y = (y_1,..., y_s)^T$ and $P_i(x, y)$ and $L_i(x, y)$, which are called production and loss functions, are given. Let us further suppose that (2) holds with $w_1 = \cdots = w_s = 1$; that is,

$$\sum_{i=1}^{s} [P_i(x, y) - y_i L_i(x, y)] = 0. \tag{11}$$

The following method has been used by several authors (see [2, 3]). Let

$$(y_i^{n+1} - y_i^n)/h = P_i^n - y_i^{n+1} L_i^n, \qquad i = 1,..., s, \tag{12}$$

where $P_i^n = P_i(x_n, y^n)$, $L_i^n = L_i(x_n, y^n)$, and $y^n = (y_1^n,..., y_s^n)^T$. Equation (12) can be rewritten as

$$y_i^{n+1} = y_i^n + h[P_i^n - y_i^n L_i^n] + hL_i^n[y_i^n - y_i^{n+1}]. \tag{13}$$

By (11), the method (12) is conservative if and only if

$$\sum_{i=1}^{s} L_i^n(y_i^n - y_i^{n+1}) = 0. \tag{14}$$

Equation (14) does not, however, hold in general, and the method is not conservative. For example, for the system

$$y_1' = -y_1, \qquad (P_1 = 0, L_1 = 0),$$
$$y_2' = +y_1, \qquad (P_2 = y_1, L_2 = 0),$$

the sum, in Eq. (14), is $y_1^{n+1} - y_1^n$.

It should be noted that method (12), can be made "almost conservative" and in the limit conservative, if one applies (12) as a successive substitution process. That is, one evaluates P_i and L_i using the latest available values for $\{y_i^{n+1}\}$ and iterates (12) until convergence. The net effect of the successive substitution process is to apply the Euler implicit scheme directly to (10). The cost of the process is likely to be several times more expensive than using the Euler implicit method with modified Newton–Raphson iterations to solve the implicit equations [1].

IV. The Effect of Iterative Methods on Conservation

Many linear multistep and Runge–Kutta methods are implicit. Consequently, it would be useful if the iterative method, which is used to solve the nonlinear equations at each time step, would preserve the conservation property.

Definition. An iterative method for solving equations is said to preserve the conservation property if all iterates, $y^{(p)}$, $p = 0, 1,...,$ satisfy $w^T y^{(p)} = M$, where $w^T = (w_1, ..., w_s)$.

In the case of predictor–corrector schemes using successive substitution, it follows from the proof of Theorem 1 that if both the predictor and corrector are conservative, then successive substitutions will preserve the conservation property. Similarly, it follows from the proof of Theorem 2 that for implicit Runge–Kutta methods, successive substitutions preserve the conservation property.

For stiff systems, Newton-type methods are usually used to solve the implicit equations at each step. In Theorem 4, we isolate a property of Newton-type methods which guarantees that the method preserves the conservation property. We then show that two common types of Newton methods, modified Newton and Broyden methods, preserve the conservation property for both linear multistep methods and Runge–Kutta methods.

Theorem 4. *Consider a system of s equations*

$$F(y) = 0 \tag{15}$$

and the Newton-type method

$$B_p\, \delta y^{(p)} = -F(y^{(p)}), \qquad p = 0, 1,..., \tag{16}$$

for solving (15), *where* B_p *is the nonsingular pth iteration matrix,* $\delta y^{(p)} = y^{(p+1)} - y^{(p)}$, *and* $y^{(p)}$ *is the pth approximation to the solution of* (15). *If* $w^T = (w_1, ..., w_s)$ *is a vector of weights such that*

$$w^T y^{(0)} = M, \tag{17}$$

$$w^T F(y) = 0, \qquad \textit{for all } y \textit{ such that } w^T y = M, \tag{18}$$

and

$$w^T B_p = \lambda w^T, \qquad p = 0, 1,..., \textit{for some scalar nonzero } \lambda, \tag{19}$$

then $w^T y^{(p)} = M$, $p = 0, 1,....$

Proof. From (16) and (18), we have that $w^T B_p\, \delta y^{(p)} = -w^T F(y^{(p)}) = 0$. Therefore, from (19), it follows that $\lambda w^T\, \delta y^{(p)} = 0$, or $w^T y^{(p+1)} = w^T y^{(p)}$, and by induction, the theorem is proved.

Consider the linear multistep method (4) applied to the differential Eqs. (1). To obtain the y^n, we must solve the nonlinear system of equations

$$F(y) \equiv y - h\beta_k f(x_n, y) + \sum_{j=1}^{k} (\alpha_{k-j} y^{n-j} - h\beta_{k-j} f^{n-j}) = 0, \tag{20}$$

where we have assumed that $\alpha_k = 1$. A "modified" Newton method for solving (20) can be written as

$$[I - h\beta_k f_y(\hat{x}^{(p)}, \hat{y}^{(p)})]\, \delta y^{n,p} = -F(y^{n,p}), \tag{21}$$

where $y^{n,p}$ is the pth approximation to y^n, $\delta y^{n,p} = y^{n,p+1} - y^{n,p}$, and $(\hat{x}^{(p)}, \hat{y}^{(p)})$ is any value of the independent and dependent variables (if we choose $(x_n, y^{n,p})$ we have the usual Newton method).

We may simultaneously consider the semi-implicit Runge–Kutta method

$$y^n = y^{n-1} + \sum_{j=1}^{m} a_j K^n, \qquad n = 1, 2,..., \tag{22}$$

where K^j is defined as the solution of the equations

$$F_j(K^j) \equiv K^j - hf\left(x_{n-1} + c_j h, y^{n-1} + \sum_{l=1}^{j} b_{jl} K^l\right) = 0, \qquad j = 1,..., m. \tag{23}$$

A Newton-type method for solving (23) can be written as

$$[I - hb_{jj} f_y(\hat{x}^{(p)}, \hat{y}^{(p)})]\, \delta K^{j,p} = -F_j(K^{j,p}) \tag{24}$$

and we define a sequence of approximations to y^n by

$$y^{n,p} = y^{n-1} + \sum_{j=1}^{m} a_j K^{j,p} \tag{25}$$

COROLLARY 1. *Suppose* $w^T y^{n,0} = M$.

(a) *If the linear multistep method* (4), *with* $\alpha_k = 1$, *is conservative* (*i.e., satisfies the hypotheses of Theorem* 1) *then all the modified Newton iterates* $y^{n,p}$, $p = 1, 2,...,$ *for the linear multistep method* (21) *satisfy* $w^T y^{n,p} = M$.

(b) *If* $w^T y^{n-1} = M$, *the semi-implicit Runge–Kutta method* (22) *and* (23), *is conservative, and* $w^T K^{j,0} = 0$, *then all the modified Newton iterates* $y^{n,p}$, $p = 1, 2,...$ *for the semi-implicit Runge–Kutta method* (24) *and* (25), *satisfy* $w^T y^{n,p} = M$.

Proof. We will proceed by an induction on p, for each fixed n. For the linear multistep method, we are given that (17) is valid and from Theorem 1 and its proof,

we have that (18) holds. From (2), we have that $w^T f_y(x, y) = 0$ for all (x, y) and, therefore, $w^T[I - h\beta_k f_y(\hat{x}, \hat{y})] = w^T$. Consequently, by Theorem 4, the corollary is proved for linear multistep methods.

For semiimplicit Runge–Kutta methods, we can similarly show that $w^T K^{j,p} = 0$. It then follows that $w^T y^{n,p} = w^T y^{n-1} + \sum_{j=1}^{m} a_j w^T K^{j,p}$ or $w^T y^{n,p} = w^T y^{n-1} = M$, and the corollary is proved.

Remark. $y^{n,0}$ can always be chosen such that $w^T y^{n,0} = M$. For instance, choose $y^{n,0} = y^{n-1}$. Similarly, we can satisfy $w^T K^{j,0} = 0$ by choosing $K^{j,0}$ to be the last set of K^j's from the calculation of y^{n-1} (or zero if $n = 1$).

Broyden's method [5] for solving (20) may be written as

$$B_p \, \delta y^{n,p} = -F(y^{n,p}), \tag{26}$$

where B_0 is a given approximation to the Jacobian of (20) and

$$B_{p+1} = B_p + ((\delta F^{n,p} - B_p \, \delta y^{n,p})(\delta y^{n,p})^T / \| \delta y^{n,p} \|^2), \tag{27}$$

$$\delta F^{n,p} = F(y^{n,p+1}) - F(y^{n,p}). \tag{28}$$

COROLLARY 2. *Suppose* $w^T y^{n,0} = M$ *and* $w^T B_0 = \lambda w^T$ *for some nonzero* λ.

(a) *If the linear multistep method* (4) *satisfies the hypotheses of Theorem* 1, *then all Broyden iterates for the linear multistep method satisfy* $w^T y^{n,p} = M$.

(b) *If the semi-implicit Runge–Kutta method* (22) *and* (23) *satisfies the hypotheses of Theorem* 2, *and* $w^T K^{j,0} = 0$, *then all Broyden iterates for the semiimplicit Runge–Kutta method satisfy* $w^T y^{n,p} = M$.

Proof. As in Corollary 1, we have that (17) and (18) are valid for the linear multistep method. We will now show that if $w^T B_p = \lambda w^T$, then $w^T B_{p+1} = \lambda w^T$. This is equivalent to showing that $w^T G_p = 0$, where

$$G_p = [\delta F^{n,p} - B_p \, \delta y^{n,p}].$$

From (28), we may write

$$\begin{aligned} w^T G_p &= [w^T F(y^{n,p+1}) - w^T F(y^{n,p}) - w^T B_p \, \delta y^{n,p}] \\ &= [-\lambda w^T \, \delta y^{n,p}] = 0. \end{aligned}$$

Therefore, by induction and Theorem 4, Corollary 2 is proved for linear multistep methods.

Proceeding as above, it can be shown that for the semi-implicit Runge–Kutta methods, the Broyden iterates for K^j satisfy $w^T K^{j,p} = 0$. Therefore, as in Corollary 1, we have that $w^T y^{n,p} = M$.

Remark. The condition $w^T B_0 = \lambda w^T$ can be satisfied if we choose, for example, $B_0 = I - \gamma f_y$ for some γ.

Corollaries 1 and 2 can be extended to include secant methods for solving the nonlinear equations for both linear multistep and semi-implicit Runge–Kutta schemes. However, the extension of the corollaries to fully implicit Runge–Kutta schemes, (7) and (8), is more complex. Successive substitutions would preserve the conservation property but Newton- and Broyden-type methods preserve the conservation if either the implicit Runge–Kutta coefficients are chosen correctly, or the nonlinear equations are solved in the appropriate ways (as described below).

For implicit Runge–Kutta schemes, Eq. (23) would have to be written as $F^j(K^1,\ldots, K^m) = 0$, a system of ms simultaneous nonlinear equations. If $w^T K^{j,0} = 0$, $j = 1,\ldots, m$, then Newton (or Broyden) iterations lead to $\sum_{j=1}^m w^T K^{j,\rho} = 0$, which does not necessarily imply that $\sum_{j=1}^m a_j w^T K^{j,\rho} = 0$ unless all the $\{a_j\}$ are equal. For example, for the method [4, p. 244]

$$y^{n+1} = y^n + (h/4)(K^1 + K^2),$$

$$K^1 = hf(x_n, y^n + (1/4)\, K^1 - (1/4)\, K^2),$$

$$K^2 = hf(x_n + (2/3)\, h, y^n + (1/4)\, K^1 + (5/12)\, K^2),$$

the use of the Newton or Broyden methods to solve the $2s$ equations for K^1 and K^2 does not preserve the conservation property.

However, if one uses a Gauss–Seidel–Newton (or Broyden or Secant) or a Jacobi–Newton (or Broyden or Secant) method [7] to solve Eq. (23), then the conservation property is preserved. The proof of this follows the same lines as Corollary 2, part (b).

Acknowledgments

The author is especially grateful to Dr. A. K. Cline and Professor C. W. Gear for their helpful suggestions, particularly for Section V. The author would also like to thank the reviewers for their suggestions and Dr. L. F. Shampine for reference [6].

References

1. C. W. Gear, "Numerical Initial Value Problems in Ordinary Differential Equations," Prentice–Hall, Englewood Cliffs, N. J., 1971.
2. T. Shimazaki and A. R. Laird, *J. Geophys. Res.* **75** (1970), 3221.
3. W. Stewart and M. J. Hoffert, Stratospheric contamination experiments with a one-dimensional athmospheric model, International Conference on Environmental Impact of Aerospace Operations in the High Athmosphere, Denver, Colorado, June, 1973.

4. J. D. LAMBERT, "Computational Methods in Ordinary Differential Equations," Wiley, New York, 1973.
5. C. BROYDEN, J. DENNIS, AND J. MORE, *J. Inst. Math. Appl.* **12** (1973), 223.
6. H. H. ROBERTSON AND M. J. MCCANN, *Computer J.* **12** (1969), 81.
7. J. M. ORTEGA AND W. C. RHEINBOLDT, "Iterative Solution of Nonlinear Equations in Several Variables," Academic Press, New York, 1970.