ELSEVIER Journal of Computational and Applied Mathematics 62 (1995) 129–154

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

Störmer–Cowell: straight, summed and split. An overview

J.F. Frankena

Department of Applied Mathematics, Twente University, P.O. Box 217, 7500 AE Enschede, Netherlands

Received 22 November1993; revised 26 July 1994

Abstract

In this paper we consider the relationship between some (forms of) specific numerical methods for (second-order) initial value problems. In particular, the Störmer–Cowell method in *second-sum form* is shown to be the *Gauss–Jackson* method (and analogously, for the sake of completeness, we relate Adams–Bashforth–Moulton methods to their first-sum forms). Furthermore, we consider the *split* form of the Störmer–Cowell method. The reason for this consideration is the fact that these summed and split forms exhibit a better behaviour with respect to rounding errors than the original method (whether in difference or in ordinate notation). Numerical evidence will support the formal proofs that have been given elsewhere.

Keywords: Ordinary differential equations; Periodic solutions; Initial value problems; Numerical methods; Multistep methods; Summed forms; Split forms

1. Introduction

Second-order ordinary differential equations have been integrated numerically ever since the 17th century, in the context of physical problems. In particular, the equations of celestial mechanics have been considered and integrated succesfully since several centuries. Present techniques in numerical astronomy date back to Gauss and have been put in a familiar form by a.o. Störmer (1907), Cowell (1910) and Jackson (1925). Cowell's method has reached an official status among well-known numerical techniques, while the Gauss–Jackson method is particularly known among astronomers. The latter method has recently (1986–1988) been used by Milani et al. to integrate the solar outer planetary system over 100 million years. Yet its features are little known among numerical analysts, let alone the relation of this method with the Störmer–Cowell method as a predictor–corrector pair. The purpose of our present investigation is to clarify this relation. Moreover, we shall compare the technique of "summation" to that of "splitting", as introduced by Spijker [11,12].

SSDI 0377-0427(94)00102-2

2. The Störmer and Cowell methods

The terminology around Cowell methods has been somewhat loose, in literature. Some authors designate all methods of the form

$$y_{n+1} - 2y_n + y_{n-1} = h^2.(\dots)$$

(where $(\dots)$ contains some linear combination of $f_{n+1}, f_n, \dots$ or their differences) as Störmer–Cowell methods. On the other hand, especially among astronomers not much distinction is made between the (original) Cowell method and the Gauss–Jackson method, which, by the way, can very simply be transformed into Cowell's, and vice versa.

To make things explicit, we shall stick to strict definitions and adopt Henrici's formulation of the Störmer and Cowell methods and Herrick's definition of the Gauss–Jackson method, see [5, 6].

We consider initial value problems for a set of second-order ODEs:

$$y''(x) = f(x, y(x)); \qquad y(x_0) = y_0, \quad y'(x_0) = y_{10}, \tag{1}$$

in which f is a continuous mapping from $I \times U \subset \mathbb{R} \times \mathbb{R}^m$ to $\mathbb{R}^m$, and $y_0, y_{10} \in U$. For this type of problems, numerical methods of the "Störmer–Cowell family" (see [2, 3]) can be used. For simplicity, we take $m = 1$.

On the x-axis we suppose an equidistant grid is given with step length h. Let x be a typical grid point and integrate (1) twice, then

$$y(x+h) - y(x) = hy'(x) + \int_x^{x+h} (x+h-t) f(t, y(t))\,\mathrm{d}t. \tag{2}$$

Doing the same with h replaced with $-h$ and adding both results we arrive at

$$y(x+h) - 2y(x) + y(x-h) = \int_x^{x+h} (x+h-t)\,\{f(t) + f(2x-t)\}\,\mathrm{d}t, \tag{3}$$

where "$f(t)$" is an abbreviation of "$f(t, y(t))$". At this stage an interpolating polynomial for f(interpolating at $q+1$ points) is introduced, whereupon it is straightforward (see [5]) to arrive at the *explicit* difference equation

$$y_{p+1} - 2y_p + y_{p-1} = h^2 \sum_{m=0}^{q} \sigma_m \nabla^m f_p, \tag{4}$$

with the coefficients of Table 1.

If we now use the relationship

$$\nabla^m f_p = \nabla^m f_{p+1} - \nabla^{m+1} f_{p+1}, \tag{5}$$

we arrive easily at the corresponding *implicit* Cowell method

$$y_{p+1} - 2y_p + y_{p-1} = h^2 \sum_{m=0}^{q} \sigma_m^* \nabla^m f_{p+1}, \tag{6}$$

in which $\sigma_0^* = \sigma_0$; $\sigma_m^* = \sigma_m - \sigma_{m-1}$ $(m \geqslant 1)$. We then get the coefficients as shown in Table 2.

Table 1
Coefficients of the Störmer method

m	0	1	2	3	4	5	6	$\cdots$
σ_m	1	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{19}{240}$	$\frac{18}{240}$	$\frac{863}{12\,096}$	$\cdots$

Table 2
Coefficients of the Cowell method

m	0	1	2	3	4	5	6	$\cdots$
σ_m^*	1	-1	$\frac{1}{12}$	0	$-\frac{1}{240}$	$-\frac{1}{240}$	$-\frac{221}{60\,480}$	$\cdots$

The concept of order is introduced in [5] in connection with the *ordinates* notation for multistep methods. The ordinate and difference notations can be transformed into one another by means of the formulas

$$\nabla^q f_p = \sum_{m=0}^{q} (-1)^m \binom{q}{m} f_{p-m} \quad (q = 0, 1, \dots) \tag{7}$$

and

$$f_{p-q} = \sum_{m=0}^{q} (-1)^m \binom{q}{m} \nabla^m f_p \quad (q = 0, 1, \dots). \tag{8}$$

It is easily seen that a Störmer method in difference form, used with maximum difference order q, is a $(q+1)$-step method when regarded in ordinate form. For Cowell methods, q and the step number k are equal.

With this in mind, we arrive at the survey of Table 3 (see [5]):

3. The first- and second-sum (Gauss–Jackson) methods

The following presentation of the Gauss–Jackson method has been inspired by the book on Astrodynamics by Herrick [6]. Extensive use will be made of the difference tables displayed in Tables 4 and 5. They will, if necessary, be supplied with "artificial" or "average" differences

$$\begin{aligned} \delta f_i &= \tfrac{1}{2}(\delta f_{i+1/2} + \delta f_{i-1/2}), \\ \delta^3 f_i &= \tfrac{1}{2}(\delta^3 f_{i+1/2} + \delta^3 f_{i-1/2}), \text{ etc.,} \end{aligned} \tag{9}$$

$$\begin{aligned} \delta^2 f_{i+1/2} &= \tfrac{1}{2}(\delta^2 f_{i+1} + \delta^2 f_i), \\ \delta^4 f_{i+1/2} &= \tfrac{1}{2}(\delta^4 f_{i+1} + \delta^4 f_i), \text{ etc.,} \end{aligned} \tag{10}$$

Table 3
Orders and error constants for Störmer and Cowell

Method		Order	Error constant
Störmer	$q = 0$	2	$\frac{1}{12}$
	$q > 0$	$k = q + 1$	σ_{q+1}
Cowell	$q = 2$	4	$-\frac{1}{240}$
	$q > 2$	$k + 1 = q + 1$	σ^*_{q+1}

Table 4
Central difference table, extended with two "sum" columns

"Time"	2nd sum	1st sum	r.h.s.	Differences			
(x)	(Σ^2)	(Σ)	(f)	(δ)	(δ^2)	(δ^3)	(δ^4)
x_{i-2}	$\Sigma^2 f_{i-2}$		f_{i-2}		$\delta^2 f_{i-2}$		$\delta^4 f_{i-2}$
		$\Sigma f_{i-3/2}$		$\delta f_{i-3/2}$		$\delta^3 f_{i-3/2}$	
x_{i-1}	$\Sigma^2 f_{i-1}$		f_{i-1}		$\delta^2 f_{i-1}$		$\delta^4 f_{i-1}$
		$\Sigma f_{i-1/2}$		$\delta f_{i-1/2}$		$\delta^3 f_{i-1/2}$	
x_i	$\Sigma^2 f_i$		f_i		$\delta^2 f_i$		$\delta^4 f_i$
		$\Sigma f_{i+1/2}$		$\delta f_{i+1/2}$		$\delta^3 f_{i+1/2}$	
x_{i+1}	$\Sigma^2 f_{i+1}$		f_{i+1}		$\delta^2 f_{i+1}$		
		$\Sigma f_{i+3/2}$		$\delta f_{i+3/2}$			
x_{i+2}	$\Sigma^2 f_{i+2}$		f_{i+2}				
		$\Sigma f_{i+5/2}$					
x_{i+3}	$\Sigma^2 f_{i+3}$						

or similar sums,

$$\begin{aligned} &\textstyle\sum^2 f_{i+1/2} = \frac{1}{2}(\sum^2 f_{i+1} + \sum^2 f_i), \\ &\textstyle\sum f_i = \frac{1}{2}(\sum f_{i+1/2} + \sum f_{i-1/2}), \text{ etc.} \end{aligned} \tag{11}$$

In this paper we use standard notations for differences: the *forward* difference is denoted by Δ, the *central* difference by δ and the *backward* difference by ∇. Moreover,

$$\nabla^0 = \Delta^0 = \delta^0 = \textit{identity} \quad \text{and} \quad \textstyle\sum = \delta^{-1}$$

Table 5
Backward difference table, extended with two "sum" columns

"Time"	2nd sum	1st sum	r.h.s.	Differences			
(x)	(∇^{-2})	(∇)	(f)	(∇)	(∇^2)	(∇^3)	(∇^4)
x_{i-2}	$\nabla^{-2}f_{i-3}$		f_{i-2}		$\nabla^2 f_{i-1}$		$\nabla^4 f_i$
		$\nabla^{-1}f_{i-2}$		∇f_{i-1}		$\nabla^3 f_i$	
x_{i-1}	$\nabla^{-2}f_{i-2}$		f_{i-1}		$\nabla^2 f_i$		$\nabla^4 f_{i+1}$
		$\nabla^{-1}f_{i-1}$		∇f_i		$\nabla^3 f_{i+1}$	
x_i	$\nabla^{-2}f_{i-1}$		f_i		$\nabla^2 f_{i+1}$		$\nabla^4 f_{i+2}$
		$\nabla^{-1}f_i$		∇f_{i+1}		$\nabla^3 f_{i+2}$	
x_{i+1}	$\nabla^{-2}f_i$		f_{i+1}		$\nabla^2 f_{i+2}$		
		$\nabla^{-1}f_{i+1}$		∇f_{i+2}			
x_{i+2}	$\nabla^{-2}f_{i+1}$		f_{i+2}				
		$\nabla^{-1}f_{i+2}$					
x_{i+3}	$\nabla^{-2}f_{i+2}$						

by definition. $\sum$ and $\sum^2$ are called *first sum* and *second sum*, respectively. The difference tables are constructed according to the rule that for every three entries situated as follows:

$\otimes\ a$

$\qquad\otimes\ b$

$\otimes\ c$

we should have $a + b = c$.

Appropriate starting values should be given in order to proceed with this scheme. Suppose, for instance, that we want to work with a difference table like in Table 4. If (starting) values f_{i-1}, f_i, f_{i+1} and f_{i+2} are given, the triangle from that column up to and including $\delta^3 f_{i+1/2}$ can be filled. With one value given in each of the first two columns in the larger triangle, both columns $(\sum^2)$ and $(\sum)$ can also be filled. In this circumstance we may call the ascending diagonal through the entry $\sum^2 f_{i+3}$ the "last known diagonal". From these values, sometimes estimates are made of some of the values in the next (not shown) upward diagonal through the entry f_{i+3}. Therefore, the latter diagonal is called the *first estimated* or *first unknown* diagonal. We shall return to the manipulation of the difference tables later (see "Comment on the use of the Gauss–Jackson formula" at the end of this section).

In the following we consider the differential equation in (1)

$$y'' = f(x, y(x)),$$

although it is possible to include first derivatives on the r.h.s. With $x - x_0 = hn$ (where h is the (fixed) step length and $n \in \mathbb{Z}$) it follows that

$$y'_{i+n} = y'_i + h \int_0^n g(\tau)\,\mathrm{d}\tau, \tag{12}$$

$$y_{i+n} = y_i + nhy_i' + h^2 \int_0^n \int_0^t g(\tau)\,\mathrm{d}\tau\,\mathrm{d}t, \tag{13}$$

with $g(\tau) = f(x_i + h\tau, y(x_i + h\tau))$ and $y_i' = y'(x_i) = y'(x_0 + hi)$.

We consider the central difference $\delta^2 y_i$; using (12) and (13) we get

$$\delta^2 y_i = y_{i+1} - 2y_i + y_{i-1} = h^2\left(\int_0^1 \int_0^t g(\tau)\mathrm{d}\tau\mathrm{d}t + \int_0^{-1} \int_0^t g(\tau)\,\mathrm{d}\tau\,\mathrm{d}t \right). \tag{14}$$

It is because of the use of the central differences that the name of Gauss remains connected with the present method. In order to interpolate the integrand however, we do not quite use the *Gauss* interpolation formula but the average of the forward- and backward-Gauss interpolation formulae, known as the *Stirling* interpolation formula,

$$f_{i+n} = f_i + n\delta f_i + S_2\delta^2 f_i + 2S_3\delta^3 f_i + S_4\delta^4 f_i + 2S_5\delta^5 f_i + \cdots, \tag{15}$$

with

$$\begin{aligned} S_2 &= \frac{n^2}{2!}, & 2S_5 &= \frac{n(n^2-1)(n^2-4)}{5!}, \\ 2S_3 &= \frac{n(n^2-1)}{3!}, & S_6 &= \frac{n^2(n^2-1)\,(n^2-4)}{6!}, \\ S_4 &= \frac{n^2(n^2-1)}{4!}, & 2S_7 &= \frac{n(n^2-1)\,(n^2-4)\,(n^2-9)}{7!}, \text{ etc.} \end{aligned} \tag{16}$$

In (15), f_i indicates $f(x_i, y(x_i))$. In the following, f_{i+t} is the value of f at the point $x_0 + ht$, where t is any real number between 0 and n. So if we substitute (16) into the integrals in (14), the coefficients S_k are supposed to be dependent upon the integration variable t and are to be integrated twice, between zero and n. This gives

$$\begin{aligned} \int_0^n \int_0^t g\,\mathrm{d}\tau\,\mathrm{d}t = {}& \tfrac{1}{2}n^2 f_i + \frac{1}{3!}n^3\delta f_i + \frac{1}{4!}n^4\delta^2 f_i + \frac{1}{5!}\left(n^5 - \frac{10}{3}n^3\right)\delta^3 f_i \\ & + \frac{1}{6!}\left(n^6 - \frac{5}{2}n^4\right)\delta^4 f_i + \frac{1}{7!}\left(n^7 - \frac{21}{2}n^5 + 28n^3\right)\delta^5 f_i + \cdots. \end{aligned} \tag{17}$$

Inserting $n = \pm 1$ we get

$$\int_0^{\pm 1} \int_0^t g\,\mathrm{d}\tau\,\mathrm{d}t = \frac{1}{2}f_i \pm \frac{1}{6}\delta f_i + \frac{1}{24}\delta^2 f_i \mp \frac{7}{360}\delta^3 f_i - \frac{1}{480}\delta^4 f_i \pm \frac{37}{10\,080}\delta^5 f_i + \cdots . \tag{18}$$

Substitution of (18) into (14) yields the "δ_c^2" or *central second difference integration formula*

$$\begin{aligned} \delta^2 y_i &= y_{i+1} - 2y_i + y_{i-1} \\ &= h^2\left(f_i + \frac{1}{12}\delta^2 f_i - \frac{1}{240}\delta^4 f_i + \frac{31}{60\,480}\delta^6 f_i - \frac{289}{3\,628\,800}\delta^8 f_i + \cdots \right). \end{aligned} \tag{19}$$

Since we have used the Stirling interpolation polynomial it is clear that the δ_c^2 -*formula, if truncated after the δ^n-term, is exact for polynomials of degree* $\leqslant n$.

Here we remark that there is an analogous formula for the first derivative, the "δ_c"- *or central first difference integration formula*

$$\delta y'_{i+1/2} = h\left(f_i + \frac{1}{2}\delta f_i + \frac{1}{6}\delta^2 f_i - \frac{1}{24}\delta^3 f_i - \frac{1}{180}\delta^4 f_i + \cdots\right), \tag{20}$$

which we would need in the case of a r.h.s. of the form $f(x, y(x), y'(x))$. Also it should be mentioned that the "δ_c^2" and its companion formula "δ_c" have their simplest form when written in terms of central differences, as in (19) and (20) (see also [2, Appendix B]).

Now it is quite easy to obtain several well-known formulae from the central difference formulae (19) and (20). The easiest way to do this is to use the formal operator calculus (see e.g. [7, Ch. 5] for details), in which

$$\delta^{-2} = \nabla^{-2}E^{-1} = \textstyle\sum^2,$$

E being the shift operator: $Ef_i = f_{i+1}$. If we apply the operator $\sum^2 E$ to both sides of (19), the result is the *Gauss–Jackson-*, "$\sum^2$"- or "*second-sum*" *integration formula* (*in central difference notation*)

$$y_{i+1} = h^2\left(\textstyle\sum^2 f_{i+1} + \frac{1}{12}f_{i+1} - \frac{1}{240}\delta^2 f_{i+1} + \frac{31}{60\,480}\delta^4 f_{i+1} - \frac{289}{3\,628\,800}\delta^6 f_{i+1} + \cdots\right). \tag{21}$$

Analogously we obtain from (20), by applying $\sum E^{1/2}$, the "$\sum_c$"- or "*first-sum*" *integration formula*

$$y'_{i+1} = h\left(\textstyle\sum f_{i+1/2} + \frac{1}{2}f_{i+1/2} + \frac{1}{6}\delta f_{i+1/2} - \frac{1}{24}\delta^2 f_{i+1/2} + \cdots\right). \tag{22}$$

Formulae (21) and (22) are the main results of this section. With the help of the operator calculus and manipulations in the difference tables however, we shall be able to elucidate the relationship of these formulae with the Störmer and Cowell formulae of the preceding section. To this end we derive, in the next section, a few other formulae connected with the first- and second-sum formulae.

Comment on the use of the Gauss–Jackson formula

Here we indicate globally how the Gauss–Jackson formula or rather the central difference table supplied with a first- and a second-sum column may be "run". This matter is discussed extensively in [6]. Suppose (see Table 4) we want to work with at most fourth-order central differences and initial values in $x = 0$ are given. Then starting values in $x_{-2}, \ldots, x_2$ must somehow be supplied (for instance, with the aid of a Runge–Kutta method), yielding $f_{-2}, \ldots, f_2$. From this, we have all values in the triangle with vertices f_{-2}, f_2 and $\delta^4 f_0$. By applying the first- and second-sum formulae *in reverse*:

$$\textstyle\sum^2 f_0 = h^{-2}y_0 - \frac{1}{12}f_0 + \frac{1}{240}\delta^2 f_0 - \frac{31}{60\,480}\delta^4 f_0 + \cdots \tag{23}$$

and

$$\Sigma f_0 = h^{-1} y'_0 + \frac{1}{12}\delta f_0 - \frac{11}{720}\delta^3 f_0 + \cdots, \tag{24}$$

we can generate starting values in the (Σ^2) and (Σ)-columns, respectively, with which we can "complete" the triangle up to and including the values $\Sigma^2 f_{-2}, \ldots, \Sigma^2 f_3$ and $\Sigma f_{-3/2}, \ldots, \Sigma f_{5/2}$. Now the values $y_{-2}, \ldots, y_2$ may be recomputed using Gauss–Jackson (21), where additional values $\delta^2 f_{-2}$ and $\delta^2 f_2$ have been estimated from the table itself. From this we might recompute $f_{-2}, \ldots, f_2$ and go through (23) and (24) again, obtaining a *revised* starting table. Going twice through the "reverse sum-formulae" might do in this case, but generally it depends upon the accuracy of the starting values $f_{-k}, \ldots, f_k$, the number of difference columns used and the desired table accuracy.

Once the table has been initiated we have a "last known" ascending diagonal, in this case e.g. the one through the entry f_{i+2}. If we, for proper accuracy reasons, agree in using (24) with the three terms given, we can proceed with the table (the "step-by-step integration procedure") in the following way.

From the table, deduce

$$\delta^3 f_{i+3/2} = \delta^3 f_{i+1/2} + \delta^4 f_i + \cdots \approx \delta^3 f_{i+1/2} + \delta^4 f_i,$$

and hence

$$\delta^2 f_{i+2} = \delta^2 f_{i+1} + \delta^3 f_{i+3/2},$$

so,

$$f_{i+3} = f_{i+2} + \delta f_{i+5/2} = f_{i+2} + \delta f_{i+3/2} + \delta^2 f_{i+2},$$

and finally

$$\delta^2 f_{i+3} = \delta^2 f_{i+2} + \delta^3 f_{i+3/2} + \cdots \approx \delta^2 f_{i+2} + \delta^3 f_{i+3/2}.$$

Now (24) gives

$$y_{i+3} \approx h^2 \left(\Sigma^2 f_{i+3} + \frac{1}{12} - \frac{1}{240}\delta^2 f_i + 3 \right) \tag{25}$$

from which a new value of f_{i+3} may be computed. With this recomputed value, the whole ascending diagonal through this entry can be computed. If a second recomputing would be necessary, it would start again with $\delta^2 f_{i+2} = \delta^2 f_{i+1} + \delta^3 f_{i+3/2}$, using the recomputed value of the entry $\delta^3 f_{i+3/2}$.

4. Störmer–Cowell, first and second sum (Gauss–Jackson), Adams–Bashforth–Moulton: all in the family

Consider the Difference Table 4 and suppose it to be filled, up to and including the upward diagonal through f_i. Then it is not possible to use (21) because f_{i+1}, $\delta^2 f_{i+1}$, $\delta^4 f_{i+1}$, etc., are not yet known. All these entries can however be "summed" to *the last known diagonal* (that is, the one through f_i):

$$\begin{aligned} f_{i+1} &= f_i + \delta f_{i+1/2} = f_i + \delta f_{i-1/2} + \delta^2 f_i \\ &= \ldots = f_i + \delta f_{i-1/2} + \delta^2 f_{i-1} + \delta^3 f_{i-3/2} + \delta^4 f_{i-2} + \delta^5 f_{i-5/2} + \cdots \end{aligned}$$

and analogously

$$\begin{aligned} \delta^2 f_{i+1} &= \delta^2 f_i + \delta^3 f_{i+1/2} \\ &= \ldots = \delta^2 f_{i-1} + 2\delta^3 f_{i-3/2} + 3\delta^4 f_{i-2} + 4\delta^5 f_{i-5/2} + 5\delta^6 f_{i-3} + \cdots, \\ \delta^4 f_{i+1} &= \ldots = \delta^4 f_{i-2} + 3\delta^5 f_{i-5/2} + 6\delta^6 f_{i-3} + \cdots, \\ \delta^6 f_{i+1} &= \ldots = \delta^6 f_{i-3} + \cdots, \end{aligned}$$

where we have neglected seventh and higher differences. When inserted into (21) this gives *the backward second-sum formula* (*in central difference notation*!) "Σ_b^2" (see Note on Notations at the end of this section):

$$\begin{aligned} y_{i+1} = h^2 \Big(&\Sigma^2 f_{i+1} + \frac{1}{12} f_i + \frac{1}{12} \delta f_{i-1/2} + \frac{19}{240} \delta^2 f_{i-1} + \frac{18}{240} \delta^3 f_{i-3/2} \\ &+ \frac{1726}{24\,192} \delta^4 f_{i-2} + \frac{1650}{24\,192} \delta^5 f_{i-5/2} + \cdots \Big), \end{aligned} \tag{26}$$

which because of its relationship to Störmer's formula, we call the "*Σ_b^2-Störmer*" *formula*. It is of course more appropriate to write this formula in *backward* differences, because it is a *backward formula*:

$$y_{i+1} = h^2 \left(\nabla^{-2} f_i + \frac{1}{12} f_i + \frac{1}{12} \nabla f_i + \frac{19}{240} \nabla^2 f_i + \frac{18}{240} \nabla^3 f_i + \frac{1726}{24\,192} \nabla^4 f_i + \frac{1650}{24\,192} \nabla^5 f_i + \cdots \right). \tag{27}$$

We note that this an *explicit* formula, and we shall call it "*∇_b^2-Störmer*" for later reference. Instead of "summing" unknown entries to the last *known diagonal*, we could rewrite all or all but the first entries in the r.h.s. *to the same diagonal through* f_{i+1}, which yields the "*Σ_b^2-Cowell*" *formula* (in central difference notation):

$$y_{i+1} = h^2 \left(\Sigma^2 f_{i+1} + \frac{1}{12} f_{i+1} - \frac{1}{240} \delta^2 f_i - \frac{1}{240} \delta^3 f_{i-1/2} - \frac{884}{241\,920} \delta^4 f_{i-1} + \cdots \right), \tag{28}$$

which reads in *backward* difference notation

$$y_{i+1} = h^2 \left(\nabla^{-2} f_i + \frac{1}{12} f_{i+1} - \frac{1}{240} \nabla^2 f_{i+1} - \frac{1}{240} \nabla^3 f_{i+1} - \frac{884}{241\,920} \nabla^4 f_{i+1} + \cdots \right). \tag{29}$$

It is an *implicit* formula, which we shall call "*∇_b^{-2}-Cowell*" (this formula and (27) have been used in [1, 9, 10].

We note that (29) also follows from (27) because of (5) and analogously, (28) follows from (26) due to

$$\delta^m f_p = \delta^m f_{p+1} - \delta^{m+1} f_{p+1/2}, \qquad \delta^m f_{p-1/2} = \delta^m f_{p+1/2} - \delta^{m+1} f_p. \tag{30}$$

So far, from the $\sum_c^2$-formula (Gauss–Jackson) we derived

- (26), $\sum_b^2$-Störmer;
- (27), ∇_b^{-2}-Störmer;
- (28), $\sum_b^2$-Cowell;
- (29), ∇_b^{-2}-Cowell.

We can handle the δ_c^2-formula (19) in completely the same way. Summing to the last known diagonal in the r.h.s. we find an *explicit* formula, δ_b^2 or *the backward second difference integration formula*, which we call "*δ_b^2-Störmer*":

$$\delta^2 y_i = h^2\Big(f_i + \frac{1}{12}\delta^2 f_{i-1} + \frac{1}{12}\delta^3 f_{i-3/2} + \frac{19}{240}\delta^4 f_{i-2} + \frac{18}{240}\delta^5 f_{i-5/2} + \frac{1726}{24\,192}\delta^6 f_{i-3} + \frac{1650}{24\,192}\delta^7 f_{i-7/2} + \cdots\Big). \tag{31}$$

Of course, also this backward formula reads more easily in the proper notation, i.e. in *backward* differences:

$$\delta^2 y_i = h^2\Big(f_i + \frac{1}{12}\nabla^2 f_i + \frac{1}{12}\nabla^3 f_i + \frac{19}{240}\nabla^4 f_i + \frac{18}{240}\nabla^5 f_i + \frac{1726}{24\,192}\delta^6 f_i + \frac{1650}{24\,192}\nabla^7 f_i + \cdots\Big), \tag{32}$$

which is the "ordinary" Störmer formula (4).

If, instead of summing (19) to the last known diagonal, we rewrite all entries to the diagonal through $\delta^2 f_i$ we get an implicit formula like in (28):

$$\delta^2 y_i = h^2\Big(f_{i+1} - \delta f_{i+1/2} + \frac{1}{12}\delta^2 f_i - \frac{1}{240}\delta^4 f_{i-1} - \frac{1}{240}\delta^5 f_{i-3/2} - \frac{884}{241\,920}\delta^6 f_{i-2} - \frac{760}{241\,920}\delta^7 f_{i-5/2} - \cdots\Big), \tag{33}$$

the "*δ_b^2-Cowell*" formula, which in backward difference notation is nothing else but the "ordinary" Cowell method (6):

$$\delta^2 y_i = h^2\Big(f_{i+1} - \nabla f_{i+1} + \frac{1}{12}\nabla^2 f_{i+1} - \frac{1}{240}\nabla^4 f_{i+1} - \frac{1}{240}\nabla^5 f_{i+1} - \frac{884}{241\,920}\nabla^6 f_{i+1} - \frac{760}{241\,920}\nabla^7 f_{i+1} - \cdots\Big). \tag{34}$$

This formula comes also from (32) by applying (5). On the other hand, (33) is easily derived from (31), using (30).

Summarizing, from (19): "δ_c^2" we derived

- (31), δ_b^2-Störmer;
- (32), "ordinary" Störmer;
- (33), δ_b^2-Cowell;
- (34), "ordinary" Cowell.

Moreover, (19): "δ_c^2" was transformed into (21): "Σ_c^2" or Gauss–Jackson, by the application of the operator $\Sigma^2 E$. Now it is easily checked that the same is true of all the following pairs of formulae:

- (31) and (26),
- (32) and (27),
- (33) and (28),
- (34) and (29).

These transformations are completely invertible and since Σ^2 and E commute, the inverse of $\Sigma^2 E$ may be computed as $\Sigma^{-2}E^{-1}$. With all of this in mind, we arrive at the diagram given in Fig. 1.

In [5, p. 343], Henrici hints to this "algebraic equivalency" of the Cowell and Σ^2-methods. The present author, however, encountered no elaboration of these ideas anywhere in the literature.

To the author's knowledge, it is not mentioned anywhere either that a similar relational scheme may be constructed, involving the first central difference-, first-sum and Adams–Bashforth–Moulton formulae. Starting with the "δ_c" or *central first difference integration formula* (20), we may proceed as follows (see the Note on Notations at the end of this section).

Suppose for convenience that we apply (20) to a *first*-order equation with r.h.s. $f(x, y(x))$ and that we consider only a few terms:

$$\delta y_{i+1/2} = h(f_i + \tfrac{1}{2}\delta f_i + \tfrac{1}{6}\delta^2 f_i - \tfrac{1}{24}\delta^3 f_i - \tfrac{1}{180}\delta^4 f_i + \cdots). \tag{35}$$

As we mentioned in Section 3 we should now apply the operator $(\Sigma^2 E)^{1/2} = \Sigma E^{1/2}$ (remember the commutation property) to (35) in order to get the "Σ_c"- or *first-sum integration formula*

$$y_{i+1} = h(\Sigma f_{i+1/2} + \tfrac{1}{2} f_{i+1/2} + \tfrac{1}{6}\delta f_{i+1/2} - \tfrac{1}{24}\delta^2 f_{i+1/2} - \cdots). \tag{36}$$

Summation of (35) and (36), respectively, to the last known diagonal in the difference table gives the "δ_b"- or *backward first difference integration formula* or "δ_b-*Adams–Bashforth*" predictor formula

$$\delta y_{i+1/2} = h(f_i + \tfrac{1}{2}\delta f_{i-1/2} + \tfrac{5}{12}\delta^2 f_{i-1} + \tfrac{3}{8}\delta^3 f_{i-3/2} + \cdots) \tag{37}$$

and the "Σ_b-*Adams–Bashforth*" or *backward first-sum integration formula*

$$y_{i+1} = h(\Sigma f_{i+1/2} + \tfrac{1}{2} f_i + \tfrac{5}{12}\delta f_{i-1/2} + \tfrac{3}{8}\delta^2 f_{i-1} + \tfrac{251}{720}\delta^3 f_{i-3/2} + \cdots) \tag{38}$$

(both in central difference notation). Both (37) and (38) are more convenient in backward difference notation. So we have, from (37), the "ordinary" Adams–Bashforth

$$\delta y_{i+1/2} = h(f_i + \tfrac{1}{2}\nabla f_i + \tfrac{5}{12}\nabla^2 f_i + \tfrac{3}{8}\nabla^3 f_i + \cdots) \tag{39}$$

and, from (38), the *backward first-sum formula in backward difference notation* or "∇_b^{-1}-*Adams–Bashforth*"

$$y_{i+1} = h(\nabla^{-1} f_i + \tfrac{1}{2} f_i + \tfrac{5}{12}\nabla f_i + \tfrac{3}{8}\nabla^2 f_i + \tfrac{251}{720}\nabla^3 f_i + \cdots). \tag{40}$$

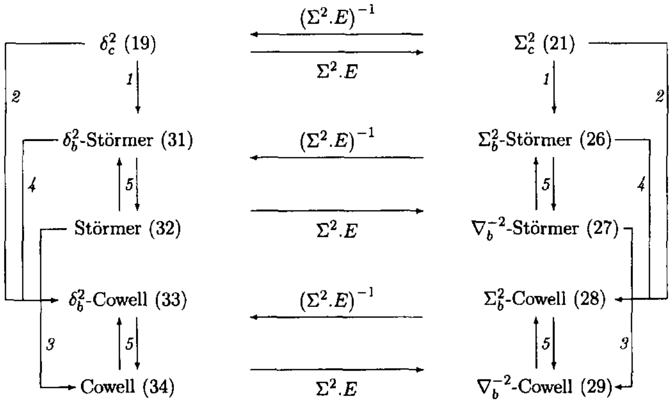

Fig. 1. Relationships in the Cowell family. *1*: sum to last known diagonal; *2*: rewrite to estimated or first unknown diagonal; *3*: apply (5); *4*: apply (30); *5*: replace central by backward differences and vice versa.

Next we can derive the implicit "δ_b-*Adams–Moulton*" either from "δ_c", rewriting the r.h.s. to the first unknown or estimated diagonal or from (37): "δ_b-*Adams–Bashforth*" by application of the relations (30):

$$\delta y_{i+1/2} = h\,(f_{i+1} - \tfrac{1}{2}\,\delta f_{i+1/2} - \tfrac{1}{12}\,\delta^2 f_i - \tfrac{1}{24}\,\delta^3 f_{i-1/2} - \cdots), \tag{41}$$

and of course we rewrite this immediately in terms of backward differences, obtaining the "ordinary" Adams–Moulton corrector formula:

$$\delta y_{i+1/2} = h\,(f_{i+1} - \tfrac{1}{2}\,\nabla f_{i+1} - \tfrac{1}{12}\,\nabla^2 f_{i+1} - \tfrac{1}{24}\,\nabla^3 f_{i+1} - \cdots). \tag{42}$$

Finally, the "summed" counterpart of (41), "Σ_b-*Adams–Moulton*" can be derived from "Σ_c", (36) by rewriting the r.h.s. to the first estimated or unknown diagonal or, alternatively, from "Σ_b-*Adams–Bashforth*" by application of (30) (or else, of course, by application of the operator $\Sigma E^{1/2}$ to "δ_b-*Adams–Moulton*" (41), as a third possibility):

$$y_{i+1} = h\,(\textstyle\sum f_{i+3/2} - \tfrac{1}{2} f_{i+1} - \tfrac{1}{12}\,\delta f_{i+1/2} - \tfrac{1}{24}\,\delta^2 f_i - \cdots), \tag{43}$$

which reads in backward difference notation, as "∇_b^{-1}-*Adams–Moulton*":

$$y_{i+1} = h\,(\nabla^{-1} f_{i+1} - \tfrac{1}{2} f_{i+1} - \tfrac{1}{12}\,\nabla f_{i+1} - \tfrac{1}{24}\,\nabla^2 f_{i+1} - \cdots), \tag{44}$$

Collecting our results for the family of "Adams-like" formulae we have:

(35): "δ_c"
(37): "δ_b-*Adams–Bashforth*"
(39): "*ordinary Adams–Bashforth*"
(41): "δ_b-*Adams–Moulton*"
(42): "*ordinary Adams–Moulton*"

(36): "Σ_c"
(38): "Σ_b-*Adams–Bashforth*"
(40): "∇_b^{-1}-*Adams–Bashforth*"
(43): "Σ_b-*Adams–Moulton*"
(44): "∇_b^{-1}-*Adams–Moulton*".

$$\begin{tikzcd}
\delta_c\ (35) \arrow[d, "1"'] \arrow[r, shift right, "(\Sigma^2.E)^{\frac{1}{2}}"'] \arrow[ddd, bend right=60, "2"'] & \Sigma_c\ (36) \arrow[l, shift right, "(\Sigma^2.E)^{-\frac{1}{2}}"'] \arrow[d, "1"'] \arrow[ddd, bend left=60, "2"] \\
\delta_b\text{-Ad.-Bas. } (37) \arrow[d, shift left, "5"] \arrow[dd, bend right=45, "4"'] & \Sigma_b\text{-Ad.-Bas. } (38) \arrow[l, "(\Sigma^2.E)^{-\frac{1}{2}}"'] \arrow[d, shift left, "5"] \arrow[dd, bend left=45, "4"] \\
\text{Ad.-Bas. } (39) \arrow[u, shift left] \arrow[r, "(\Sigma^2.E)^{\frac{1}{2}}"'] \arrow[dd, bend right=45, "3"'] & \nabla_b^{-1}\text{-Ad.-Bas. } (40) \arrow[u, shift left] \arrow[dd, bend left=45, "3"] \\
\delta_b\text{-Ad.-Moult. } (41) \arrow[d, shift left, "5"] & \Sigma_b\text{-Ad.-Moult. } (43) \arrow[l, "(\Sigma^2.E)^{-\frac{1}{2}}"'] \arrow[d, shift left, "5"] \\
\text{Ad.-Moult. } (42) \arrow[u, shift left] \arrow[r, "(\Sigma^2.E)^{\frac{1}{2}}"'] & \nabla_b^{-1}\text{-Ad.-Moult. } (44) \arrow[u, shift left]
\end{tikzcd}$$

Fig. 2. Relationships in the Adams family. *1*: sum to last known diagonal; *2*: rewrite to estimated or first unknown diagonal; *3*: apply (5); *4*: apply (30); *5*: replace central by backward differences and vice versa.

These results may be put into a diagram, analogous to that for the relationships between the Störmer–Cowell and second-sum formulae; see Fig. 2. Coefficients of the formulae mentioned in Figs. 1 and 2 are listed in [2, Appendix B].

Note on Notations

		r.h.s. of the formula		
Formula	l.h.s.	starts with	has character	notation used:
δ_c, δ_c^2	δ, δ^2	f	central	central difference
δ_b, δ_b^2	δ, δ^2	f	backwards	central difference
$\sum_c, \sum_c^2$	y	Σ, Σ^2	central	central difference
Σ_b, Σ_b^2	y	Σ, Σ^2	backwards	central difference
$\nabla_b^{-1}, \nabla_b^{-2}$	y	∇^{-1}, ∇^{-2}	backwards	backward difference
"ordinary"	δ, δ^2	f	backwards	backward difference

5. Summed forms, viewed constructively

In the preceding section we investigated several summed methods of the "Adams family" and of the "Störmer–Cowell family" from a rather formal point of view, using operators to establish correspondences between some multistep methods and their summed forms. Here we shall inspect the summed methods from a more constructive point of view, using their "definition" or generating principle. For single summation, the process has been described in [5, Section 6.4] (see also [3, Section 2]). We follow Henrici's formulation in the following outline.

Suppose

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h^2 \sum_{i=0}^{k} \beta_i f_{n+i} \tag{45}$$

(with stability polynomial $\rho(\zeta) + \lambda^2 h^2 \sigma(\zeta)$ for the test case $f = -\lambda^2 y$) is a consistent method, i.e. $\alpha_0 + \alpha_1 + \cdots + \alpha_k = 0$ (in which case α_k can be expressed as $-(\alpha_0 + \cdots + \alpha_{k-1})$ and similarly for $\alpha_k + \cdots + \alpha_{k-i}$, $i = 1, \dots, k-1$). Suppose we *sum* (45) from $n = 0$ to $n = N$ and let S_N be the result on both sides. Upon introducing new coefficients $\alpha'_{k-1}, \dots, \alpha'_0$ (using the consistency) by means of

$$\begin{aligned} \alpha'_{k-1} &= \alpha_k = -\sum_{v=0}^{k-1} \alpha_v, \\ \alpha'_{k-2} &= \alpha_k + \alpha_{k-1} = -\sum_{v=0}^{k-2} \alpha_v, \\ &\vdots \\ \alpha'_0 &= \alpha_k + \cdots + \alpha_1 = -\alpha_0, \end{aligned} \tag{46}$$

the left-hand side of the summation can be expressed in terms of "starting" and "final" values, as follows:

$$S_N = (\alpha'_{k-1} y_{N+k} + \cdots + \alpha'_0 y_{N+1}) - (\alpha'_{k-1} y_{k-1} + \cdots + \alpha'_0 y_0). \tag{47}$$

The right-hand side in the summation of (45) can be written as

$$\begin{aligned} S_N = h^2 \{&\beta_k(f_k + f_{k-1} + \cdots + f_{k+N}) + \beta_{k-1}(f_{k-1} + \cdots + f_{k-1+N}) \\ &+ \cdots + \beta_0(f_0 + \cdots + f_N)\}. \end{aligned} \tag{48}$$

Now we introduce the "indefinite sum" $\mathscr{F}_n$ with "summation constant" H by

$$\frac{h}{\alpha} \sum_{\mu=0}^{n} f_\mu = \mathscr{F}_n - H, \tag{49}$$

where α has to be chosen properly. Then (48) can be rewritten as

$$\begin{aligned} S_N = \alpha h \{&(\beta_k \mathscr{F}_{N+k} + \beta_{k-1} \mathscr{F}_{N+k-1} + \cdots + \beta_0 \mathscr{F}_N) \\ &- (\beta_k \mathscr{F}_{k-1} + \beta_{k-1} \mathscr{F}_{k-2} + \cdots + \beta_1 \mathscr{F}_0 + \beta_0 H)\}. \end{aligned} \tag{50}$$

The freedom we have in choosing H will now be exploited to get rid of the starting values in (47) and (50): we choose H such that

$$\begin{aligned} \alpha'_{k-1} y_{k-1} + \cdots + \alpha'_0 y_0 &= \alpha h (\beta_k \mathscr{F}_{k-1} + \beta_{k-1} \mathscr{F}_{k-2} + \cdots + \beta_1 \mathscr{F}_0 + \beta_0 H) \\ &= h^2 \{\beta_k f_{k-1} + (\beta_k + \beta_{k-1}) f_{k-2} + \cdots + (\beta_k + \beta_{k-1} + \cdots + \beta_1) f_0\} \\ &\quad + \alpha h (\beta_k + \cdots + \beta_0) H. \end{aligned} \tag{51}$$

Defining new coefficients β'_i similar to α'_i, the right-hand side of (51) may be written as

$$h^2\{\beta'_{k-1}f_{k-1} + \beta'_{k-2}f_{k-2} + \cdots + \beta'_0 f_0\} + \alpha h\sigma(1)H$$

if we denote, as usual, by $\sigma(\zeta)$ the polynomial corresponding to the right-hand side of (45). Since $\sigma(1) \neq 0$ it follows from (47)–(50) that we have chosen H as a function of the known starting values, leaving, from (47)–(51),

$$\begin{aligned} &\alpha'_{k-1}y_{N+k} + \cdots + \alpha'_0 y_{N+1} = \alpha h\{\beta_k \mathscr{F}_{N+k} + \cdots + \beta_0 \mathscr{F}_N\}, \\ &\mathscr{F}_{N+k} - \mathscr{F}_{N+k-1} = \frac{h}{\alpha} f_{N+k}, \quad \mathscr{F}_{-1} = H. \end{aligned} \tag{52}$$

This is a difference equation of the type used for *first*-order differential equations, with, e.g., the Adams methods as special cases. Once starting values for $y_0, \ldots, y_k$ are given, the values of $\mathscr{F}_0, \ldots, \mathscr{F}_k$ may be recursively computed and (52) may be solved.

We remark that the form (52) is *consistent*: one can show that its first characteristic polynomial is $\rho_\Sigma(\zeta) = \rho'(\zeta)$ and since the original method is supposed to be consistent, we have

$$\rho_\Sigma(1) = \rho'(1) = 0; \quad \rho'_\Sigma(1) = \rho''(1) = 2\sigma(1) \neq 0,$$

which gives exactly the consistency condition for the "first-order" method. The meaning of this is that we can follow the same procedure to establish a *second* summation. The *exact* solution of (52) satisfies the original Eq. (45) and vice versa. If both methods are used to approximate the solution of the initial value problem (1), they would yield the same truncation error. *Henrici* [5] *proved, however, that the accumulated rounding error of the summed form* (52) *is a factor of $\mathcal{O}(h)$ better than that of the original method, provided that ρ and σ define a stable and consistent method and $\zeta = 1$ is the only double zero of ρ on the unit circle.*

We illustrate this procedure by applying it to a simple case of the "δ_c^2"-method (see (19)), which is known as the *Numerov* method (in difference notation):

$$\delta^2 y_i = h^2(f_i + \tfrac{1}{12}\delta^2 f_i) \tag{53}$$

or (in ordinate notation)

$$y_{i+1} - 2y_i + y_{i-1} = h^2\{f_i + \tfrac{1}{12}(f_{i+1} - 2f_i + f_{i-1})\}. \tag{54}$$

Summing from $i = 0$ to $i = n$ we get (choose $\alpha = 1$)

$$\begin{aligned} y_{-1} - y_0 - y_n + y_{n+1} &= h^2\left\{\sum_{\nu=0}^{n} f_\nu + \tfrac{1}{12}(f_{-1} - f_0 - f_n + f_{n+1})\right\} \\ &= h^2\{h^{-1}(\mathscr{F}_n - H) + \tfrac{1}{12}(\delta f_{n+1/2} - \delta f_{-1/2})\}. \end{aligned}$$

If we now choose H such that

$$y_0 - y_{-1} = \delta y_{-1/2} = hH + \tfrac{1}{12}h^2\,\delta f_{-1/2}$$

then there remains the *summed δ_c^2-method*

$$\delta y_{n+1/2} = h\mathscr{F}_n + \tfrac{1}{12}h^2\,\delta f_{n+1/2}, \tag{55}$$

which we shall call "$\Sigma\delta_c^2$" and which is more appropriately written (see [2, 6]) in the form

$$\delta y_{n+1/2} = h^2\,\{\Sigma f_{n+1/2} + \tfrac{1}{12}\,\delta f_{n+1/2}\}. \tag{56}$$

The symbol $\Sigma f_{n+1/2}$ is called "the first sum of f at $n+\frac{1}{2}$"; it obviously equals $h^{-1}\mathscr{F}$.

Remark. Eq. (56) could have been established by applying the operator $E^{1/2}\Sigma$ to (53), which amounts to the formal "summation" of Section 4, last part, especially Fig. 2. On the other hand, by applying ∇ to (56) or "differencing" the latter equation, we get back (53). This shows that the formal and the constructive approach to "summation" are basically equivalent.

We can repeat the whole procedure to obtain the so-called "second-sum formula", in this case $\Sigma^2\delta_c^2$, by summing again, from $n=0$ to $n=N$. Writing (55) in ordinate form and summing from 0 to N we obtain

$$y_{N+1} - y_0 = h\sum_{\nu=0}^{N}\mathscr{F}_\nu + \tfrac{1}{12}h^2\,(f_{N+1} - f_0). \tag{57}$$

We introduce a second summation constant $\tilde{H}$ with the definition

$$h\sum_{\nu=0}^{N}\mathscr{F}_\nu = \tilde{\mathscr{F}}_N - \tilde{H} \tag{58}$$

and determine $\tilde{H}$ such that the starting values disappear from (57) (that is, in first instance):

$$\tilde{H} = y_0 - \tfrac{1}{12}h^2 f_0.$$

What remains is the apparently simple formula for $\Sigma^2\delta_c^2$:

$$y_{N+1} = \tilde{\mathscr{F}}_N + \tfrac{1}{12}h^2 f_{N+1}. \tag{59}$$

For this formula we have also the more appropriate notation

$$y_{N+1} = h^2\,\{\Sigma^2 f_{N+1} + \tfrac{1}{12}f_{N+1}\}, \tag{60}$$

in which the first term on the right-hand side is the so-called "second sum of f at $N+1$". It can easily be shown that, *in general*,

$$\Sigma^2 f_{N+1} = (N+1)\Sigma f_{-1/2} + \Sigma^2 f_0 + \Sigma_{\nu=0}^{N}(N+1-\nu)f_\nu, \tag{61}$$

where in our case H and $\tilde{H}$ have been chosen such that

$$\Sigma f_{-1/2} = h^{-2}\,(y_0 - y_{-1}) - \tfrac{1}{12}(f_0 - f_{-1}) \quad\text{and}\quad \Sigma^2 f_0 = h^{-2}y_0 - \tfrac{1}{12}f_0 \tag{62}$$

are the two starting values of the summation process.

Remark. Again, Eq. (60) can be "differenced" twice to yield (53), showing the equivalence of the formal and the constructive approach. The other way around requires proper choice of the summation constants H and $\tilde{H}$.

It follows from the definitions of Σf_n and $\Sigma^2 f_n$ that, for $n \geqslant 0$,

$$\Sigma^2 f_{n+1} = \Sigma^2 f_n + \Sigma f_{n+1/2}. \tag{63}$$

If we now *define* $\Sigma^2 f_{-1}$ by the same relation, it follows from (62) that (61) can also be written in the form

$$\Sigma^2 f_{N+1} = (N+2)\Sigma f_{-1/2} + \Sigma^2 f_{-1} + \sum_{\nu=0}^{N} (N+1-\nu) f_\nu. \tag{64}$$

Formula (63), extended if necessary to $n \geqslant -1$ (or even further back, if desired, always using (63)), can be used to generate the "first-sum" and "second-sum" columns on the left in the difference table, showing otherwise the values of f, δf, $\delta^2 f$, etc. (see Table 4 and, for backward differences, [2, Appendix A]). With sufficient starting values for f, *and* the above-mentioned starting values for Σf and $\Sigma^2 f$, a complete "sum and difference table" will be built up, using (63) and the corresponding relation for Σ:

$$\Sigma f_{n+1/2} = \Sigma f_{n-1/2} + f_n$$

and, of course, the well-known relations between f and its differences used in difference tables. From (61) or (62) it follows that, in using the second-sum formula (60), the starting values have ever growing coefficients with increasing N, and that *the complete history is "dragged along" with increasing weights for past values of f, as N increases.*

In more complicated cases like a sixth-order Gauss–Jackson method, the initiation of the sum columns proceeds along slightly different lines. Instead of the determination of H and $\tilde{H}$ to obtain the starting values (62), one uses the first- and second-sum formulas *in reverse*, see Section 3: "Comment on the use of the Gauss–Jackson formula", and also [6, 2].

As pointed out earlier, Henrici proved that summed forms have a computational advantage over the corresponding original multistep formulation: the effects of the propagation of round-off errors are diminished, thus stabilizing the original method. Thus, using second-sum methods, one might expect to have a benefit from the double summation over the corresponding multistep methods, which is quite useful in integration over long time intervals. However, although one might have expected a further increase of the stability by summing *twice*, this is not the case. Summing twice improves the stability not more than summing *once*. In fact, analogously to [5, Theorem 6.11], one can prove the following.

Theorem. *Suppose the numerical values y_n^*, $\mathscr{F}_n^*$ and $\tilde{\mathscr{F}}_n^*$, calculated with a second-sum algorithm satisfy the relations*

$$\begin{aligned} &\alpha''_{k-2} y^*_{n+k} + \cdots + \alpha''_0 y^*_{n+2} = \alpha^2 \{\beta_k \tilde{\mathscr{F}}^*_{n+k} + \cdots + \tilde{\mathscr{F}}^*_n\} + \bar{\varepsilon}_{n+k}, \\ &\nabla \tilde{\mathscr{F}}^*_{n+k} = \frac{h}{\alpha} \mathscr{F}(x_{n+k}, y^*_{n+k}) + \tilde{\eta}_{n+k}, \\ &\nabla \mathscr{F}^*_{n+k} = \frac{h}{\alpha} f(x_{n+k}, y^*_{n+k}) + \eta_{n+k}. \end{aligned} \tag{65}$$

If the only double root of ρ is on the unit circle and the local round-off errors satisfy

$$|\tilde{\varepsilon}_n| \leqslant \tilde{\varepsilon}, \quad |\eta_n| \leqslant \eta, \quad |\tilde{\eta}_n| \leqslant \tilde{\eta}, \quad n = 1, 2, \dots ,$$

then the accumulated round-off error r_n in the numerical solution of $y'' = f(x, y)$ by the second-sum formula satisfies, for $a \leqslant x_n \leqslant x_q$ and $h^2 \leqslant L^{-1} \| \alpha_k \beta_k^{-1} \|$,

$$|r_n| \leqslant K^* \exp\{(x_q - a^*)^2 \Gamma^* L B\},$$

with

$$K^* = \alpha B \frac{(x_q - a^*)}{h} \left\{ \Gamma^* (x_q - a^*) \frac{\eta}{2} + 2\Gamma^* \tilde{\varepsilon} + \alpha(\Gamma^* + \gamma^*)\tilde{\eta} \right\},$$

$$a^* = a - 2h \frac{\gamma^*}{\Gamma^*},$$

where most quantities have the same meaning as in [5, *Theorem* 6.11] *and where $K^* = \mathcal{O}(h^{-1})$. Both K^* and the exponential argument grow quadratically with increasing $x_q - a$. The factor K^* contains one more complex of terms which is linear in $(x_q - a)$ than the corresponding factor in* [5, *Theorem* 6.11].

The proof of this theorem, which runs along lines similar to those of Theorem 6.11 in [5], will be given elsewhere.

The last result mentioned in the above theorem indicates that summing twice might even be disadvantageous compared to summing once! Numerical evidence sustains this supposition (see Section 7). The second-sum forms have the further general disadvantage that one necessarily needs a (summed form of a) consistent difference equation for y' in order to initiate the first-sum column (there is no such *necessity* for the first-sum form, although it works quite well).

For the Störmer–Cowell family of methods, such a consistent difference equation is at hand (see Section 3 and (66), but for other methods (modified Cowell, modified Numerov, symmetric methods, etc.) a separate and somewhat lengthy derivation is necessary. Therefore, using second-sum forms, we should better restrict ourselves to the second-sum Cowell method, which is nothing else but *Gauss–Jackson* written in backward difference form, i.e. ∇_b^{-2}-Cowell (29), to be used in combination with the ∇_b^{-1}-formula for y',

$$y'_{i+1} = h(\nabla^{-1} f_{i+1} - \tfrac{1}{2} f_{i+1} - \tfrac{1}{12} \nabla f_{i+1} - \tfrac{1}{24} \nabla^2 f_{i+1} - \tfrac{19}{720} \nabla^3 f_{i+1} - \cdots). \tag{66}$$

In practical calculations, this second-sum method (used in the backward difference form in [1, 9, 10]), with

$$\nabla^{-2} f_{i+1} - \nabla^{-1} f_{i+1} = \nabla^{-2} f_i$$

appears indeed to be more stable than the corresponding Cowell method (see Milani et al. [10], who conducted the LONGSTOP project, and [3]). But as the above theorem shows, there is no gain at all in summing *twice* instead of *once*. Moreover, the first-sum form appears to be somewhat *more stable and faster* than the second-sum form. It seems that historical reasons have determined the continuing use of the Gauss–Jackson form.

Apart from summing only once, there is another way out of the restriction to Gauss–Jackson, as we shall see in the next section.

6. The split form according to Spijker

6.1. Definition of the split form

Spijker [11, 12] introduced a very useful alternative for the summed forms: the *split* form, which has the same benefits as the *summed* form with respect to the reduction of the error accumulation, but which is easier to implement and more generally applicable. Instead of Henrici's root condition (see Section 2) the split form requires only that "the roots of ρ have a modulus at most 1 and the multiplicity of the roots with modulus equal to 1 is at most 2". Under this condition any consistent method of the form (45) will admit of a splitting which produces, as do the summed forms, an error accumulation of the order $\mathcal{O}(h^{-1})$ as opposed to the $\mathcal{O}(h^{-2})$ of the original form (45) of the method.

The method (45) for the second-order IVP (1) will be written in the form

$$\rho(E)y_n = h^2\sigma(E)f_n, \tag{67}$$

where E is the shift operator and ρ and σ are, as usual, the polynomials corresponding to the left- and right-hand sides of (45). It is supposed that the method (67) is convergent. Now suppose that we have a *splitting* of the polynomials ρ and σ as follows:

$$\rho(\zeta) = \rho_2(\zeta)\rho_1(\zeta), \qquad \sigma(\zeta) = \sigma_1(\zeta)\sigma_2(\zeta) \tag{68}$$

and that $p, q \in \mathbb{R}$ are such that $p + q = 2$. Suppose furthermore that y_n and z_n satisfy the *split form* of Eq. (67), i.e.

$$\begin{aligned} \rho_1(E)y_n &= h^p\sigma_1(E)z_n, \\ \rho_2(E)z_n &= h^q\sigma_2(E)f(x_n, y_n). \end{aligned} \tag{69}$$

Then

$$\begin{aligned} \rho(E)y_n &= \rho_2(E)\rho_1(E)\ y_n = \rho_2(E)h^p\sigma_1(E)z_n \\ &= h^p\sigma_1(E)\sigma_2(E)z_n = h^{p+q}\sigma_1(E)\sigma_2(E)f(x_n, y_n), \end{aligned}$$

and therefore y_n satisfies (67) (if there are no round-off errors). So, with regard to convergence and truncation error, (67) and (69) are equivalent.

In the sequel, it is supposed that the unimodular roots of $\rho_1(\zeta) = 0$ are simple; there are no restrictions with regard to the roots of $\rho_2(\zeta) = 0$. Suppose a is a fixed but otherwise arbitrary positive number and we consider the interval $0 \leqslant x \leqslant a$.

Let y_n^* and z_n^* again be calculated values. If ξ_n and η_n are local round-off errors (satisfying $|\xi_n| \leqslant \xi$, $|\eta_n| \leqslant \eta$ for some positive real ξ and η) occurring in

$$\left.\begin{aligned} \rho_1(E)\ y_n^* &= h^p\sigma_1(E)z_n^* + \xi_n \\ \rho_2(E)z_n^* &= h^q\sigma_2(E)f(x_n, y_n^*) + \eta_n \end{aligned}\right. \quad (n = 0, 1, 2, \ldots),$$

then, as Spijker proved, starting with exact solutions of (69) on the nth subinterval, *there exist constants γ and h_1 such that*

$$|y_n^* - y_n| \leqslant \gamma h^{-1}(\xi + h^{p-1}\eta) \quad (n = 1, 2, \ldots; nh \leqslant a) \tag{70}$$

for all $h \in (0, h_1]$ (here γ and h_1 depend upon a and the Lipschitz constant L of f w.r.t. y). This is an improvement of a factor h compared with the corresponding round-off error bound for the original method (67) (see [5], formula (6-103) and also (6-155) for a result analogous to (70)). The proof is quite involved and is given in [11].

In the following, we give a brief survey of the split form of the Störmer–Cowell predictor–corrector method. For implicit methods like Cowell's, their always remains an implicit system of difference equations to be solved, which can be done iteratively in a quite natural way. With reference to (70) we remark that the freedom in the choice of the power p has been exploited to take $p = q = 1$, which seems to be optimal. (For more applications of Spijker's split form, see Frankena [3].)

6.2. *The split Störmer–Cowell method*

Referring to the ordinate version

$$y_{n+5} - 2y_{n+4} + y_{n+3} = \frac{h^2}{240}(18 f_{n+5} + 209 f_{n+4} + 4 f_{n+3} + 14 f_{n+2} - 6 f_{n+1} + f_n) \tag{71}$$

of Cowell's method it is clear that here

$$\rho(\zeta) = \zeta^3(\zeta - 1)^2,$$

$$\sigma(\zeta) = \frac{1}{240}(18\zeta^5 + 209\zeta^4 + 4\zeta^3 + 14\zeta^2 - 6\zeta + 1).$$

Since ρ_2 should have a simple zero $\zeta = 1$, we select the following split form of (71), with $p = q = 1$:

$$\begin{aligned} &\rho_1(\zeta) = \zeta(\zeta - 1), \quad \sigma_1(\zeta) = 1, \\ &\rho_2(\zeta) = \zeta^2(\zeta - 1), \quad \sigma_2(\zeta) = \tfrac{1}{240}(18\zeta^5 + 209\zeta^4 + 4\zeta^3 + 14\zeta^2 - 6\zeta + 1). \end{aligned} \tag{72}$$

This amounts to solving the following set of difference equations:

$$\begin{aligned} &y_n = y_{n-1} + h.z_{n-2}, \\ &z_{n-2} = z_{n-3} + \frac{h}{240}(18 f_n + 209 f_{n-1} + 4 f_{n-2} + 14 f_{n-3} - 6 f_{n-4} + f_{n-5}), \end{aligned} \tag{73}$$

while for the (sixth-order) Störmer predictor we get in the same way,

$$y_n = y_{n-1} + h.z_{n-2},$$

$$z_{n-2} = z_{n-3} + \frac{h}{240}(317 f_{n-1} - 266 f_{n-2} + 374 f_{n-3} - 276 f_{n-4} + 109 f_{n-5} - 18 f_{n-6}), \tag{74}$$

($n = 1, 2, \dots$), with appropriate initial values $y(-5), \dots, y(0)$, and z_{n-3} calculated from the first equation (74).

7. Numerical verification

The theoretical results of the preceding sections have been verified with two test problems:

Test 1. The unperturbed harmonic oscillator:

$$x'' + \omega^2 x = 0, \quad x(0) = 1, \quad x'(0) = 0;$$

$$y'' + \omega^2 y = 0, \quad y(0) = 0, \quad y'(0) = \omega;$$

with exact solution: $x(t) = \cos \omega t$, $y(t) = \sin \omega t$.

Test 2. The unperturbed Kepler equations in the plane (with $e =$ *eccentricity*, $0 < e < 1$):

$$x'' = -x/r^3, \quad x(0) = 1 - e, \quad x'(0) = 0\ ;$$

$$y'' = -y/r^3, \quad y(0) = 0, \quad y'(0) = \sqrt{\left(\frac{1+e}{1-e)}\right)}.$$

Using the *eccentric anomaly E*, defined by the set of equations

$$x = r\cos\theta, y = r\sin\theta \ (\theta \text{ being the } \textit{true anomaly}),$$

$$\tan\frac{\theta}{2}\sqrt{\left(\frac{1-e}{1+e}\right)} = \tan\frac{E}{2},$$

the exact solution can be written in the form:

$$x(E) = \cos E - e, \quad y(E) = \sqrt{(1-e^2)}\sin E.$$

The following numerical procedure has been followed. Tests 1 and 2 have been run in Turbo Pascal with sixth-order predictor–corrector formulae, step-sizes $h = \pi/500$ and $h = \pi/1000$ over 10^7 steps, for each of the following forms.

(A) *Difference form used*:
- Störmer–Cowell (method "0");
- First-sum form of Störmer–Cowell (method "1") in the form of Σ_b;
- Second-sum form of Störmer–Cowell (method "2"), in the form of Σ_b^2.

(B) *Ordinate form used*:
- Störmer–Cowell (method "or");
- Split form of Störmer–Cowell (method "sp").

The results have been compared to the exact solutions at $N = 10^n, n = 3, \dots, 7$. This resulted in an accumulated *rounding* error. (The only Turbo Pascal mode which admits of nonextended

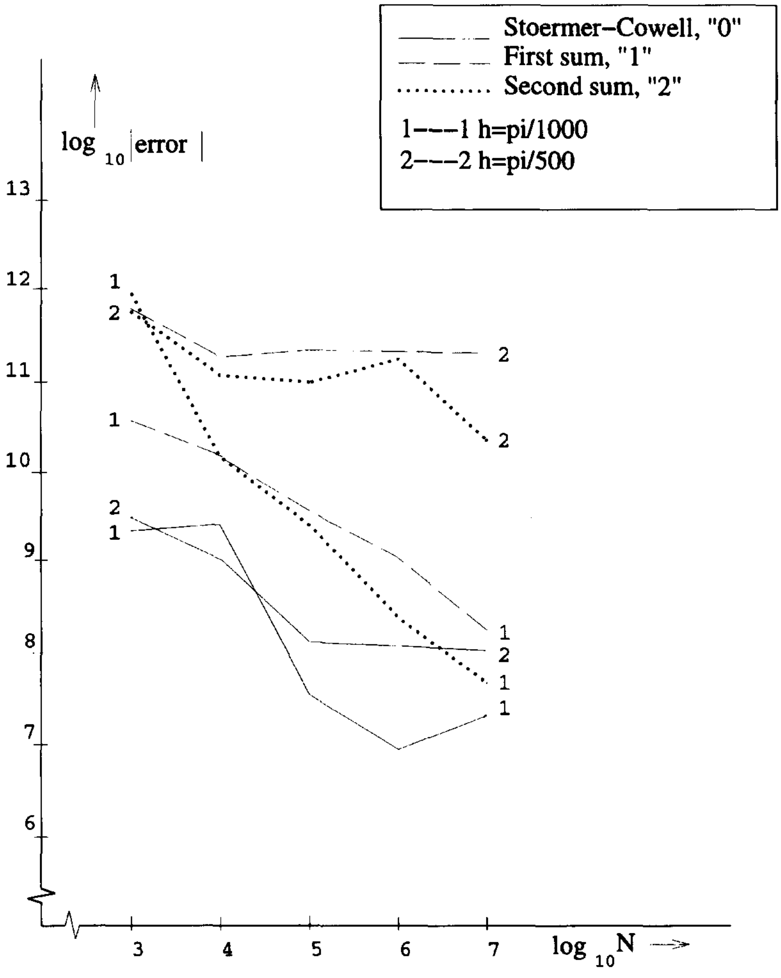

Fig. 3. Accumulated rounding errors, Test 1, $N = 10^7$, difference notation.

internal representation, however, is the *real* mode with a relative machine precision of about 10^{-12}. We use sixth-order methods with an accumulated truncation error of the order of magnitude $\mathcal{O}(10^{-16})$.)

There are three main conclusions from the theorem:

- *h-dependency*: For constant step number N, the theoretical error bounds are inversely proportional to the step length h;
- *N-dependency*: For constant step length h, the theoretical error bounds grow (faster than or at least) quadratically with increasing step number N;
- *summing once instead of twice*: The first-sum form may be expected to perform somewhat better than the second-sum form, because the factor K^* of the the latter's error bound is more complicated.

The experimental verification of these items is difficult because the error bounds are highly pessimistic. Assuming that the local errors $\tilde{\varepsilon}_n$, η_n and $\tilde{\eta}_n$ are of the order of the iteration tolerance (10^{-12}), a rough evaluation of the formula presented in the theorem indicates an error bound of the

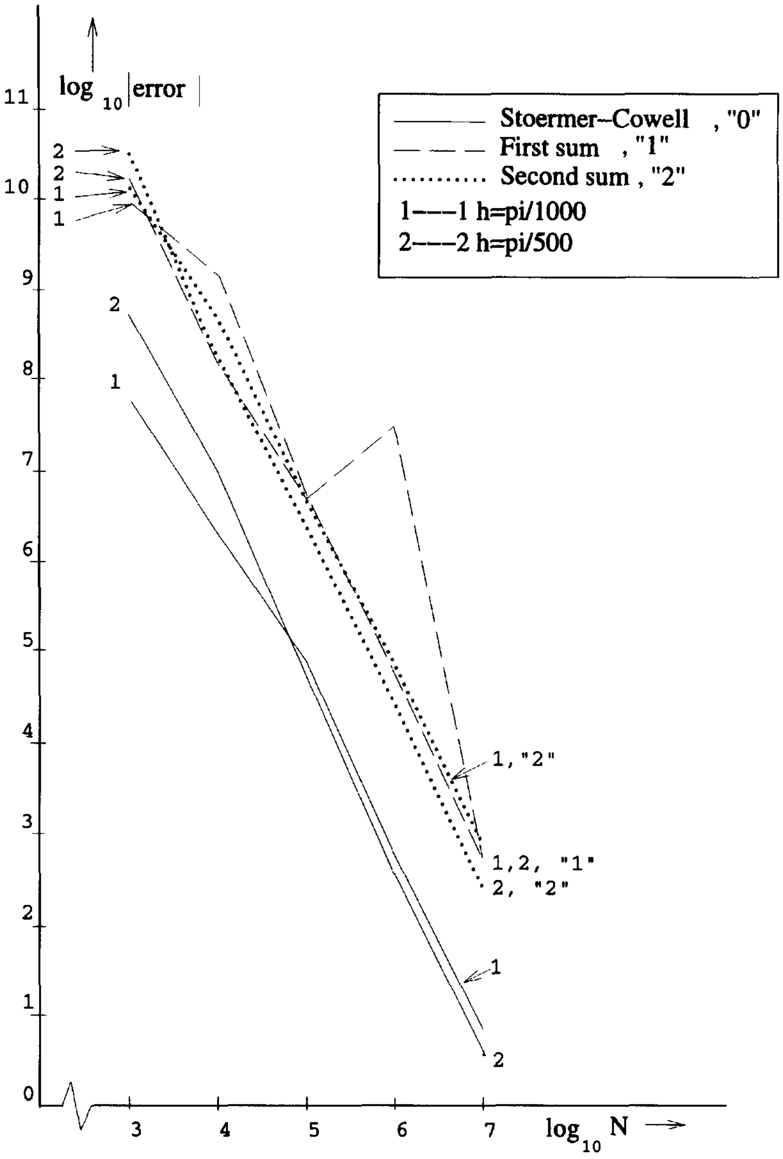

Fig. 4. Accumulated rounding errors, Test 2, $N = 10^7$ difference and ordinate notation, respectively.

order 10^{-5} in the case $h = \pi/500$, $x_q - a^* \approx 600$, for the first-sum Cowell form. The choice $h = \pi/500$ is good enough to suppress the *truncation errors*, in the time interval considered, relative to the accumulated rounding errors, but it leaves a considerable gap between the theoretical error bounds and the actual rounding errors. Nevertheless, from the experiments the following conclusions may be drawn, see Figs. 3 and 4.

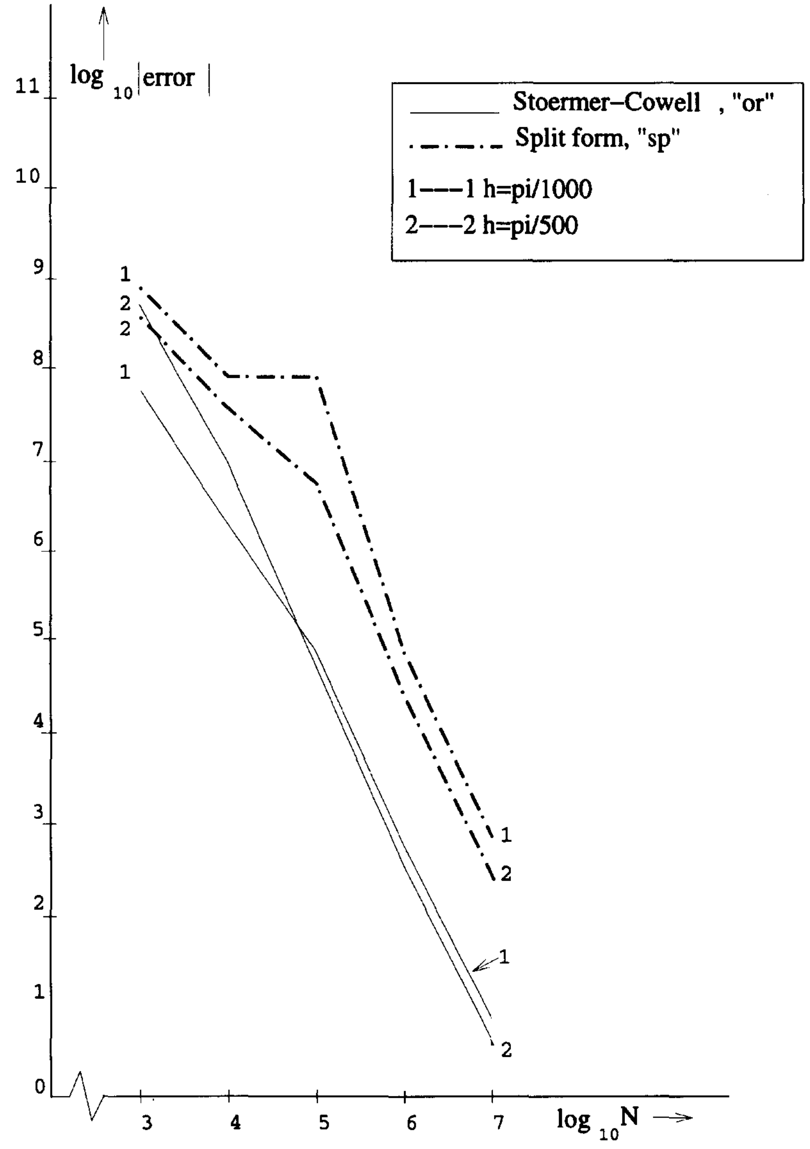

Fig. 4. Continued.

Unfortunately, the harmonic oscillator (Test 1) generates too few rounding errors to yield conclusive results, see Fig. 3.

1. *h-dependency*: In most cases, for large enough N, the error increased somewhat (in the order of a factor 2) upon halving the stepsize;

2. *N-dependency*: The actual rounding errors grow (except in the case of the unperturbed harmonic oscillator) predominantly quadratically;

Table 6
Test 2, $N = 10^7$, computation times in h:m:s

	Step size h	
Method/form	$\frac{\pi}{500}$	$\frac{\pi}{1000}$
A) *Difference notation*		
Störmer–Cowell	2:16:58	2:16:38
First sum	2:18:08	2:17:04
Second sum	2:30:07	2:17:50
(B) *Ordinate notation*		
Störmer–Cowell	2:42:54	2:42:59
Split form	2:53:28	2:53:30

3. *summing once vs. summing twice*: For small to moderate values of the step number, the first sum is somewhat faster and more accurate than the second sum. The differences diminish if $N \to \infty$.

In addition, we note that there is no difference between the difference and ordinate versions of the Störmer–Cowell method, as far as the rounding errors are concerned. The ordinate notation, however, is considerably slower, as is clear from Table 6. This table also shows that the split form is slower than the summed forms, of which the first sum is fastest.

Clearly, summed and split forms are superior over the original forms (this holds in general, see also [3, 11, 12]). Among the original forms, the difference form is faster than the ordinate version.

8. Conclusions

Multistep methods and their summed forms can easily be derived from one another by means of schemes of the type of Figs. 1 and 2, which represent the "Cowell" and "Adams" families, respectively. The Gauss–Jackson method, which is really the $\sum_c^2$-form of the "Cowell-family", is well known among astronomers. In this paper it is shown, however, that there is no reason to prefer this second-sum form over the first-sum form. This has been demonstrated with the aid of theoretical and numerical evidence. In general, summed and split forms are to be preferred over the original forms.

Acknowledgements

The author wishes to thank Prof. Dr. Cees Traas for his valuable comments during the preparation of this paper.

References

[1] K. Fox, Numerical integration of the equations of motion of celestial mechanics, *Celestial Mech.* **33** (1984) 127–142.

[2] J.F. Frankena, Störmer–Cowell and related methods for the numerical solution of second order periodic initial value problems for ODE's I: survey of methods, Memorandum no. 966, Dept. of Appl. Math., Twente University, Enschede, Netherlands, 1991.

[3] J.F. Frankena, Störmer–Cowell and related methods for the numerical solution of second order periodic initial value problems for ODE's II: numerical characteristics of the methods, Memorandum no. 1082, Dept. of Appl. Math., Twente University, Enschede, Netherlands, 1992.

[4] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Differential Equations, Vol. I: Nonstiff Problems* (Springer, Berlin, 1987).

[5] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations* (Wiley, New York, 1962).

[6] S. Herrick, *Astrodynamics I and II* (Van Nostrand Reinhold, London, 1971, 1972).

[7] F.B. Hildebrand, *Introduction to Numerical Analysis* (McGraw-Hill, New York, 1974).

[8] J.D. Lambert, *Numerical Methods for Ordinary Differential Systems* (Wiley, New York, 1991).

[9] R.H. Merson, Numerical integration of the differential equations of celestial mechanics, Tech. Report TR 74184, Royal Aircraft Establishment, 1975.

[10] A. Milani, and A.M. Nobili, Integration error over very long time spans, *Celestial Mech.* **43** (1988) 1–34.

[11] M.N. Spijker, Round-off error in the numerical solution of second order differential equations, in: Lecture Notes in Math. **109** (Springer, New York, 1969) 249–254.

[12] M.N. Spijker, Reduction of roundoff error by splitting of difference formulas, *SIAM J. Numer. Anal.* **8** (1971) 345–357.