SIAM J. Sci. Comput.
Vol. 18, No. 1, pp. 257–269, January 1997

REVERSIBLE LONG-TERM INTEGRATION WITH VARIABLE STEPSIZES[*]

ERNST HAIRER[†] AND DANIEL STOFFER[‡]

Abstract. The numerical integration of reversible dynamical systems is considered. A backward analysis for variable stepsize one-step methods is developed, and it is shown that the numerical solution of a symmetric one-step method, implemented with a reversible stepsize strategy, is formally equal to the exact solution of a perturbed differential equation, which again is reversible. This explains geometrical properties of the numerical flow, such as the nearby preservation of invariants. In a second part, the efficiency of symmetric implicit Runge–Kutta methods (linear error growth when applied to integrable systems) is compared with explicit nonsymmetric integrators (quadratic error growth).

Key words. symmetric Runge–Kutta methods, extrapolation methods, long-term integration, Hamiltonian problems, reversible systems

AMS subject classifications. 65L05, 34C35

PII. S1064827595285494

1. Introduction.

We consider the numerical treatment of systems

$$y' = f(y), \qquad y(0) = y_0, \tag{1.1}$$

where the evolution of dynamical features over long time intervals is of interest. In such a situation it is important to use a numerical method that is *numerically stable* in the sense of [17], i.e., the dynamical properties are inherited by the numerical approximation. Typical examples are that (1.1) is a Hamiltonian system and a symplectic integration method is applied with constant stepsizes or a reversible differential equation is integrated with a symmetric method (constant stepsizes). In these situations numerical stability can be explained by a backward error analysis, because the numerical solution can be formally interpreted as the exact solution of a perturbed differential equation which again is Hamiltonian, respectively, reversible. (This is related to a classical question called interpolation; see, e.g., H. Shniad [13]. In the context of numerical integrators we refer the reader to chapter 10 of [12] and the references given there.)

The aim of this article is to extend these results to variable stepsize integrations. At present no such extension is known for Hamiltonian problems. Therefore we shall concentrate on reversible systems and on the stepsize strategy of [15] (see also [2]). Section 2 presents a backward analysis for variable stepsize methods, where the step lengths are determined by a nonlinear equation involving a small parameter (the user-supplied tolerance). Particular attention is paid to symmetric methods. Practical aspects for an implementation of the reversible stepsize strategy are discussed in section 3. Finally, in section 4 we give a comparison between the symmetric Lobatto IIIA method (which preserves the geometric structure of a reversible problem and for which the global error grows linearly in time when applied to integrable systems) with

[*] Received by the editors May 3, 1995; accepted for publication (in revised form) September 9, 1995. This work was supported by project 2000-040665.94 of Swiss National Science Foundation.
http://www.siam.org/journals/sisc/18-1/28549.html

[†] Dept. de mathématiques, Université de Genève, CH-1211 Genève 24, Switzerland (hairer@divsun.unige.ch).

[‡] Dept. of Mathematics, ETH Zürich, CH-8092 Zürich, Switzerland (stoffer@math.ethz.ch).

the explicit GBS extrapolation method (applied with a more stringent tolerance, so that the same accuracy is achieved over a long time interval).

1.1. Reversible differential equations. The differential equation (1.1) is called *ρ-reversible* (ρ is an invertible linear transformation in the phase space) if

$$f(\rho y) = -\rho f(y) \qquad \text{for all } y. \tag{1.2}$$

This implies that the flow of the system, denoted by $\phi_t(y)$, satisfies $\rho\phi_t(y) = \phi_t^{-1}(\rho y)$, hence the name ρ- or time-reversible. A typical example is the partitioned system

$$p' = f_1(p,q), \qquad q' = f_2(p,q), \tag{1.3}$$

where $f_1(-p,q) = f_1(p,q)$ and $f_2(-p,q) = -f_2(p,q)$. Here the transformation ρ is given by $\rho(p,q) = (-p,q)$. Hamiltonian systems with a Hamiltonian satisfying $H(-p,q) = H(p,q)$ are of this type and all second-order differential equations $p' = g(q)$, $q' = p$.

1.2. Symmetric integration methods. For the numerical integration of (1.1) we consider Runge–Kutta methods

$$\begin{aligned} Y_i &= y_0 + h\sum_{j=1}^{s} a_{ij} f(Y_j), \qquad i = 1,\dots,s, \\ y_1 &= y_0 + h\sum_{i=1}^{s} b_i f(Y_i). \end{aligned} \tag{1.4}$$

Such a method defines implicitly a discrete-time map $y_0 \mapsto y_1$, denoted by $y_1 = \Phi_h(y_0)$. By the implicit function theorem it is defined for positive and negative (sufficiently small) values of h. Method (1.4) is called *symmetric* if

$$y_1 = \Phi_h(y_0) \qquad \Leftrightarrow \qquad y_0 = \Phi_{-h}(y_1). \tag{1.5}$$

If the coefficients of the Runge–Kutta method satisfy

$$a_{s+1-i,s+1-j} + a_{ij} = b_j \qquad \text{for all } i, j, \tag{1.6}$$

then the method is symmetric (see [5, p. 221]). For irreducible methods it can be shown (using the ideas of the proof of Lemma 4 of [4]) that after a suitable permutation of the stages, condition (1.6) is also necessary for symmetry.

The symmetry of a method is beneficial for the integration of ρ-reversible differential equations, in particular, if the main interest is in a qualitatively correct, but not necessarily accurate, simulation over a long time interval (see, e.g., [15]). A further explanation of this fact can be obtained by a backward analysis argument as follows: from [4] it is known that the numerical solution $y_0, y_1, y_2, \dots$ (obtained with a constant stepsize h) can be formally interpreted as the exact solution at times $0, h, 2h, \dots$ of the perturbed differential equation

$$\widetilde{y}\,' = f_h(\widetilde{y}\,), \qquad \widetilde{y}\,(0) = y_0, \tag{1.7}$$

where

$$f_h(y) = f(y) + \sum_{r(\tau)\ge p+1} \frac{h^{r(\tau)-1}}{r(\tau)!}\,\alpha(\tau)\,b(\tau)\,F(\tau)(y). \tag{1.8}$$

Here, p denotes the order of the method, the sum is over all rooted trees with at least $p+1$ vertices (we write $r(\tau)$ for the number of vertices of the tree τ), $\alpha(\tau)$ is an integer coefficient, $b(\tau)$ is a real coefficient depending on the integration method, and $F(\tau)(y)$ denotes the elementary differential corresponding to the tree τ. It follows from (1.2) and from the definition of the elementary differentials (see [4] or [5]) that

$$F(\tau)(\rho y) = (-1)^{r(\tau)} \rho F(\tau)(y). \tag{1.9}$$

Consequently, the function $f_h(y)$ of (1.8) is ρ-reversible if and only if $b(\tau)$ vanishes for all trees with an even number of vertices. But this is precisely the case for symmetric methods, because they are characterized by the fact that the global error has an asymptotic expansion in even powers of h. All properties of ρ-reversible systems are thus inherited by the numerical approximation, and Kolomogorov–Arnold–Moser (KAM) theory (see [8]) can be applied. If system (1.1) is integrable (see [9]), this implies the nearby conservation of invariants and the linear growth of the global error (in contrast to a quadratic growth if a nonreversible method were used; see [2]).

1.3. Reversible stepsize strategy. We consider an embedded method and denote the difference of the numerical solutions by

$$D(y_0, h) = h \sum_{i=1}^{s} e_i f(Y_i). \tag{1.10}$$

Using (1.4), the Taylor expansion around $h = 0$ shows that $D(y_0, h) = \mathcal{O}(h^q)$ with some $q \geq 1$. Usually, some heuristic formula based on this relation is used for stepsize prediction, and the step is accepted if $\|D(y_0, h)\| \leq \mathit{Tol}$. Such a stepsize strategy destroys the above-mentioned properties (conservation of invariants and linear error growth) as can be observed by numerical experiments.

To overcome this inconvenience, new stepsize strategies have been proposed in [7], [14], and [15]. Whereas the strategies of [7], [14] are problem dependent, the strategy of [15] is in the spirit of embedded methods. The idea is to use a *symmetric error estimate* satisfying

$$\|D(y_0, h)\| = \|D(y_1, -h)\| \qquad \text{with } \; y_1 = \Phi_h(y_0) \tag{1.11a}$$

and to require *equality* in

$$\|D(y_0, h)\| = \mathit{Tol}. \tag{1.11b}$$

Condition (1.11b) then determines uniquely (for small *Tol* and small h) the stepsize h as a function of the initial value y_0. If the coefficients e_i in (1.10) satisfy

$$e_{s+1-i} = e_i \quad \text{for all } i \qquad\qquad \text{or} \qquad\qquad e_{s+1-i} = -e_i \quad \text{for all } i, \tag{1.12}$$

and if the Runge–Kutta method is symmetric (i.e., condition (1.6) is satisfied), then condition (1.11a) is satisfied. This follows from the fact that the internal stage vectors Y_i of the step from y_0 to y_1 and the stage vectors $\overline{Y}_i$ of the step from y_1 to y_0 (negative stepsize $-h$) are related by $\overline{Y}_i = Y_{s+1-i}$. The stepsize determined by (1.11b) is thus the same for both steps.

As an illustration of the effect of the different stepsize strategies let us consider the modified Kepler problem (as presented in [12])

$$\begin{aligned} q_1' &= p_1, \qquad & p_1' &= -\frac{q_1}{(q_1^2+q_2^2)^{3/2}} - \frac{3\varepsilon q_1}{2(q_1^2+q_2^2)^{5/2}}, \\ q_2' &= p_2, \qquad & p_2' &= -\frac{q_2}{(q_1^2+q_2^2)^{3/2}} - \frac{3\varepsilon q_2}{2(q_1^2+q_2^2)^{5/2}} \end{aligned} \tag{1.13}$$

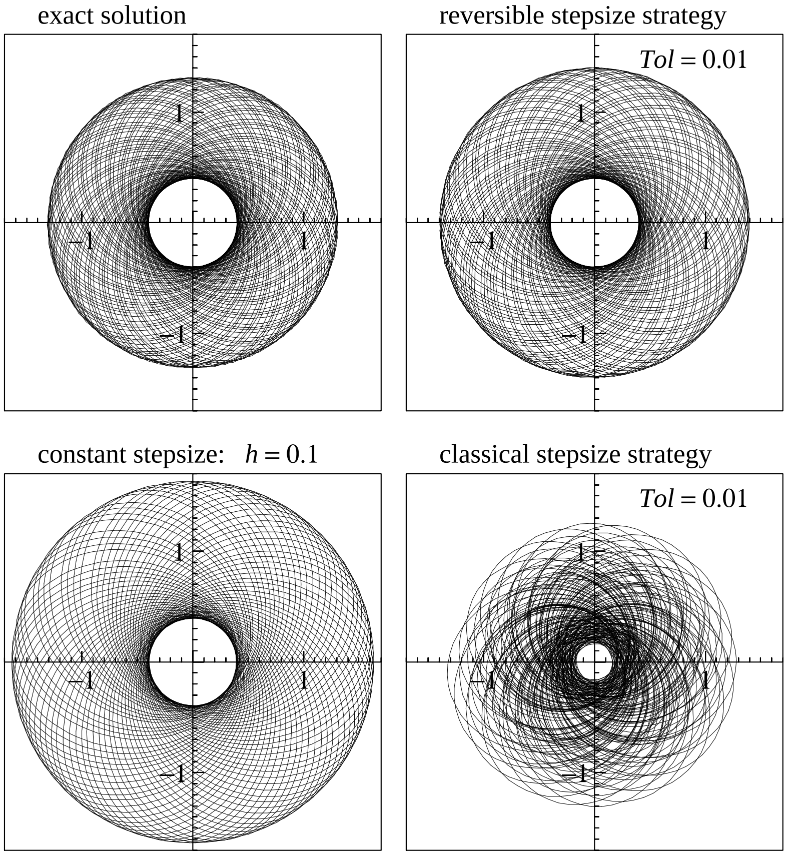

FIG. 1. *Comparison of different stepsize strategies.*

($\varepsilon = 0.01$) with initial values

$$q_1(0) = 1 - e, \quad q_2(0) = 0, \quad p_1(0) = 0, \quad p_2(0) = \sqrt{\frac{1+e}{1-e}} \tag{1.14}$$

(eccentricity $e = 0.6$). This problem has several symmetries and in particular satisfies condition (1.2) with $\rho(q_1, q_2, p_1, p_2) = (q_1, -q_2, -p_1, p_2)$. As numerical integrator we take the trapezoidal rule

$$y_1 = y_0 + \frac{h}{2}\left(f(y_0) + f(y_1)\right), \tag{1.15}$$

which is a symmetric method. For stepsize selection we consider the expression

$$D(y_0, h) = \frac{h}{2}\left(f(y_1) - f(y_0)\right). \tag{1.16}$$

It satisfies $\|D(y_0, h)\| = \|D(y_1, -h)\|$. In Fig. 1 we show the projection of the exact solution onto the (q_1, q_2)-plane for a time interval of length 500. We further present the numerical solution obtained with constant stepsize $h = 0.1$ and those obtained with variable stepsizes—the classical strategy (see, e.g., [5, p. 168]) and the reversible strategy—both applied with $Tol = 10^{-2}$. The value $h = 0.1$ for the constant stepsize implementation has been chosen so that the number of function evaluations is approximately the same as that for the reversible stepsize strategy. Whereas the numerical solution with classical stepsize strategy approaches the center and shows a wrong behavior, the results of the constant stepsize implementation and of the reversible stepsize implementation are qualitatively correct. This phenomenon will be explained in the next section.

2. Backward analysis of variable stepsize methods. In this section we shall extend the backward analysis of [4] to variable stepsize methods, where the step length is determined by a nonlinear equation involving a small parameter ε. We shall show that for symmetric methods the perturbed differential equation has an expansion in even powers of ε and that such methods are suitable for the integration of ρ-reversible systems.

We consider a *one-step method*, written in the general form

$$y_1 = \Phi_h(y_0), \tag{2.1a}$$

together with a *stepsize function*

$$h = \varepsilon\, s(y_0, \varepsilon). \tag{2.1b}$$

In [16] it was shown that a large class of variable stepsize methods are asymptotically equivalent to integration methods of the form (2.1).

The Runge–Kutta method (1.4) is clearly of the form (2.1a) and the reversible stepsize strategy, explained in the introduction, can be brought to the form (2.1b). Indeed, expanding $\|D(y,h)\|$ into powers of h and setting $Tol = \varepsilon^q$, condition (1.11b) becomes

$$\|D(y,h)\| = h^q d_q(y) + h^{q+1} d_{q+1}(y) + \cdots = \varepsilon^q. \tag{2.2}$$

Assuming that $d_q(y) > 0$ in the considered domain, we can take the qth root of (2.2) and solve the resulting equation for h. This yields a relation of the form (2.1b). The parameter ε is equal to $Tol^{1/q}$ and can be interpreted as a mean stepsize.

We denote by p the order of the one-step method (2.1a), and we expand its local error into a Taylor series as follows:

$$\Phi_h(y) - \phi_h(y) = h^{p+1} F_{p+1}(y) + h^{p+2} F_{p+2}(y) + \cdots. \tag{2.3}$$

Here, $F_j(y)$ is a linear combination of elementary differentials corresponding to trees with j vertices. We also assume that the stepsize function admits an expansion of the form

$$h = \varepsilon\, s(y,\varepsilon) = \varepsilon\, s_0(y) + \varepsilon^2\, s_1(y) + \varepsilon^3\, s_2(y) + \cdots. \tag{2.4}$$

THEOREM 1. *Let the one-step method* (2.1a) *be of order* p. *Then there exist functions* $f_j(y)$ (*for* $j \ge p$) *such that the numerical solution* $y_1 = \Phi_{\varepsilon s(y_0,\varepsilon)}(y_0)$ *is formally equal to the exact solution at time* $\varepsilon s(y_0,\varepsilon)$ *of the perturbed differential equation*

$$\widetilde{y}\,' = f(\widetilde{y}\,) + \varepsilon^p f_p(\widetilde{y}\,) + \varepsilon^{p+1} f_{p+1}(\widetilde{y}\,) + \ldots \tag{2.5}$$

with initial value $\widetilde{y}(0) = y_0$, *i.e.,* $y_1 = \widetilde{y}(\varepsilon s(y_0,\varepsilon))$.

Remark. The series in (2.5) does not converge in general. Nevertheless, in order to give a precise meaning to the statement of Theorem 1, we truncate the series after the ε^N-term and denote the solution of the truncated perturbed differential equation by $\widetilde{y}_N(t)$. In this way we get a family of solutions, which satisfy $y_1 = \widetilde{y}_N(\varepsilon s(y_0,\varepsilon)) + \mathcal{O}(\varepsilon^{N+1})$ (for a more detailed estimate of this error we refer to [1]). In order to emphasize the main ideas and to keep the notations as simple as possible, we omit in the following the subscript N and the remainder terms.

Proof. Following an idea of [10] we shall recursively construct the functions $f_j(y)$. We insert h from (2.4) into (2.3), and we see that the local error of the method is $\varepsilon^{p+1}(s_0(y))^{p+1}F_{p+1}(y) + \mathcal{O}(\varepsilon^{p+2})$. If we put

$$f_p(y) := (s_0(y))^p F_{p+1}(y), \tag{2.6}$$

the difference of the solution of the differential equation

$$\overline{y}\,' = f(\overline{y}) + \varepsilon^p f_p(\overline{y}), \qquad \overline{y}(0) = y_0$$

to that of (1.1) satisfies $\overline{y}(\varepsilon s(y_0,\varepsilon)) - y(\varepsilon s(y_0,\varepsilon)) = \varepsilon^{p+1}(s_0(y_0))^{p+1}F_{p+1}(y_0) + \mathcal{O}(\varepsilon^{p+2})$. Since the leading term of this difference is exactly the same as that of the local error, we get

$$\Phi_{\varepsilon s(y,\varepsilon)}(y) - \overline{y}(\varepsilon s(y,\varepsilon)) = \varepsilon^{p+2}G_{p+2}(y) + \cdots,$$

where the function $G_{p+2}(y)$ depends on $F_{p+2}(y)$, $F_{p+1}(y)$, $s_0(y)$, $s_1(y)$ and on the derivative of $f_p(y)$. In the second step of the proof we put $f_{p+1}(y) := G_{p+2}(y)$, consider the differential equation

$$\overline{\overline{y}}\,' = f(\overline{\overline{y}}) + \varepsilon^p f_p(\overline{\overline{y}}) + \varepsilon^{p+1} f_{p+1}(\overline{\overline{y}}), \qquad \overline{\overline{y}}(0) = y_0,$$

and conclude as above that the difference between $\overline{\overline{y}}(\varepsilon s(y_0,\varepsilon))$ and the numerical solution is of size $\mathcal{O}(\varepsilon^{p+3})$. Continuing this procedure leads to the desired expansion (2.5). □

Remark. If we apply the statement of Theorem 1 to successive integration steps, we obtain formally $y_n = \widetilde{y}(t_n)$ where t_n is recursively defined by $t_n = t_{n-1} + \varepsilon s(y_{n-1},\varepsilon)$ and $t_0 = 0$. It then follows from the nonlinear variation of constants formula (see Theorem I.14.5 of [5]) that

$$y_n - y(t_n) = \varepsilon^p e_p(t_n) + \varepsilon^{p+1} e_{p+1}(t_{n+1}) + \cdots, \tag{2.7}$$

giving an asymptotic expansion in powers of ε for the global error of the variable stepsize method (2.1). Since the leading term contains the factor $\varepsilon^p = Tol^{p/q}$, we get "tolerance proportionality" if we replace condition (1.11) by

$$\|D(y_0,h)\| = Tol^{q/p}. \tag{2.8}$$

This modification is implemented in all of our computations.

2.1. Adjoint and symmetric methods. Our next aim is to prove that for symmetric methods the expansion (2.5) contains only even powers of ε. For this purpose we extend the results of [5, Sect. II.8] to variable stepsize methods.

Let us start with the definition of the adjoint method of (2.1). For this we replace in (2.1) ε by $-\varepsilon$, denote $h^* = -h$, and exchange the notations of y_0 and y_1. This yields

$$y_0 = \Phi_{-h^*}(y_1), \qquad h^* = \varepsilon\, s(y_1, -\varepsilon), \tag{2.9}$$

a system which by the implicit function theorem defines implicitly y_1 and h^* as functions of y_0 and ε. We denote these functions by

$$y_1 = \Phi^*_{h^*}(y_0), \qquad h^* = \varepsilon\, s^*(y_0, \varepsilon) \tag{2.10}$$

and call the resulting formulas the *adjoint method* of (2.1).

DEFINITION 2. *The variable stepsize method* (2.1) *is called symmetric if it is equal to its adjoint method, i.e., if* $\Phi^* = \Phi$ *and* $s^* = s$.

The condition $\Phi^* = \Phi$ is the same as for the symmetry of the fixed stepsize method (2.1a). If the stepsize strategy (1.11) is used in connection with a method satisfying $\Phi^* = \Phi$, then $s(y_1, -\varepsilon) = s(y_0, \varepsilon)$ holds and hence also $s^* = s$. Consequently, a Runge–Kutta method satisfying (1.6) together with a stepsize strategy (1.11) satisfying (1.12) is symmetric in the sense of Definition 2.

THEOREM 3. *If the variable stepsize method* (2.1) *is symmetric, then the perturbed differential equation* (2.5) *has an expansion in even powers of* ε, *i.e.,* $f_j(y) = 0$ *for odd* j.

Proof. We first look for the perturbed differential equation of the adjoint method. Recall that the adjoint method is obtained from (2.1) by replacing ε by $-\varepsilon$ and by exchanging y_0 and y_1. If we apply these operations to the statement of Theorem 1, we obtain

$$\widetilde{y}\,' = f(\widetilde{y}\,) + (-\varepsilon)^p f_p(\widetilde{y}\,) + (-\varepsilon)^{p+1} f_{p+1}(\widetilde{y}\,) + \cdots \tag{2.11}$$

with $\widetilde{y}(0) = y_1$ and $\widetilde{y}(-\varepsilon s(y_1, -\varepsilon)) = y_0$ or, after a time translation, $\widetilde{y}(\varepsilon s(y_1, -\varepsilon)) = y_1$ and $\widetilde{y}(0) = y_0$. By definition of s^* we thus obtain $y_1 = \widetilde{y}(\varepsilon s^*(y_0, \varepsilon))$, where $\widetilde{y}(t)$ is the solution of (2.11) with initial value $\widetilde{y}(0) = y_0$. The perturbed differential equation of the adjoint method is therefore equation (2.11).

For symmetric methods the expansions of (2.5) and (2.11) have to be equal. But this is only the case if the terms with odd powers of ε vanish. ☐

2.2. Application to reversible systems. Now let the differential equation (1.1) satisfy (1.2) and apply the Runge–Kutta method (1.4) with stepsize strategy (2.1b). At the moment neither the method nor the stepsize function are assumed to be symmetric. We are interested in the structure of the perturbed differential equation (2.5).

Multiplying (1.4) by ρ, it follows from (1.2) that

$$\rho Y_i = \rho y_0 - h \sum_{j=1}^{s} a_{ij} f(\rho Y_j), \qquad \rho y_1 = \rho y_0 - h \sum_{i=1}^{s} b_i f(\rho Y_i).$$

This shows that the numerical solution of method (1.4), applied with stepsize $-h$ to $\overline{y}_0 = \rho y_0$, yields $\overline{y}_1 = \rho y_1$ and the stage vectors are $\overline{Y}_i = \rho Y_i$. Consequently, we have $D(\rho y_0, -h) = \rho D(y_0, h)$ and under the additional assumption that ρ is an orthogonal transformation (with respect to the inner product norm used in (1.11)) we also have

$$s(\rho y_0, -\varepsilon) = s(y_0, \varepsilon). \tag{2.12}$$

THEOREM 4. *Suppose that the differential equation* (1.1) *is* ρ*-reversible with an orthogonal transformation* ρ. *For a Runge–Kutta method* (1.4) *with a stepsize function satisfying* (2.12), *the coefficient functions of the perturbed differential equation* (2.5) *satisfy*

$$f_j(\rho y) = -(-1)^j \rho\, f_j(y).$$

Proof. Let y_1 be given by (1.4) with $h = \varepsilon s(y_0, \varepsilon)$. The discussion before the formulation of Theorem 4 shows that ρy_1 is then the numerical solution of (1.4) obtained with initial value ρy_0 and negative stepsize $-h = -\varepsilon s(\rho y_0, -\varepsilon)$. As a consequence of Theorem 1 we have $\rho y_1 = z(-\varepsilon s(\rho y_0, -\varepsilon))$, where $z(t)$ is the solution of

$$z' = f(z) + (-\varepsilon)^p f_p(z) + (-\varepsilon)^{p+1} f_{p+1}(z) + \cdots$$

with initial value $z(0) = \rho y_0$. Introducing the new variable $v(t) := \rho^{-1} z(-t)$ we obtain

$$v' = -\rho^{-1} f(\rho v) - (-\varepsilon)^p \rho^{-1} f_p(\rho v) - (-\varepsilon)^{p+1} \rho^{-1} f_{p+1}(\rho v) + \cdots \tag{2.13}$$

with $v(0) = y_0$ and $v(\varepsilon s(y_0, \varepsilon)) = y_1$ (here we have again used the relation (2.12)). Since the differential equation (2.13) is the same for all values of y_0, we obtain the statement of the theorem by comparing the series of (2.5) and (2.13). □

Combining the statements of Theorems 3 and 4 we get the following interesting result.

COROLLARY 5. *If in addition to the assumptions of Theorem* 4 *the method is symmetric (in the sense of Definition* 2)*, then the perturbed differential equation* (2.5) *is also ρ-reversible.*

We are now back to the same situation as for symmetric fixed stepsize methods: the numerical solution is formally equal to the exact solution of a perturbed differential equation, which is also ρ-reversible. Hence, the same conclusions concerning the preservation of invariants and the linear error growth can be drawn (see section 3 of [2] for a formal explanation). In order to get rigorous estimates also in the case where the series in (2.5) does not converge, one has to estimate the error $y_1 - \widetilde{y}_N(\varepsilon s(y_0, \varepsilon))$ (see the remark following Theorem 1) and to study its propagation and accumulation. Under suitable assumptions this allows us to conclude that the formally obtained results are valid on "exponentially long" time intervals (see [1] for more details).

3. On the implementation of reversible stepsize strategies.

In the preceding analysis we have assumed that the nonlinear equations (1.4) and (2.8) are solved exactly. In practice, however, these equations have to be solved iteratively, and it is important to have suitable stopping criteria. For implicit Runge–Kutta methods it is natural to solve equations (1.4) and (2.8) simultaneously. An iterative method based on a dense output formula is proposed in [2]. For certain problems, simplified Newton iterations for system (1.4) and convergence acceleration strategies for the stepsizes may be more suitable. It is not the purpose of the present paper to discuss such details (which are still under investigation), but we shall concentrate on aspects that are common to all iteration techniques.

We have problems in mind where KAM theory is applicable. This implies that the global error of symmetric methods grows linearly in time. Obviously, the errors due to the approximate solution of the nonlinear equations should not be larger than the error due to discretization.

3.1. Stopping criterion for the iteration of the stage vectors.

The error in the stage vectors that is due to the iterative solution of the system (1.4) is not correlated to the discretization error. This means that its dominant part has also components orthogonal to the direction of the flow of the problem. Hence the contribution of this error to the numerical solution grows quadratically in time.

Suppose now that the considered problem is well scaled (meaning that a characteristic time interval such as the period or quasi period is equal to one) and that an approximation of the solution is searched over an interval of length T with an accuracy δ. We are interested in the situation where T is much larger than one. Due to the linear error growth of the method we have to apply the stepsize strategy (2.8) with

$$Tol = \delta/T. \tag{3.1}$$

If we denote the stopping error of the iterative process by $\varepsilon_{\text{iter}}$, then its contribution to the error at time t will be of size $\mathcal{O}(t^2\varepsilon_{\text{iter}})$. To guarantee that it is below δ on the considered interval, we require $T^2\varepsilon_{\text{iter}} \le \delta$. This justifies the stopping criterion

$$\|\Delta Y_i\| \le Tol^2/\delta \qquad \text{for } i = 1, \dots, s. \tag{3.2}$$

Since $Tol \le \delta$, the nonlinear equations (1.4) have to be solved with a higher precision than when the method is applied to unstructured problems.

For stringent tolerances the influence of round-off errors may be significant. Since they are random and the propagation of perturbations is linear for our model problem, the round-off contribution to the error at time t is of size $\mathcal{O}(t^{3/2}eps)$, where eps is the unit roundoff. For $\varepsilon_{\text{iter}} \ge eps$ the round-off errors can be neglected, because they grow slower than the iteration errors. What happens if $\varepsilon_{\text{iter}} < eps$? In this situation it is not adequate to use the stopping criterion $\|\Delta Y_i\| \le \varepsilon_{\text{iter}}$, but we can proceed as follows: as long as $\|\Delta Y_i\| > 10eps$ (here 10 is an arbitrary safety factor) we use the increments to estimate the convergence rate κ. If $\|\Delta Y_i\| \le 10eps$ for the first time, we continue the iteration until the theoretical iteration error (assuming that from now on one gains a factor of κ at each iteration) is smaller than $\varepsilon_{\text{iter}}$. Numerical experiments revealed that such a procedure allows the iteration error to be pushed below the round-off error.

3.2. Stopping criterion for the iteration of the stepsize. The situation changes completely for the errors in the stepsize. Let $\ell(y_0, h) = \Phi_h(y_0) - \phi_h(y_0)$ denote the local truncation error of the method. Then, perturbing the stepsize h to $h + \Delta h$, the numerical solution of (1.4) becomes $y_1 + \Delta y_1$ where $\Delta y_1 = y(t_0 + h + \Delta h) - y(t_0 + h) + \ell(y_0, h + \Delta h) - \ell(y_0, h)$. For a method of order p, the difference $\ell(y_0, h + \Delta h) - \ell(y_0, h)$ is of size $\mathcal{O}(h^p\Delta h)$, which by (2.8) can be expected to be of size $\mathcal{O}(Tol \cdot \Delta h)$. Hence the dominant part of this error is in direction of the flow of the differential equation and induces a time shift of the solution. The sum of these errors results in a linear error growth in time. If we require that in each step this error is bounded by Tol, i.e.,

$$|\Delta h| \cdot \|f(y_0)\| \le Tol, \tag{3.3}$$

then its contribution to the final error is comparable with that of the discretization error. The component of the error, orthogonal to the flow, can grow quadratically and usually leads to a $\mathcal{O}(T^2 \cdot Tol \cdot \Delta h) = \mathcal{O}(T^2 \cdot Tol^2)$ contribution to the error at time T. If the constant symbolized by $\mathcal{O}(T^2 \cdot Tol^2)$ is not larger than δ^{-1} (see (3.1)), then these components can be neglected. We therefore propose (3.3) as a stopping criterion for the iterative solution of (2.8).

3.3. A reversible lattice stepsize strategy. As pointed out in [2] the computation of the stepsize h from (2.8) is ill conditioned. Indeed, due to round-off errors in the computation of $f(Y_i)$, the expression $D(y_0, h)$ is affected by a relative error of size $eps \cdot h \cdot \|f(y_0)\|/\|D\|$. Since $\|D\| \approx Ch^q$, this leads to a relative error in h of size

$$\frac{\Delta h}{h} \approx \frac{1}{q}\,\frac{\Delta\|D\|}{\|D\|} \approx eps\,\frac{h \cdot \|f(y_0)\|}{q \cdot Tol^{q/p}}, \tag{3.4}$$

because $\|D\| \approx Tol^{q/p}$. Difficulties will therefore arise when the right-hand expression of (3.4) is larger than $Tol/(h \cdot \|f(y_0)\|)$.

As a remedy to this fact (inspired by the symplectic lattice methods of [3]) we propose to restrict the stepsizes to the values of a fixed lattice L; e.g., we consider only stepsizes that are integer multiples of say 2^{-m} where m is a fixed number. Obviously, condition (2.8) can no longer be satisfied exactly. The essential idea is now to take as stepsize h the largest element of L satisfying $\|D(y_0, h)\| \le Tol^{q/p}$. In this way the stepsize is (locally) uniquely determined and can again be written as $h = \varepsilon s(y_0, \varepsilon)$. Under assumptions (1.6) and (1.12) we still have $s(y_1, -\varepsilon) = s(y_0, \varepsilon)$, so that this variable stepsize method is symmetric in the sense of Definition 2. However, $s(y, \varepsilon)$ no longer has an expansion (2.4) and the results of section 2 cannot be applied. Nevertheless, experiments have shown that the numerical solution of this approach still shares the nice properties (nearby preservation of invariants, linear error growth) of the continuous stepsize strategy.

4. Comparison between reversible and nonreversible methods. It is well known that for the numerical integration of nonstiff differential equations, explicit methods are superior to implicit methods. Unfortunately, symmetric Runge–Kutta methods cannot be explicit. Hence the question arises whether the symmetry of a method can compensate for its nonexplicitness when it is applied to a reversible differential equation over a very long time interval. In order to achieve the accuracy δ on an interval of length T, a symmetric method can be applied with $Tol = \delta/T$ whereas a general nonsymmetric method has to be applied with the more stringent tolerance $\delta/T^2 = Tol^2/\delta$.

This section is devoted to a comparison of two classes of integration methods: the symmetric collocation method based on the Lobatto quadrature formula, named Lobatto IIIA, and the explicit Gragg–Bulirsch–Stoer (GBS) extrapolation method (for their precise definition, see, for example, [6]). Both classes contain methods of arbitrarily high order. As a measure of comparison we consider the *work per unit step* $W = A/h$, where A counts the number of function evaluations of one step and h is the stepsize such that the local error of the method is Tol and Tol^2/δ, respectively. We restrict our theoretical comparison to linear problems $y' = Qy$, because in this case the error constants are available and a reasonable comparison is possible. Some limited numerical experiments with the Kepler problem have given similar results.

4.1. Work per unit step for Lobatto IIIA. The collocation method based on the s-stage Lobatto quadrature formula is a Runge–Kutta method of order $p = 2s - 2$ (see [6, p. 80]). For a linear problem the numerical solution is $y_1 = R_{s-1,s-1}(hQ)y_0$, where $R_{s-1,s-1}(z)$ is the diagonal Padé approximation of degree $s - 1$ (see [6, p. 50]). The local error is given by

$$err = \frac{k!\,k!}{(2k)!(2k+1)!}\, h^{2k+1} Q^{2k+1} y_0 + \mathcal{O}(h^{2k+2}), \tag{4.1}$$

where, in view of the comparison with the GBS method, we have put $k = s - 1$. The hypothesis $\|err\| \approx Tol$ allows us to express the stepsize h as a function of Tol. Neglecting higher-order terms and assuming that $\|Q^{2k+1}y_0\| \approx 1$ for all k we obtain for the stepsize

$$h = \left(\frac{(2k)!(2k+1)!}{k!\,k!} \cdot Tol \right)^{1/(2k+1)}. \tag{4.2}$$

For the computation of the number of function evaluations we have to specify an iteration process for the solution of nonlinear system (1.4). Suppose that we apply fixed point iteration with starting approximation $Y_i^{(0)} = y_0 + hc_i f(y_0)$ where $c_i = \sum_j a_{ij}$ (in practice one would use the extrapolated dense output solution; some iterations can be saved in this way). For linear problems $y' = Qy$ the error E_0 of the vector $(Y_1^0, \ldots, Y_s^{(0)})^T$ is equal to $E_0 = h^2 A^2 \mathbb{1} \otimes Q^2 y_0 + \mathcal{O}(h^3)$. Here $A = (a_{ij})$ is the $s \times s$ matrix whose elements are the Runge–Kutta coefficients of (1.4) and $\mathbb{1}$ stands for the vector $(1, \ldots, 1)^T$. Each fixed point iteration contributes a factor $hA \otimes Q$, so that after r iterations the error is equal to

$$E_r = h^{r+2} A^{r+2} \mathbb{1} \otimes Q^{r+2} y_0 + \mathcal{O}(h^{r+3}).$$

We shall stop the iteration as soon as $\|E_r\| \approx \mathit{Tol}^2/\delta$ in a suitable norm (see (3.2)). In order to estimate $\|A^{r+2}\mathbb{1}\|$ we shall use the spectral radius of A. Since the eigenvalues of A are the inverse values of the poles of the stability function, we have to bound from below the zeros of the denominator of $R_{s-1,s-1}(z)$. A result of [11] states that these zeros lie outside the parabolic region $\{z = x + iy \,|\, y^2 \le 4s(x+s)\}$, which contains the disc with center $(0,0)$ and radius s. Consequently, the spectral radius of A satisfies $\rho(A) \le 1/s$. The number r of required iterations can therefore be approximated by the relation

$$\left(\frac{h}{k+1}\right)^{r+2} = \frac{\mathit{Tol}^2}{\delta}. \tag{4.3}$$

The work per unit step for the Lobatto IIIA method of order $2k$ is therefore given by

$$W_k = \frac{1 + kr_k}{h_k}, \tag{4.4}$$

where h_k is the value given by (4.2) and r_k is the value given by (4.3).

4.2. Work per unit step for GBS extrapolation. The GBS extrapolation algorithm is an explicit method for the integration of (1.1). It is based on the explicit midpoint rule, whose error has an asymptotic expansion in even powers of the stepsize. If we apply $k-1$ extrapolations with the most economic stepsize sequence (the "harmonic sequence" of Deuflhard) to the linear problem $y' = Qy$, the numerical approximation is of the form $y_1 = P_{2k}(hQ)y_0$, where $P_{2k}(z)$ is a polynomial of degree $2k$. Since the order of this approximation is $2k$, it has to be the truncated series for $\exp(z)$. Consequently, the local error of the GBS method is given by

$$\mathit{err} = \frac{1}{(2k+1)!}\, h^{2k+1} Q^{2k+1} y_0 + \mathcal{O}(h^{2k+2}). \tag{4.5}$$

We again neglect higher-order terms and assume that $\|Q^{2k+1}y_0\| \approx 1$. The condition $\|\mathit{err}\| \approx \mathit{Tol}^2/\delta$, which, due to the quadratic error growth, has to be imposed in order to achieve a precision of δ over the considered interval, yields

$$h = \left((2k+1)! \cdot \frac{\mathit{Tol}^2}{\delta}\right)^{1/(2k+1)}. \tag{4.6}$$

The number of required function evaluations is $k^2 + 1$, so that the work per unit step of the GBS method is

$$W_k = \frac{k^2 + 1}{h_k}, \tag{4.7}$$

where the stepsize h_k is given by (4.6).

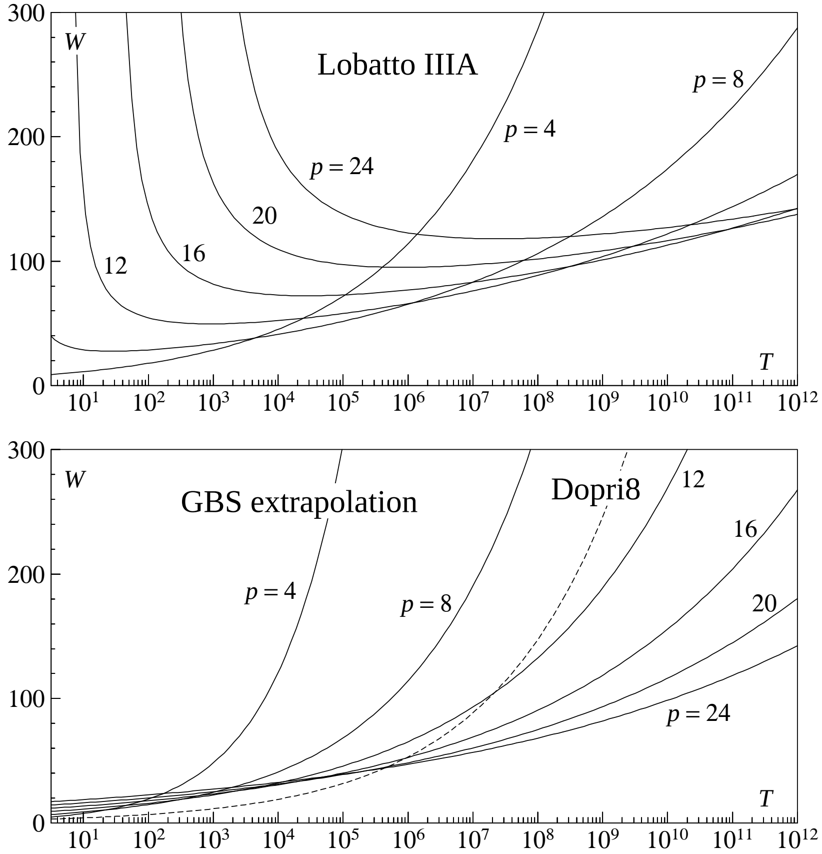

FIG. 2. *Work per unit step as a function of the length of the interval.*

4.3. Comparison between Lobatto IIIA and GBS. Figure 2 shows the values W_k (for methods of order $p = 2k$) plotted as functions of T, the length of the considered interval. We have chosen $\delta = 10^{-1}$ and *Tol* from (3.1) so that the numerical solution is accurate to one digit on the interval $[0, T]$.

In the upper picture (Lobatto IIIA) the plotted curves show a singularity for small values of T. This can be explained as follows: if a high-order method is used with a relatively large *Tol* (which is the case for small values of T), then the stepsize determined by (4.2) is rather large. If it is so large that $h \geq k+1$ (see (4.3)), then the fixed point iteration does not converge and the work per unit step tends to infinity. We also observe that for long integration intervals it is important to use high-order methods, but an order higher than $p = 20$ does not improve the performance.

The lower picture (GBS method) shows a completely different qualitative behavior. Since the method is explicit, there is no singularity for small values of T. Again, high order is essential for an efficient integration over a long time interval. Here we observe that high order improves the performance of the method on large intervals but still remains efficient also for small T (coarse tolerances). We have also included the corresponding curve for the explicit Runge–Kutta method of order 8 due to Dormand and Prince (see [5, p. 181ff]). Due to its smaller error constant it lies below the curve of the eighth-order GBS method, but its order is too low for very large time intervals.

Comparing the two pictures we see that (at least for large T) the curves for order p in the Lobatto IIIA case are close to those for order $2p$ in the GBS case. This indicates that the application of an explicit (nonsymmetric) method needs an order twice as high as that of a symmetric method in order to achieve the same efficiency. It should be mentioned that the pictures of Fig. 2 do not take into account the influence of round-off errors. The numerical solution of the Lobatto IIIA method is obtained by an iterative process, so that only the last iteration contributes to the round-off error of one step. Since the weights of the Lobatto quadrature formulas are all positive, this error will be close to the unit round-off *eps*. The situation changes for the GBS method. Since extrapolation is a numerically unstable process, the round-off error of one step increases with increasing order. For example, about two digits are lost for the method of order 18 (see [5, p. 242]).

REFERENCES

[1] G. Benettin and A. Giorgilli, *On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms*, J. Statist. Phys., 74 (1994), pp. 1117–1143.

[2] M. P. Calvo and E. Hairer, *Accurate long-term integration of dynamical systems*, Appl. Numer. Math., 18 (1995), pp. 95–105.

[3] D. J. D. Earn, *Symplectic integration without roundoff error*, in Ergodic Concepts in Stellar Dynamics, V.G. Gurzadyan and D. Pfenniger, eds., Lecture Notes in Physics 430, Springer-Verlag, Berlin, New York, 1994, pp. 122–130.

[4] E. Hairer, *Backward analysis of numerical integrators and symplectic methods*, Ann. Numer. Math., 1 (1994), pp. 107–132.

[5] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I. *Nonstiff Problems. Second Revised Edition*, Springer Series in Computational Mathematics 8, Springer-Verlag, Berlin, New York, 1993.

[6] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II. *Stiff and Differential-Algebraic Problems*, Springer Series in Computational Mathematics 14, Springer-Verlag, Berlin, New York, 1991.

[7] P. Hut, J. Makino, and S. McMillan, *Building a Better Leapfrog*, Astrophysical Journal, 443 (1995), pp. L93–L96.

[8] J. Moser, *Stable and Random Motions in Dynamical Systems*, Annnals of Math. Studies 77, Princeton University Press, Princeton, NJ, 1973.

[9] A. M. Perelomov, *Integrable Systems of Classical Mechanics and Lie Algebras*, A.G. Reyman, translator, Birkhäuser-Verlag, Basel, 1990.

[10] S. Reich, *Numerical Integration of the Generalized Euler Equations*, Tech. Report 93-20, Univ. of British Columbia, Vancouver, BC, 1993.

[11] E. B. Saff and R. S. Varga, *On the sharpness of theorems concerning zero-free regions for certain sequences of polynomials*, Numer. Math., 26 (1976), pp. 345–354.

[12] J. M. Sanz-Serna and M. P. Calvo, *Numerical Hamiltonian Problems*, Applied Mathematics and Mathematical Computation 7, Chapman and Hall, London, 1994.

[13] H. Shniad, *The equivalence of von Zeipel mappings and Lie transforms*, Cel. Mech., 2 (1970), pp. 114–120.

[14] D. Stoffer, *On Reversible and Canonical Integration Methods*, SAM-Report 88-05, ETH-Zürich, 1988.

[15] D. Stoffer, *Variable steps for reversible integration methods*, Computing, 55 (1995), pp. 1–22.

[16] D. Stoffer and K. Nipp, *Invariant curves for variable stepsize integrators*, BIT, 31 (1991), pp. 169–180.

[17] A. M. Stuart and A. R. Humphries, *Model problems in numerical stability theory for initial value problems*, SIAM Rev., 36 (1994), pp. 226–257.