IMA Journal of Numerical Analysis, **19**, (1999) 169–190.

Qualitative Properties of Modified Equations

O. GONZALEZ, [1] D.J. HIGHAM [2] AND A.M. STUART [3]

Suppose that a consistent one-step numerical method of order r is applied to a smooth system of ordinary differential equations. Given any integer $m \geq 1$, the method may be shown to be of order $r+m$ as an approximation to a certain modified equation. If the method and the system have a particular qualitative property then it is important to determine whether the modified equations inherit this property. In this article, a technique is introduced for proving that the modified equations inherit qualitative properties from the method and the underlying system. The technique uses a straightforward contradiction argument applicable to arbitrary one-step methods and does not rely on the detailed structure of associated power series expansions. Hence the conclusions apply, but are not restricted, to the case of Runge–Kutta methods. The new approach unifies and extends results of this type that have been derived by other means: results are presented for integral preservation, reversibility, inheritance of fixed points, Hamiltonian problems and volume preservation. The technique also applies when the system has an integral that the method preserves not exactly, but to order greater than r. Finally, a negative result is obtained by considering a gradient system and gradient numerical method possessing a global property that is not shared by the associated modified equations.

[1] Department of Mathematics, Swiss Federal Institute of Technology, Lausanne, Switzerland.
Work supported by a National Science Foundation Graduate Fellowship.

[2] Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, Scotland.
Work supported by the Engineering and Physical Sciences Research Council of the UK under grant GR/K80228.

[3] Program in Scientific Computing and Computational Mathematics & Division of Mechanics and Computation, Durand Building 257, Department of Mechanical Engineering, Stanford University, Stanford, CA 94305-4040, USA.
Work supported by the National Science Foundation under grant DMS-9201727 and by the Office of Naval Research under grant N00014-92-J-1876.

1. Introduction.

In this article we consider the relationship between solutions to a given system of ordinary differential equations, numerical approximations to them, and solutions to associated modified equations. Our goal is to show that if an underlying system and an approximation scheme possess solutions sharing certain qualitative properties, then there is a family of associated modified equations possessing solutions that also share these properties. Although many results of this type may already be found in the literature, we provide a general framework and a unified method of proof. Our method of proof, which is analytical, is not as succinct as other more algebraic approaches, but it is readily accessible to readers with a basic knowledge of numerical methods for initial value problems.

The presentation is structured as follows. In Section 2 we establish the notation for the semigroups generated by the underlying differential equation and the numerical method. We work within a general class of one-step methods that satisfy a certain local approximation property: the expansions of the true and approximate semigroups in powers of the time-step Δt agree up to order r. (Since we are dealing with ordinary differential equations the local semigroup property may be extended to a local group property, but in many interesting applications global existence is only assured in forward time so that we retain the concept of *semi*group.) Standard error estimates for such methods show that the error over a finite time interval is of $\mathcal{O}(\Delta t^r)$. Runge-Kutta methods are included in our framework, together with a variety of non-standard one-step methods used in practice, such as those that ensure conservation of invariants for Hamiltonian systems or volume for systems defined by divergence-free vector fields.

In Section 3 we discuss modified equations. Here, given an integer $m \geq 1$, the idea is to find an $\mathcal{O}(\Delta t^r)$ modification of the original ordinary differential equation with the property that the numerical method is $\mathcal{O}(\Delta t^{r+m})$ accurate as an approximation of this *modified equation*. We prove a general result concerning the existence and approximation properties of modified equations for the general class of one-step methods introduced in Section 2. This straightforward result is well-known in the numerical analysis literature, and also has related counterparts in the mathematical physics literature where interpolation of near-identity maps is important. Our proof simply sets the notation and methodology used in the remainder of the article. Modified equations were first studied in detail in the numerical analysis community by Warming & Hyett [1974] within the context of partial differential equations. In this area the modified equation approach is often useful in interpreting the qualitative properties of errors introduced by numerical approximation, such as numerical dissipation or dispersion for wave propagation problems. For further results on the usefulness and applicability of the modified equation approach, especially in the context of ordinary differential equations, see Griffiths & Sanz-Serna [1986], Beyn [1991], Reich [1993], [1996] Calvo, Murua & Sanz-Serna [1994], Hairer [1994], Reddien [1995], Fiedler & Scheurle [1996] and Sanz-Serna & Murua [1997]. A major application has been the derivation of exponentially small error estimates, starting with the work of Neishtadt [1984] and continued in, for example, Benettin & Giorgilli [1994], Hairer & Lubich [1997], and Reich [1996].

The main contribution of this article is contained in Section 4 where the qualitative properties of modified equations are studied. By use of a straightforward contradiction argument we show that if the numerical method shares a certain structural property with the underlying system, then the family of associated modified equations inherits this property. Such results motivate the use of numerical methods that respect qualitative features of the ordinary differential equation. The specific structural properties that we consider are preservation of a scalar function, reversibility, inheritance of fixed points, conservation of the canonical symplectic two-form for Hamiltonian systems and conservation of volume. A negative example is also given: we show that the global limit set behaviour of gradient systems is not necessarily shared by modified equations for gradient numerical methods.

The fact that structure-preserving numerical methods may possess modified equations with analogous structure has been known for some time. For early references in the numerical analysis literature see Mackay [1992], Sanz-Serna [1992] and Sanz-Serna & Calvo [1994] and for some early applications see Auerbach & Friedman [1991] and Yoshida [1993]. More recently, modified equations for problems with special structure have been explored in greater detail using several different approaches, as in Hairer [1994], Calvo, Murua & Sanz-Serna [1994], Benettin & Giorgilli [1994], Reich [1993], [1996], Hairer & Stoffer [1997] and Sanz-Serna & Murua [1997].

The work of Reich, in particular, describes a very general algebraic approach to the study of structure in modified equations which we now outline for the purposes of comparison with our analytic approach. For differential equations whose vector fields lie in a certain linear subspace g of the infinite dimensional Lie algebra of smooth vector fields, the semigroup generated lies in a corresponding subset G of the Fréchet manifold of smooth diffeomorphisms. Assuming G is a submanifold and that the tangent space to G at the identity is g, Reich shows that general one-step methods with semigroups in G possess modified equations with vector fields in g. In this sense, the modified equations possess the same structure as the underlying differential equation. The primary goal of this article is to demonstrate that a straightforward contradiction argument may be applied to a general one-step method to obtain similar results, without directly relying upon the geometrical relationship between G and g. Our approach is less succinct than that of Reich, but fits more naturally into a traditional numerical analysis framework. The approach also allows us to prove some new results: Theorems 4.2 and 4.4. Theorem 4.2 concerns methods that "almost" preserve an integral of the system; that is, they preserve the integral to a higher order than the classic order suggests. Examples of such methods have been proposed by Calvo, Iserles & Zanna [1996]. Theorem 4.4 concerns the preservation of fixed points.

Most of the work mentioned in the references above, and also the analysis presented here, applies to fixed-stepsize implementations. Reich [1996], Hairer & Stoffer [1997] and Hairer [1997] have recently explored the idea of developing customized variable-stepsize strategies for which an appropriate modified equation theory exists.

2. Background.

Consider a system of ordinary differential equations in $\mathbb{R}^p$ of the form

$$\frac{du}{dt} = f(u) \tag{2.1}$$

where the vector field $f : \mathbb{R}^p \to \mathbb{R}^p$ is assumed to be of class C^∞. For any $u_0 \in \mathbb{R}^p$ we denote by $S : B \times [0, T] \to \mathbb{R}^p$ the local evolution semigroup generated by (2.1) where B is a closed ball at u_0 and $T > 0$. In particular, for any $U \in B$ the curve

$$u(t) = S(U, t) = S_t(U) \tag{2.2}$$

is a solution to (2.1) with initial condition $u(0) = U$, defined for all $t \in [0, T]$. Furthermore, for each $t \in [0, T]$ the mapping $S_t : B \to \mathbb{R}^p$ is a C^∞ diffeomorphism onto its image, and we denote its derivative at a point $U \in B$ by $dS_t(U) \in \mathbb{R}^{p\times p}$. We will use the fact that the mapping $B \times [0, T] \ni (U, t) \mapsto dS_t(U) \in \mathbb{R}^{p\times p}$ is continuous in U and continuously differentiable in t, and we note that $dS_t(U)$ is invertible for each $U \in B$ and $t \in [0, T]$. Hence, by compactness, there exists real numbers $C_i > 0$ $(i = 1, \dots, 4)$ such that

$$C_1 \le |||dS_t(U)||| \le C_2 \quad \text{and} \quad C_3 \le |||dS_t(U)^{-1}||| \le C_4, \tag{2.3}$$

for all $U \in B$ and $t \in [0, T]$, where $||| \cdot |||$ denotes the Frobenius norm on $\mathbb{R}^{p\times p}$.

We will consider one-step numerical methods for (2.1) of the form

$$\mathcal{G}_{\Delta t}(U_n, U_{n+1}) = 0, \tag{2.4}$$

where $\mathcal{G}_{\Delta t} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^p$ is a given C^∞ map that depends smoothly on the parameter Δt. For any $u_0 \in \mathbb{R}^p$ we assume the numerical scheme generates a local evolution semigroup in the sense that there is a closed ball $\mathcal{B}$ at u_0, real numbers $h, \mathcal{T} > 0$, and a mapping $\bar{S}_{\Delta t} : \mathcal{B} \to \mathbb{R}^p$ such that for any $U \in \mathcal{B}$ and $\Delta t \in [0, h]$ the sequence $\{U_n\}$ generated by

$$U_n = \bar{S}^n_{\Delta t}(U) \tag{2.5}$$

satisfies (2.4) for all $n\Delta t \in [0, \mathcal{T}]$. Here $\bar{S}^n_{\Delta t}$ denotes the n-fold composition of the map $\bar{S}_{\Delta t}$.

Given any $u_0 \in \mathbb{R}^p$ we assume without loss of generality that $\mathcal{B} = B$ and $\mathcal{T} = T$. Furthermore, we assume the numerical scheme is consistent of order r as an approximation to (2.1); that is, for any $U \in B$ we have

$$\frac{\partial^i}{\partial t^i}\Big|_{t=0} \bar{S}_t(U) = \frac{\partial^i}{\partial t^i}\Big|_{t=0} S_t(U), \qquad i = 1, \dots, r, \tag{2.6}$$

where $r \geq 1$ by consistency.

For any $n\Delta t \in [0,T]$ with $\Delta t \in [0,h]$ let $d\bar{S}^n_{\Delta t}(U) \in \mathbb{R}^{p\times p}$ denote the derivative of $\bar{S}^n_{\Delta t} : B \to \mathbb{R}^p$ at a point $U \in B$, and let $||\cdot||$ denote the Euclidean norm on $\mathbb{R}^p$. Then, by standard results from the numerical analysis of ordinary differential equations (see, for example, Stuart & Humphries [1996, Theorem 6.2.1]) there exist real numbers $C_5 > 0$ and $C_6 > 0$ depending on $U \in B$ and T such that

$$||S_t(U) - \bar{S}^n_{\Delta t}(U)|| \leq C_5 \Delta t^r \tag{2.7}$$

and

$$|||dS_t(U) - d\bar{S}^n_{\Delta t}(U)||| \leq C_6 \Delta t^r \tag{2.8}$$

for any $t = n\Delta t \in [0,T]$ with $\Delta t \in [0,h]$. Additionally, in view of (2.3) and (2.8), there is a real number $C_7 > 0$ depending on $U \in B$ and T such that the derivative of the mapping $\bar{S}^n_{\Delta t} : B \to \mathbb{R}^p$ satisfies

$$|||d\bar{S}^n_{\Delta t}(U)||| \leq C_7. \tag{2.9}$$

3. Associated Modified Equations.

To any ordinary differential equation of the form (2.1), and numerical approximation scheme (2.4) of order r, we can associate a *modified equation* of index N of the form

$$\frac{dv}{dt} = \tilde{f}^{(N)}_{\Delta t}(v), \tag{3.1}$$

where $N \geq 1$ is an integer and the *modified vector field* $\tilde{f}^{(N)}_{\Delta t} : \mathbb{R}^p \to \mathbb{R}^p$ is defined as

$$\tilde{f}^{(N)}_{\Delta t}(v) = f(v) + \sum_{i=1}^{N} \Delta t^{r+i-1} q_i(v) \tag{3.2}$$

for some functions $q_i : \mathbb{R}^p \to \mathbb{R}^p$ $(i = 1, \ldots, N)$. It is convenient to define the modified equation of index zero to be the original equation (2.1). Thus we have

$$\tilde{f}^{(0)}_{\Delta t}(u) = f(u), \qquad \forall u \in \mathbb{R}^p. \tag{3.3}$$

For any $v_0 \in \mathbb{R}^p$ we denote by $\tilde{S}^{(N)} : \tilde{B} \times [0,\tilde{T}] \times [0,\tilde{h}] \to \mathbb{R}^p$ the local evolution semigroup generated by (3.1) where $\tilde{B}$ is a closed ball at v_0 and $\tilde{h}, \tilde{T} > 0$. In particular, for any $V \in \tilde{B}$ and $\Delta t \in [0,\tilde{h}]$ the curve defined by

$$v_{\Delta t}(t) = \tilde{S}^{(N)}_{\Delta t}(V,t) = \tilde{S}^{(N)}_{t,\Delta t}(V) \tag{3.4}$$

is a solution to (3.1) with initial condition $v_{\Delta t}(0) = V$, defined for all $t \in [0,\tilde{T}]$. For any $t \in [0,\tilde{T}]$ and $\Delta t \in [0,\tilde{h}]$ we denote by $d\tilde{S}^{(N)}_{t,\Delta t}(V) \in \mathbb{R}^{p\times p}$ the derivative of the mapping $\tilde{S}^{(N)}_{t,\Delta t} : \tilde{B} \to \mathbb{R}^p$ at a point $V \in \tilde{B}$. As for the underlying system, we will use the fact that, for any $\Delta t \in [0, \tilde{h}]$, the mapping $\tilde{B} \times [0, \tilde{T}] \ni (V, t) \mapsto d\tilde{S}^{(N)}_{t,\Delta t}(V) \in \mathbb{R}^{p\times p}$ is continuous in V and continuously differentiable in t, and we note that $d\tilde{S}^{(N)}_{t,\Delta t}(V)$ is invertible for each $V \in \tilde{B}$ and $t \in [0, \tilde{T}]$. In what follows we will consider the semigroups generated by (2.1), (2.4) and (3.1) in the neighborhood of a common point in $\mathbb{R}^p$. In this case, without loss of generality, we take $\tilde{B} = B$ and $\tilde{T} = T$.

With the above notation in hand we next show that, given an integer $N \geq 1$, it is possible to construct C^∞ functions q_i $(i = 1, \dots, N)$ such that the numerical scheme (2.4) is an order $r + N$ approximation to (3.1). The local construction of the functions q_i is outlined in the lemma below and is based on the observation that the N^{th} modified vector field (3.2) differs from the $(N + 1)^{st}$ by a single term, that is

$$\tilde{f}^{(N+1)}_{\Delta t}(v) = \tilde{f}^{(N)}_{\Delta t}(v) + \Delta t^{r+N} q_{N+1}(v).$$

With this observation the basic strategy for constructing the functions q_i becomes clear: choose the q_i such that each new term in (3.2) improves the order of approximation by one power of Δt.

Lemma 3.1. *Given $u_0 \in \mathbb{R}^p$ suppose that for some integer $s > r$ the expansions*

$$S_{\Delta t}(u_0) = u_0 + \sum_{i=1}^{s} \Delta t^i F_i(u_0) + \mathcal{O}(\Delta t^{s+1})$$

$$\bar{S}_{\Delta t}(u_0) - S_{\Delta t}(u_0) = \sum_{i=r+1}^{s} \Delta t^i F^{(0)}_i(u_0) + \mathcal{O}(\Delta t^{s+1})$$

are valid, where $\{F_i\}_{i=1}^s$ and $\{F^{(0)}_i\}_{i=r+1}^s$ are C^∞. Then, for every integer N such that $0 \leq N \leq s - r - 1$, there exists a modified equation (3.1) *with the property*

$$\bar{S}_{\Delta t}(u_0) - \tilde{S}^{(N)}_{\Delta t,\Delta t}(u_0) = \sum_{i=r+N+1}^{s} \Delta t^i F^{(N)}_i(u_0) + \mathcal{O}(\Delta t^{s+1}) \tag{3.5}$$

where $\{F^{(N)}_i\}_{i=r+N+1}^s$ are C^∞.

Proof. By assumption, the result is true for $N = 0$. Assuming $s > r + 1$ we proceed by induction. Suppose the result is true for some N with $0 \leq N < s - r - 1$ and let $q_{N+1}(v) := F^{(N)}_{r+N+1}(v)$, so that

$$\tilde{f}^{(N+1)}_{\Delta t}(v) = \tilde{f}^{(N)}_{\Delta t}(v) + \Delta t^{r+N} F^{(N)}_{r+N+1}(v). \tag{3.6}$$

Expanding $\tilde{S}^{(N+1)}_{t,\Delta t}(u_0)$ for small t and setting $t = \Delta t$ leads to an expression of the form

$$\tilde{S}^{(N+1)}_{\Delta t,\Delta t}(u_0) = u_0 + \sum_{i=1}^{s} \Delta t^i \widehat{F}_i(u_0) + \mathcal{O}(\Delta t^{s+1}) \tag{3.7}$$

for some $\{\widehat{F}_i\}_{i=1}^{s}$ in C^∞. By comparing Taylor expansions of $\tilde{S}^{(N)}_{t,\Delta t}(u_0)$ and $\tilde{S}^{(N+1)}_{t,\Delta t}(u_0)$ with $t = \Delta t$, we find from (3.6) that

$$\tilde{S}^{(N)}_{\Delta t,\Delta t}(u_0) - \tilde{S}^{(N+1)}_{\Delta t,\Delta t}(u_0) = -\Delta t^{r+N+1} F^{(N)}_{r+N+1}(u_0) + \mathcal{O}(\Delta t^{r+N+2}).$$

Hence, using the induction hypothesis (3.5),

$$\begin{aligned}\bar{S}_{\Delta t}(u_0) - \tilde{S}^{(N+1)}_{\Delta t,\Delta t}(u_0) &= [\bar{S}_{\Delta t}(u_0) - \tilde{S}^{(N)}_{\Delta t,\Delta t}(u_0)] + [\tilde{S}^{(N)}_{\Delta t,\Delta t}(u_0) - \tilde{S}^{(N+1)}_{\Delta t,\Delta t}(u_0)] \\ &= \Delta t^{r+N+1} F^{(N)}_{r+N+1}(u_0) - \Delta t^{r+N+1} F^{(N)}_{r+N+1}(u_0) + \mathcal{O}(\Delta t^{r+N+2}) \\ &= \mathcal{O}(\Delta t^{r+N+2}).\end{aligned}$$

This, along with (3.7), gives the required result. ∎

Henceforth, we will always assume that the expansions in the statement of Lemma 3.1 are valid for all integers $s > r$ and that the functions q_i, and hence the modified equations (3.1), are chosen so that (3.5) holds. Since the original vector field f is C^∞ this assumption automatically holds for $S_{\Delta t}$ and will also hold for most methods $\bar{S}_{\Delta t}$ used in practice. The following result then follows from a standard Gronwall-based convergence analysis.

Theorem 3.2. *Given any $u_0 \in \mathbb{R}^p$ and any integer $N \geq 1$ there exists a ball B at u_0, real numbers $h, T > 0$, and smooth functions q_i $(i = 1, \ldots, N)$ such that the local evolution semigroups $S_t, \bar{S}_{\Delta t}, \tilde{S}^{(N)}_{t,\Delta t} : B \to \mathbb{R}^p$ for (2.1), (2.4) and (3.1), respectively, are defined for all $t = n\Delta t \in [0,T]$ with $\Delta t \in [0,h]$. Furthermore, there is a constant $C_8 = C_8(T,N,B) > 0$ such that for each $U \in B$*

$$|||d\tilde{S}^{(N)}_{t,\Delta t}(U) - d\bar{S}^{n}_{\Delta t}(U)||| + ||\tilde{S}^{(N)}_{t,\Delta t}(U) - \bar{S}^{n}_{\Delta t}(U)|| \leq C_8 \Delta t^{r+N}, \tag{3.8}$$

for all $t = n\Delta t \in [0,T]$ with $\Delta t \in [0,h]$.

Remarks

(1) Note that combining (2.7), (2.8) and (3.8) (and without loss of generality taking the same constant) gives the bound

$$|||d\tilde{S}^{(N)}_{t,\Delta t}(U) - dS_t(U)||| + ||\tilde{S}^{(N)}_{t,\Delta t}(U) - S_t(U)|| \leq C_8 \Delta t^{r}, \tag{3.9}$$

which compares solutions of the original and modified equations.

(2) The key result (3.8) shows that the numerical method applied to the original system (2.1) behaves like an order $r + N$ method with respect to the modified equation of index N. In other words, the method computes a very accurate approximation to a perturbed problem. This idea has links with the concept of backward error in numerical linear algebra, see Golub & Van Loan [1996].

It is natural to ask whether the perturbed problem has the same structure as the original problem; in the linear algebra context this is known as structured backward error analysis. In the next section we address this question for a variety of different families of ordinary differential equations.

(3) Note that the constant C_8 in (3.8) and (3.9) depends upon T, N and B. Hence, in particular, the bounds are valid only for finite time intervals and they are not uniform in the index N. However, by optimizing over the index Neishtadt [1984], Benettin & Giorgilli [1994], Hairer & Lubich [1997] and Reich [1996] have shown that the difference between the numerical approximation and a modified equation remains exponentially small over arbitrarily long time intervals as $\Delta t \to 0$ for a variety of problems of interest.

4. Qualitative Properties of the Modified Equations.

In this section we employ an induction on N to establish various properties of the modified equation (3.1). In view of Theorem 3.2 we see that, given any $u_0 \in \mathbb{R}^p$, the ball B at u_0 and the values $h, T > 0$ will in general depend upon N. In the following induction arguments we will choose a ball B and numbers $h, T > 0$ such that all the local evolution semigroups $S_t, \bar{S}_{\Delta t}, \tilde{S}^{(m)}_{t,\Delta t} : B \to \mathbb{R}^p$ $(m = 1, \ldots, N+1)$ are defined for any $t = n\Delta t \in [0, T]$ with $\Delta t \in [0, h]$. Note that h may shrink to zero as $N \to \infty$, but will be non-zero for every fixed integer $N \geq 1$.

By virtue of Theorem 3.2 we may assume, without loss of generality, that the same constants C_1, C_2, C_3 and C_4 which appear in (2.3) may be used to bound the derivatives of the semigroups for the modified equations up to index $N + 1$. Thus, for $1 \leq m \leq N + 1$,

$$C_1 \leq |||d\tilde{S}^{(m)}_{t,\Delta t}(U)||| \leq C_2 \quad \text{and} \quad C_3 \leq |||d\tilde{S}^{(m)}_{t,\Delta t}(U)^{-1}||| \leq C_4. \tag{4.1}$$

4.1. *Integrals for the Modified Semigroup.*

Suppose that the underlying system (2.1) and the approximation scheme (2.4) share an integral $\mathcal{F} \in C^1(\mathbb{R}^p, \mathbb{R})$. That is, for any $u_0 \in \mathbb{R}^p$ the function $\mathcal{F}$ is invariant under the local semigroups S_t and $\bar{S}_{\Delta t}$ in the sense that, for any $U \in B$ and $\Delta t \in [0, h]$, we have $\mathcal{F}(S_t(U)) = \mathcal{F}(U)$ and $\mathcal{F}(\bar{S}^n_{\Delta t}(U)) = \mathcal{F}(U)$ for all $t \in [0, T]$ and $n\Delta t \in [0, T]$. We then have the following result, which is also proved in Reich [1993], [1996] by Lie algebraic methods.

Theorem 4.1. *Suppose the underlying system* (2.1) *and the approximation scheme* (2.4) *share an integral* $\mathcal{F} \in C^1(\mathbb{R}^p, \mathbb{R})$*. Then* $\mathcal{F}$ *is an integral for the associated modified equation* (3.1) *of index* N *for any integer* $N \geq 1$*. Hence, the modified* *equation* (3.1) *has the form*

$$\frac{dv}{dt} = f(v) + \Delta t^r \sum_{i=1}^{N} \Delta t^{i-1} q_i(v),$$

where

$$\nabla \mathcal{F}(v) \cdot q_i(v) = 0, \quad \forall v \in \mathbb{R}^p,\ i = 1, \ldots, N.$$

Proof. For induction assume the modified equation of index N has $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$ as an integral. Note that this is true for $N = 0$ since the modified equation of order zero is the original equation (2.1).

Consider any $u_0 \in \mathbb{R}^p$. Then, for any $U \in B$ and $\Delta t \in [0, h]$ we have

$$\mathcal{F}(\tilde{S}^{(N)}_{t,\Delta t}(U)) = \mathcal{F}(U), \tag{4.2}$$

for all $t \in [0, T]$. Equivalently, for any $\Delta t \in [0, h]$, we have

$$\nabla \mathcal{F}(u) \cdot \tilde{f}^{(N)}_{\Delta t}(u) = 0, \qquad \forall u \in \mathcal{I}m\big(\tilde{S}^{(N)}_{\Delta t}\big), \tag{4.3}$$

where

$$\mathcal{I}m\big(\tilde{S}^{(N)}_{\Delta t}\big) = \{u \in \mathbb{R}^p \mid u = \tilde{S}^{(N)}_{t,\Delta t}(U), \quad U \in B, \quad t \in [0, T]\}. \tag{4.4}$$

Now assume, for contradiction, that $\mathcal{F}$ is not an integral for the modified equation of index $N + 1$, which is of the form

$$\frac{dv}{dt} = \tilde{f}^{(N+1)}_{\Delta t}(v) = \tilde{f}^{(N)}_{\Delta t}(v) + \Delta t^{r+N} q_{N+1}(v). \tag{4.5}$$

Then there exists $u_0 \in \mathbb{R}^p$ such that

$$\nabla \mathcal{F}(u_0) \cdot q_{N+1}(u_0) \neq 0. \tag{4.6}$$

Otherwise, $\nabla \mathcal{F}(u) \cdot \tilde{f}^{(N+1)}_{\Delta t}(u) = 0$ for all $u \in \mathbb{R}^p$ and $\mathcal{F}$ would be an integral.

Let $C_9(u_0) = \nabla \mathcal{F}(u_0) \cdot q_{N+1}(u_0)/2 \neq 0$ and assume, without loss of generality, that $C_9 > 0$; otherwise, if $C_9 < 0$, then one can redefine $\mathcal{F}$ by changing sign. By continuity there is a closed ball D at u_0 such that

$$\nabla \mathcal{F}(U) \cdot q_{N+1}(U) \geq C_9 > 0, \qquad \forall U \in D. \tag{4.7}$$

Consider a point $U \in D \cap B$ and let $h, T > 0$ be such that, for any $\Delta t \in [0, h]$, the evolution semigroups satisfy $\tilde{S}^{(N+1)}_{t,\Delta t}(U) \in D$ for all $t \in [0, T]$ and $U_n = \bar{S}^n_{\Delta t}(U) \in D$

for all $n\Delta t \in [0,T]$. Then, for any $\Delta t \in [0,h]$ and $t \in [0,T]$, we have by (4.3) and (4.5)

$$\begin{aligned} \frac{\partial}{\partial\tau}\Big|_{\tau=t} \mathcal{F}(\tilde{S}^{(N+1)}_{\tau,\Delta t}(U)) &= \nabla\mathcal{F}(\tilde{S}^{(N+1)}_{t,\Delta t}(U)) \cdot \tilde{f}^{(N+1)}_{\Delta t}(\tilde{S}^{(N+1)}_{t,\Delta t}(U)) \\ &= \Delta t^{r+N} \nabla\mathcal{F}(\tilde{S}^{(N+1)}_{t,\Delta t}(U)) \cdot q_{N+1}(\tilde{S}^{(N+1)}_{t,\Delta t}(U)) \\ &\geq C_9 \Delta t^{r+N}, \end{aligned} \tag{4.8}$$

which implies

$$|\mathcal{F}(\tilde{S}^{(N+1)}_{T,\Delta t}(U)) - \mathcal{F}(U)| \geq C_9 T \Delta t^{r+N}, \tag{4.9}$$

for all $\Delta t \in [0,h]$.

By compactness of the closed ball D, since $\mathcal{F} \in C^1(\mathbb{R}^p, \mathbb{R})$, there is a real number $C_{10} > 0$ such that

$$|\mathcal{F}(U) - \mathcal{F}(V)| \leq C_{10}||U - V||, \quad \forall U, V \in D. \tag{4.10}$$

Furthermore, in view of (3.8), the modified equation of index $N+1$ and the numerical scheme (2.4) have solutions satisfying

$$||\tilde{S}^{(N+1)}_{T,\Delta t}(U) - \bar{S}^n_{\Delta t}(U)|| \leq C_8 \Delta t^{r+N+1}, \tag{4.11}$$

for all $\Delta t = T/n$ and $n \geq n^*$, where n^* is any positive integer such that $T/n^* \in [0,h]$. Since by hypothesis $\mathcal{F}$ is an integral for the local numerical semigroup we use (4.10) and (4.11) to write

$$\begin{aligned} |\mathcal{F}(\tilde{S}^{(N+1)}_{T,\Delta t}(U)) - \mathcal{F}(U)| &= |\mathcal{F}(\tilde{S}^{(N+1)}_{T,\Delta t}(U)) - \mathcal{F}(\bar{S}^n_{\Delta t}(U))| \\ &\leq C_{10}||\tilde{S}^{(N+1)}_{T,\Delta t}(U) - \bar{S}^n_{\Delta t}(U)|| \\ &\leq C_8 C_{10} \Delta t^{r+N+1}, \end{aligned} \tag{4.12}$$

for all $\Delta t = T/n$ and $n \geq n^*$. This yields a contradiction, since for $\Delta t < TC_9/C_8C_{10}$ both (4.9) and (4.12) cannot hold. Hence $\mathcal{F}$ must be an integral for the modified equation of index $N+1$. The result follows by induction. ■

The same argument can be used in the case where a method fails to preserve an integral exactly, but preserves it to higher accuracy than the classic order would suggest. See, for example, Calvo, Iserles & Zanna [1996] for instances of such methods.

Theorem 4.2. *Suppose that the system* (2.1) *preserves an integral* $\mathcal{F} \in C^1(\mathbb{R}^p, \mathbb{R})$, *and that given any* $U \in B$, *with* B *compact, there exist* $h, T > 0$ *such that*

$$|\,\mathcal{F}(\bar{S}^n_{\Delta t}(U)) - \mathcal{F}(U)\,| \leq C_{11}\Delta t^l, \quad \forall \Delta t \in [0,h],\ n\Delta t \in [0,T],$$

for some constant $C_{11} = C_{11}(B,T)$ *and integer* $l > r$. *Then all modified equations of index up to and including* $l-r$ *also preserve* $\mathcal{F}$*; that is,*

$$\nabla\mathcal{F}(v)\cdot q_i(v) = 0, \quad \forall v \in \mathbb{R}^p,\ i = 1,\ldots,l-r.$$

Proof. Applying the induction argument used in the previous theorem, we find that the bound (4.12) is degraded by a term of $\mathcal{O}(\Delta t^l)$, and hence the contradiction remains for indices up to $l-r$. ■

We note that Theorem 3.2 yields two corollaries to the above result:

(i) Over any finite time interval the numerical solution approximates to $\mathcal{O}(\Delta t^l)$ a modified equation that preserves $\mathcal{F}$ (namely, the modified equation of index $l-r$).

(ii) For any $N \geq 1$ the modified equation of index N approximates the numerical solution to $\mathcal{O}(\Delta t^{r+N})$ and preserves $\mathcal{F}$ to within $\mathcal{O}(\Delta t^l)$ over any finite time interval.

4.2. *Reversibility and the Modified Semigroup.*

The system (2.1) is said to be *reversible* if there exists an invertible linear transformation $\rho : \mathbb{R}^p \to \mathbb{R}^p$ such that

$$f\cdot\rho(y) = -\rho\cdot f(y), \quad \forall y \in \mathbb{R}^p, \tag{4.13}$$

where $\cdot$ denotes composition. (This definition, which is also used by Stoffer [1995] and Hairer & Stoffer [1997], is more general than the definition of reversibility that is found in some texts, such as Strogatz [1994]). The implication of property (4.13) on local semigroups can be summarized as follows. Consider two compact sets B_1 and B_2 such that $B_2 = \rho(B_1)$, and let $S_{1,t}$, $t \in [0,T_1]$ and $S_{2,t}$, $t \in [0,T_2]$ be the local semigroups for (2.1) defined on B_1 and B_2. (Note that $S_{1,t}$ and $S_{2,t}$ can be identified by extending their domain to a compact set including both B_1 and B_2 and possibly reducing the time-interval on which they are defined; similar considerations apply to the numerical method and to the modified equations considered below). The reversibility property (4.13) implies that the time domains of $S_{1,t}$ and $S_{2,t}$ can be extended to $[-T_2,T_1]$ and $[-T_1,T_2]$, respectively, and implies the semigroups enjoy the property

$$\rho\cdot S_{1,t}(y) = S_{2,-t}\cdot\rho(y), \quad \forall y \in B_1,\ t\in[-T_2,T_1]. \tag{4.14}$$

Moreover, for any $y \in B_1$ and $T > 0$ such that $S_{1,t}(y) \in B_1$ for all $t \in [-T,T]$, we have

$$S_{2,t}\cdot\rho\cdot S_{1,t}(y) = \rho(y), \quad \forall t \in [-T,T]. \tag{4.15}$$

Another way to characterize the above relations that will prove useful in our developments is the following. Define a function $\psi : [-T, T] \to \mathbb{R}^p$ by

$$\psi(t) = \rho \cdot S_{1,t}(y) - S_{2,-t} \cdot \rho(y). \tag{4.16}$$

Differentiating with respect to time and using (4.16) we see that $\psi(t)$ satisfies an equation of the form

$$\frac{d\psi}{dt} = g(t, \psi) \quad \text{where} \quad g(t, \psi) = \rho \cdot f(S_{1,t}(y)) + f(\rho \cdot S_{1,t}(y) - \psi). \tag{4.17}$$

Since $\psi(0) = 0$, and (4.13) implies $g(t, 0) = 0$ for all $t \in [-T, T]$, we deduce that $\psi(t) \equiv 0$ is the unique solution to (4.17). Hence (4.14) holds, and (4.15) follows from the properties of the semigroup.

Extending the notation established above, a one-step method such as (2.4) is said to be reversible if, whenever the structure (4.13) is present, we have

$$\bar{S}_{2,\Delta t} \cdot \rho \cdot \bar{S}_{1,\Delta t}(y) = \rho(y)$$

for any $y \in B_1$ and $\Delta t \in [0, h]$ such that $\bar{S}_{1,\Delta t}(y) \in B_1$. Here $\bar{S}_{1,\Delta t}$ and $\bar{S}_{2,\Delta t}$ are the local semigroups on B_1 and B_2 generated by the method. Stoffer [1995] showed that all symmetric Runge-Kutta methods are reversible in this sense, and Hairer & Stoffer [1997] showed that when a symmetric Runge-Kutta method is applied to a reversible system, all modified equations are reversible. Below, we use a different technique to show that the same result holds for all one-step methods of the form given in Section 2. A similar result can also be proved by modifying the techniques presented in Reich [1996].

Theorem 4.3. *If the equation* (2.1) *and the numerical method* (2.4) *are reversible, then so are all modified equations. More precisely, assume* (4.13) *holds and assume for any compact sets B_1 and B_2 such that $B_2 = \rho(B_1)$ the method defined by* (2.4) *has the property*

$$\bar{S}_{2,\Delta t} \cdot \rho \cdot \bar{S}_{1,\Delta t}(y) = \rho(y) \tag{4.18}$$

for any $y \in B_1$ and $\Delta t \in [0, h]$ such that $\bar{S}_{1,\Delta t}(y) \in B_1$. Then, for any $N \geq 1$,

$$q_i \cdot \rho(y) = -\rho \cdot q_i(y), \quad \forall y \in \mathbb{R}^p, \ i = 1, \ldots, N$$

so that

$$\tilde{f}^{(N)}_{\Delta t} \cdot \rho(y) = -\rho \cdot \tilde{f}^{(N)}_{\Delta t}(y) \quad \forall y \in \mathbb{R}^p.$$

Proof. For induction assume the modified equations up to index N are reversible, that is

$$\tilde{f}^{(i)}_{\Delta t} \cdot \rho(y) = -\rho \cdot \tilde{f}^{(i)}_{\Delta t}(y), \qquad 0 \leq i \leq N, \ y \in \mathbb{R}^p. \tag{4.19}$$

Note that this is true for $N = 0$ since the modified equation of index zero is the original equation (2.1).

Now suppose the modified equation of index $N+1$ is not reversible. This implies that

$$q_{N+1} \cdot \rho(u_0) \neq -\rho \cdot q_{N+1}(u_0) \tag{4.20}$$

at some point $u_0 \in \mathbb{R}^p$. By continuity, there exists a ball D around u_0 and some constant $C_{12} > 0$ such that

$$\| q_{N+1} \cdot \rho(y) + \rho \cdot q_{N+1}(y) \|_\infty \geq C_{12}, \quad \forall y \in D. \tag{4.21}$$

Let $B_1 = D$ and $B_2 = \rho(B_1)$, and for $i = 1, 2$ let $S_{i,t}, \bar{S}^n_{i,\Delta t}, \tilde{S}^{(m)}_{i,t,\Delta t} : B_i \to \mathbb{R}^p$ $(m = 1, \ldots, N+1)$ be the semigroups generated by (2.1), (2.4) and (3.1), respectively. Given $y = u_0$ let $h, T > 0$ be sufficiently small such that, for any $\Delta t \in [0, h]$, we have $S_{1,t}(y), \tilde{S}^{(N+1)}_{1,t,\Delta t}(y) \in B_1$ for all $t \in [-T, T]$ and $\bar{S}^n_{1,\Delta t}(y) \in B_1$ for any $n\Delta t \in [0, T]$.

Define a function $\psi : [-T, T] \to \mathbb{R}^p$ by

$$\psi(t) = \rho \cdot \tilde{S}^{(N+1)}_{1,t,\Delta t}(y) - \tilde{S}^{(N+1)}_{2,-t,\Delta t} \cdot \rho(y).$$

Then differentiating with respect to time gives

$$\frac{d\psi}{dt}(t) = \rho \cdot \tilde{f}^{(N+1)}_{\Delta t}(\tilde{S}^{(N+1)}_{1,t,\Delta t}(y)) + \tilde{f}^{(N+1)}_{\Delta t}(\tilde{S}^{(N+1)}_{2,-t,\Delta t} \cdot \rho(y)),$$

and by definition of $\psi(t)$ we obtain

$$\frac{d\psi}{dt}(t) = \rho \cdot \tilde{f}^{(N+1)}_{\Delta t}(\tilde{S}^{(N+1)}_{1,t,\Delta t}(y)) + \tilde{f}^{(N+1)}_{\Delta t}(\rho \cdot \tilde{S}^{(N+1)}_{1,t,\Delta t}(y) - \psi(t)). \tag{4.22}$$

Using (3.2), (4.13) and (4.19) we have

$$\rho \cdot \tilde{f}^{(N+1)}_{\Delta t}(v) = \Delta t^{r+N}[\rho \cdot q_{N+1}(v) + q_{N+1} \cdot \rho(v)] - \tilde{f}^{(N+1)}_{\Delta t} \cdot \rho(v)$$

and substituting into (4.22) yields

$$\begin{aligned}\frac{d\psi}{dt}(t) = {} & \Delta t^{r+N} \left[\rho \cdot q_{N+1}(\tilde{S}^{(N+1)}_{1,t,\Delta t}(y)) + q_{N+1} \cdot \rho(\tilde{S}^{(N+1)}_{1,t,\Delta t}(y))\right] \\ & + [\tilde{f}^{(N+1)}_{\Delta t}(\rho \cdot \tilde{S}^{(N+1)}_{1,t,\Delta t}(y) - \psi(t)) - \tilde{f}^{(N+1)}_{\Delta t} \cdot \rho(\tilde{S}^{(N+1)}_{1,t,\Delta t}(y))].\end{aligned}$$

This expression may be written as

$$\frac{d\psi}{dt}(t) = \Delta t^{r+N} \left[\rho \cdot q_{N+1}(\tilde{S}^{(N+1)}_{1,t,\Delta t}(y)) + q_{N+1} \cdot \rho(\tilde{S}^{(N+1)}_{1,t,\Delta t}(y))\right] + A(t)\psi(t),$$

where $A(t)$ is obtained from the integral form of the mean value theorem for vector-valued functions. If $B(t)$ is the fundamental matrix solving

$$\frac{dB}{dt}(t) = A(t)B(t), \quad B(0) = I,$$

then

$$\psi(t) = \Delta t^{r+N} B(t) \int_0^t B(s)^{-1} \left[\rho \cdot q_{N+1}(\tilde{S}^{(N+1)}_{1,s,\Delta t}(y)) + q_{N+1} \cdot \rho(\tilde{S}^{(N+1)}_{1,s,\Delta t}(y))\right] ds.$$

Now assume for contradiction that

$$B(t) \int_0^t B(s)^{-1} \left[\rho \cdot q_{N+1}(S_{1,s}(y)) + q_{N+1} \cdot \rho(S_{1,s}(y))\right] \, ds = 0$$

for all $t \in [0, T]$, with T sufficiently small. Since $B(t)$ is close to the identity for small t, a straightforward continuity argument shows that

$$\rho \cdot q_{N+1}(S_{1,s}(y)) + q_{N+1} \cdot \rho(S_{1,s}(y)) = 0, \quad \forall s \in [0, T].$$

This contradicts (4.21). Thus there are $\tau \in [0, T]$ and $\delta > 0$ such that

$$||B(\tau) \int_0^\tau B(s)^{-1} \left[\rho \cdot q_{N+1}(S_{1,s}(y)) + q_{N+1} \cdot \rho(S_{1,s}(y))\right] \, ds|| \geq 2\delta.$$

In view of (3.9) we have

$$||B(\tau) \int_0^\tau B(s)^{-1} \left[\rho \cdot q_{N+1}(\tilde{S}^{(N+1)}_{1,s,\Delta t}(y)) + q_{N+1} \cdot \rho(\tilde{S}^{(N+1)}_{1,s,\Delta t}(y))\right] \, ds|| \geq \delta$$

for all $\Delta t \in [0, h]$ with h sufficiently small, and thus

$$||\psi(\tau)|| \geq \Delta t^{r+N}\delta, \quad \forall \Delta t \in [0, h]. \tag{4.23}$$

Now define

$$\Phi(\tau) = \tilde{S}^{(N+1)}_{2,\tau,\Delta t} \cdot \rho \cdot \tilde{S}^{(N+1)}_{1,\tau,\Delta t}(y) - \rho(y)$$

so that

$$\psi(\tau) = \tilde{S}^{(N+1)}_{2,-\tau,\Delta t}(\Phi(\tau) + \rho(y)) - \tilde{S}^{(N+1)}_{2,-\tau,\Delta t} \cdot \rho(y).$$

Since $\tilde{S}^{(N+1)}_{2,-t,\Delta t}$ is close to the identity mapping for t small, the mean value theorem and (4.23) yield

$$||\Phi(\tau)|| \geq \frac{\Delta t^{r+N}\delta}{2}, \quad \forall \Delta t \in [0, h], \tag{4.24}$$

after reducing T if necessary. However, since the numerical method is reversible, we have

$$\Phi(\tau) = \tilde{S}^{(N+1)}_{2,\tau,\Delta t} \cdot \rho \cdot \tilde{S}^{(N+1)}_{1,\tau,\Delta t}(y) - \bar{S}^n_{2,\Delta t} \cdot \rho \cdot \bar{S}^n_{1,\Delta t}(y)$$

for $n\Delta t = \tau$ with $\Delta t \in [0, h]$. By Theorem 3.2 we deduce that

$$||\Phi(\tau)|| \leq C_{13}\Delta t^{r+N+1}, \quad \forall \Delta t \in [0, h]. \tag{4.25}$$

Since (4.24) and (4.25) give a contradiction, we deduce that (4.20) cannot hold. This completes the inductive step. ■

4.3. *Fixed Points of the Modified Semigroup.*

The next theorem shows that any numerical method which inherits the equilibria of (2.1) as fixed points has modified equations which inherit these as equilibria.

Theorem 4.4. *Suppose that there is a point $y^* \in \mathbb{R}^p$ such that $f(y^*) = 0$ and $\bar{S}_{\Delta t}(y^*) = y^*$ for all $\Delta t > 0$. Then, for any $N \geq 1$, we have $q_i(y^*) = 0$ $(i = 1, \ldots, N)$ and hence*

$$\tilde{f}_{\Delta t}^{(N)}(y^*) = 0, \quad \forall \Delta t > 0.$$

Proof. For induction assume the modified vector fields up to index N have the property $\tilde{f}_{\Delta t}^{(N)}(y^*) = 0$ for all $\Delta t > 0$. Note that this is true for $N = 0$ since $\tilde{f}_{\Delta t}^{(0)} = f$. Now suppose $\tilde{f}_{\Delta t}^{(N+1)}(y^*) \neq 0$ for *some* $\Delta t > 0$. Since

$$\tilde{f}_{\Delta t}^{(N+1)}(v) = \tilde{f}_{\Delta t}^{(N)}(v) + \Delta t^{r+N} q_{N+1}(v)$$

this implies that $q_{N+1}(y^*) \neq 0$, and thus $\tilde{f}_{\Delta t}^{(N+1)}(y^*) \neq 0$ for *all* $\Delta t > 0$.

To obtain a contradiction, let D be a closed ball around y^* and let $h, T > 0$ be sufficiently small such that $\tilde{S}_{t,\Delta t}^{(N+1)}(y^*) \in D$ for all $t \in [0, T]$ and $\Delta t \in [0, h]$. Introduce a maximum Lipschitz constant in D by

$$L := \max\{\mathrm{Lip}[f], \mathrm{Lip}[q_1], \ldots, \mathrm{Lip}[q_{N+1}]\},$$

and let $y(t) = \tilde{S}_{t,\Delta t}^{(N+1)}(y^*)$ and $z(t) = y(t) - y^*$.

Now, since $\bar{S}_{\Delta t}^n(y^*) = y^*$, Theorem 3.2 shows that

$$||z(t)|| \leq C_8 \Delta t^{r+N+1} \tag{4.26}$$

for any $t = n\Delta t \in [0, T]$ with $\Delta t \in [0, h]$. By definition of L we thus have

$$||f(y(t)) - f(y^*)|| \leq L||z(t)|| \leq LC_8 \Delta t^{r+N+1} \tag{4.27}$$

and

$$||q_i(y(t)) - q_i(y^*)|| \leq LC_8 \Delta t^{r+N+1}, \quad 1 \leq i \leq N+1. \tag{4.28}$$

Using the fact that

$$\tilde{f}_{\Delta t}^{(N+1)}(y(t)) - \tilde{f}_{\Delta t}^{(N+1)}(y^*) = f(y(t)) - f(y^*) + \sum_{i=1}^{N} \Delta t^{r+i-1}[q_i(y(t)) - q_i(y^*)],$$

together with (4.27) and (4.28), it follows that there is a constant C_{14} such that

$$||\tilde{f}_{\Delta t}^{(N+1)}(y(t)) - \tilde{f}_{\Delta t}^{(N+1)}(y^*)|| \leq C_{14} \Delta t^{r+N+1}. \tag{4.29}$$

Finally, since $dz(t)/dt = \tilde{f}_{\Delta t}^{(N+1)}(y(t))$ and $z(0)=0$, we have from (4.29) that

$$\begin{aligned}||z(T)|| &= ||\int_0^T \tilde{f}_{\Delta t}^{(N+1)}(y(t))\,dt|| \\ &= ||\int_0^T \tilde{f}_{\Delta t}^{(N+1)}(y^*) + [\tilde{f}_{\Delta t}^{(N+1)}(y(t)) - \tilde{f}_{\Delta t}^{(N+1)}(y^*)]\,dt|| \\ &\geq ||\int_0^T \tilde{f}_{\Delta t}^{(N+1)}(y^*)\,dt|| - \int_0^T ||\tilde{f}_{\Delta t}^{(N+1)}(y(t)) - \tilde{f}_{\Delta t}^{(N+1)}(y^*)||\,dt \\ &\geq T\Delta t^{r+N}||q_{N+1}(y^*)|| - TC_{14}\Delta t^{r+N+1}.\end{aligned}$$

Hence, after reducing h if necessary, there exists a constant $C_{15}>0$ such that

$$||z(T)|| \geq C_{15}\Delta t^{r+N}.$$

This contradicts (4.26), and thus we must have $\tilde{f}_{\Delta t}^{(N+1)}(y^*) = 0$ for all $\Delta t > 0$. ■

4.4. *Symplecticity of the Modified Semigroup.*

Suppose now that the dimension p is even, say $p=2m$, and the vector field $f:\mathbb{R}^p \to \mathbb{R}^p$ is Hamiltonian with respect to the canonical symplectic structure; that is,

$$f(u) = J\nabla H(u) \tag{4.30}$$

for some smooth function $H:\mathbb{R}^p \to \mathbb{R}$, where $J \in \mathbb{R}^{p\times p}$ is of the form

$$J = \begin{pmatrix} O_m & I_m \\ -I_m & O_m \end{pmatrix}. \tag{4.31}$$

Here O_m and I_m denote the zero and identity matrices in $\mathbb{R}^{m\times m}$, respectively. We now show that when the numerical approximation scheme generates a symplectic semigroup, the associated modified equations share the same property. This result is already well-known (see Mackay [1992], Sanz-Serna [1992], Reich [1993], [1996], Benettin & Giorgilli [1994], Calvo, Murua & Sanz-Serna [1994] and Hairer [1994] for example), and we simply provide a new proof using the unifying technique of this article.

Theorem 4.5. *Suppose the underlying system* (2.1) *and the approximation scheme* (2.4) *both generate symplectic semigroups. Then the associated modified equation* (3.1) *of index N generates a symplectic semigroup for any integer $N \geq 1$. Thus the modified equation* (3.1) *has the form*

$$\frac{dv}{dt} = J\nabla[\,H(v) + \Delta t^r Q^{(N)}(v;\Delta t)\,]. \tag{4.32}$$

Proof. For any $u_0 \in \mathbb{R}^p$ let $S : B \times [0,T] \to \mathbb{R}^p$ denote the local semigroup generated by (4.30) with image defined as

$$\mathcal{I}m(S) = \{u \in \mathbb{R}^p \mid u = S(U,t), \quad U \in B, \quad t \in [0,T]\}, \tag{4.33}$$

and let $df(u) \in \mathbb{R}^{p\times p}$ denote the derivative at any point $u \in \mathbb{R}^p$ of the Hamiltonian vector field $f : \mathbb{R}^p \to \mathbb{R}^p$ given by (4.30).

Since the vector field we are considering is Hamiltonian, the mapping $S_t : B \to \mathbb{R}^p$ is symplectic for each $t \in [0,T]$ in the sense that

$$dS_t(U)^T J dS_t(U) = J \tag{4.34}$$

for all $U \in B$. An equivalent statement to (4.34) is that $df(u) \in \mathbb{R}^{p\times p}$ is *infinitesimally symplectic* for each $u \in \mathcal{I}m(S)$; that is,

$$df(u)^T J + J df(u) = 0 \tag{4.35}$$

for all $u \in \mathcal{I}m(S)$. That (4.35) holds follows from (4.30) since $R^T J + JR = 0$ for any matrix $R = JA$ where $A^T = A$.

For induction assume that, given any $u_0 \in \mathbb{R}^p$, the modified equation of index N generates a local evolution semigroup $\tilde{S}^{(N)}$ which is symplectic, and note that this is true for $N = 0$, the unperturbed equation (2.1). Then, for any $\Delta t \in [0,h]$ and $t \in [0,T]$, the mapping $\tilde{S}^{(N)}_{t,\Delta t} : B \to \mathbb{R}^p$ satisfies

$$d\tilde{S}^{(N)}_{t,\Delta t}(U)^T J d\tilde{S}^{(N)}_{t,\Delta t}(U) = J \tag{4.36}$$

for all $U \in B$. Equivalently, for any $\Delta t \in [0,h]$, the modified vector field $\tilde{f}^{(N)}_{\Delta t}$ satisfies

$$d\tilde{f}^{(N)}_{\Delta t}(u)^T J + J d\tilde{f}^{(N)}_{\Delta t}(u) = 0 \tag{4.37}$$

for all $u \in \mathcal{I}m(\tilde{S}^{(N)}_{\Delta t})$.

Now assume, for contradiction, that the modified equation of index $N+1$, which is of the form

$$\frac{dv}{dt} = \tilde{f}^{(N+1)}_{\Delta t}(v) = \tilde{f}^{(N)}_{\Delta t}(v) + \Delta t^{r+N} q_{N+1}(v), \tag{4.38}$$

generates a semigroup which is not symplectic. Then there exists $u_0 \in \mathbb{R}^p$ such that

$$\Phi(u_0) \neq 0 \tag{4.39}$$

where

$$\Phi(u) := dq_{N+1}(u)^T J + J dq_{N+1}(u). \tag{4.40}$$

Otherwise, $d\tilde{f}^{(N+1)}_{\Delta t}(u)$ would be infinitesimally symplectic for all $u \in I\!\!R^p$ and, for any $u_0 \in I\!\!R^p$, the local semigroup $\tilde{S}^{(N+1)}_{t,\Delta t}$ would be symplectic.

Let $C_{16} = C_{16}(u_0) > 0$ be such that

$$|||\Phi(u_0)||| = 2C_{16} > 0. \tag{4.41}$$

Since q_{N+1} is C^∞ smooth, it follows by continuity that there is a closed ball D at u_0 such that

$$|||\Phi(U)||| \geq C_{16} > 0, \qquad \forall U \in D. \tag{4.42}$$

Given $U \in D \cap B$ let $h, T > 0$ be such that, for any $\Delta t \in [0, h]$, the semigroups satisfy $S(U,t), \tilde{S}^{(N+1)}_{\Delta t}(U,t) \in D$ for all $t \in [0,T]$ and $U_n = \bar{S}^n_{\Delta t}(U) \in D$ for all $n\Delta t \in [0,T]$. Furthermore, let $V(t) = d\tilde{S}^{(N+1)}_{\Delta t}(U,t)$ and note that $V(t)$ satisfies the matrix equation

$$\frac{dV}{dt} = d\tilde{f}^{(N+1)}_{\Delta t}(u_{\Delta t}(t))V, \quad V(0) = I_p \tag{4.43}$$

where we have used the notation $u_{\Delta t}(t) = \tilde{S}^{(N+1)}_{\Delta t}(U,t)$. Using this relation together with (4.37) and (4.38) it follows that, for any $\Delta t \in [0,h]$, we have

$$\begin{aligned}
&\frac{\partial}{\partial \tau}\bigg|_{\tau=t} \left(d\tilde{S}^{(N+1)}_{\Delta t}(U,\tau)^T J d\tilde{S}^{(N+1)}_{\Delta t}(U,\tau)\right) \\
&\quad = d\tilde{S}^{(N+1)}_{\Delta t}(U,t)^T \left(d\tilde{f}^{(N+1)}_{\Delta t}(u_{\Delta t}(t))^T J + J d\tilde{f}^{(N+1)}_{\Delta t}(u_{\Delta t}(t))\right) d\tilde{S}^{(N+1)}_{\Delta t}(U,t) \\
&\quad = \Delta t^{r+N} d\tilde{S}^{(N+1)}_{\Delta t}(U,t)^T \left(dq_{N+1}(u_{\Delta t}(t))^T J + J dq_{N+1}(u_{\Delta t}(t))\right) d\tilde{S}^{(N+1)}_{\Delta t}(U,t).
\end{aligned} \tag{4.44}$$

For convenience we introduce the notation

$$\mathcal{A}_{\Delta t}(U,t) = d\tilde{S}^{(N+1)}_{\Delta t}(U,t)^T J d\tilde{S}^{(N+1)}_{\Delta t}(U,t) \tag{4.45}$$

and for future reference we note that $\mathcal{A}_{\Delta t}(U,0) = J$. By (4.44) and (3.9) we deduce that

$$\frac{\partial}{\partial \tau}\bigg|_{\tau=t} \mathcal{A}_{\Delta t}(U,\tau) = \Delta t^{r+N} dS(U,t)^T \Phi(S(U,t)) dS(U,t) + \widehat{r}(U,t) \tag{4.46}$$

where

$$|||\widehat{r}(U,t)||| \leq K(U,t)\Delta t^{2r+N} \tag{4.47}$$

for some $K = K(U,t)$ independent of Δt. Hence, for any $t \in [0,T]$ we have

$$\mathcal{A}_{\Delta t}(U,t) = J + \Delta t^{r+N} \int_0^t dS(U,s)^T \Phi(S(U,s)) dS(U,s)\, ds + \int_0^t \widehat{r}(U,s)\, ds. \tag{4.48}$$

If we assume that

$$\int_0^t dS(U,s)^T \Phi(S(U,s)) dS(U,s)\, ds = 0 \tag{4.49}$$

for all $t \in [0, T]$, then standard continuity arguments lead to the conclusion

$$dS(U,s)^T \Phi(S(U,s)) dS(U,s) = 0 \tag{4.50}$$

for all $s \in [0, T]$. By the invertibility of $dS(U, s)$, we then deduce that

$$\Phi(S(U,s)) = 0 \tag{4.51}$$

for all $s \in [0, T]$. However, since $S(U, s) \in D$ for all $s \in [0, T]$, we have a contradiction with (4.42), and so there must exist $\tau \in [0, T]$ (independent of Δt) and $\delta > 0$ such that

$$||| \int_0^\tau dS(U,s)^T \Phi(S(U,s)) dS(U,s)\, ds\, ||| \geq \delta. \tag{4.52}$$

Using the above result in (4.48) gives

$$|||\mathcal{A}_{\Delta t}(U,\tau) - J||| \geq \Delta t^{r+N}\delta - ||| \int_0^\tau \widehat{r}(U,s)\, ds\, |||, \tag{4.53}$$

and by (4.47) it follows that

$$|||\mathcal{A}_{\Delta t}(U,\tau) - J||| \geq \Delta t^{r+N}\delta/2, \tag{4.54}$$

for all $\Delta t \in [0, h]$, possibly by further reduction of h.

Now, using the identity

$$\begin{aligned} &d\tilde{S}_{\Delta t}^{(N+1)}(U,t)^T J d\tilde{S}_{\Delta t}^{(N+1)}(U,t) - d\bar{S}_{\Delta t}^n(U)^T J d\bar{S}_{\Delta t}^n(U) \\ &\quad = \tfrac{1}{2}\left(d\tilde{S}_{\Delta t}^{(N+1)}(U,t) + d\bar{S}_{\Delta t}^n(U)\right)^T J \left(d\tilde{S}_{\Delta t}^{(N+1)}(U,t) - d\bar{S}_{\Delta t}^n(U)\right) \\ &\qquad + \tfrac{1}{2}\left(d\tilde{S}_{\Delta t}^{(N+1)}(U,t) - d\bar{S}_{\Delta t}^n(U)\right)^T J \left(d\tilde{S}_{\Delta t}^{(N+1)}(U,t) + d\bar{S}_{\Delta t}^n(U)\right) \end{aligned} \tag{4.55}$$

we obtain the bound

$$\begin{aligned} &|||d\tilde{S}_{\Delta t}^{(N+1)}(U,\tau)^T J d\tilde{S}_{\Delta t}^{(N+1)}(U,\tau) - d\bar{S}_{\Delta t}^n(U)^T J d\bar{S}_{\Delta t}^n(U)||| \\ &\qquad \leq (C_2 + C_7) C_{17} |||d\tilde{S}_{\Delta t}^{(N+1)}(U,\tau) - d\bar{S}_{\Delta t}^n(U)||| \\ &\qquad \leq (C_2 + C_7) C_{17} C_8 \Delta t^{r+N+1} \end{aligned} \tag{4.56}$$

for all $\Delta t = \tau/n$ with $n \geq n^*$, where n^* is any positive integer such that $\tau/n^* \in [0, h]$. The above expression follows from the bounds in (2.9), Theorem 3.2 and (4.1), and the notation $C_{17} = |||J||| > 0$. By the definition of $\mathcal{A}_{\Delta t}$, and the fact that the numerical scheme generates a symplectic semigroup, we have

$$|||\mathcal{A}_{\Delta t}(U,\tau) - J||| \leq (C_2 + C_7) C_{17} C_8 \Delta t^{r+N+1}$$

for all $\Delta t = \tau/n$ with $n \geq n^*$. This contradicts (4.54). Hence $dq_{N+1}(u)$ must be infinitesimally symplectic for all $u \in \mathbb{R}^p$, which implies that the modified equation of index $N+1$ generates a symplectic semigroup. The case $N = 0$ holds since it gives the equation (2.1), and thus by induction the modified equation of index N generates a symplectic semigroup for any integer $N \geq 1$. By the results of Dragt & Finn [1976] we deduce that, for any integer $N \geq 1$, the modified equation is Hamiltonian as claimed. ■

4.5. *Volume Preservation and the Modified Semigroup.*

All equations (2.1) with f divergence-free have semigroups which preserve phase volume. Since Hamiltonian vector fields are divergence-free this result holds for all Hamiltonian problems; indeed, for dimension $p = 2$, it is equivalent to symplecticity of the semigroup. It is possible to construct numerical methods that automatically inherit the property of volume preservation; see, for example, Feng & Wang [1994], Feng & Shang [1995], Shang [1994], Quispel [1995] and Suris [1996] (and earlier references cited therein to the Soviet literature). We now give a result that applies to the modified equations of one-step volume-preserving methods; this result may also be proved by the techniques in Reich [1996].

Theorem 4.6. *If the system and method preserve volume, then so do all modified equations. More precisely, assume that for any compact set B we have*

$$\int_{S_t(B)} dv = \int_{\bar{S}^n_{\Delta t}(B)} dv = \int_B dv$$

for all $t = n\Delta t \in [0, T]$ and $\Delta t \in [0, h]$ where $h, T > 0$ are sufficiently small. Then, for any $N \geq 1$, we have $\nabla \cdot \tilde{f}^{(N)}_{\Delta t}(y) = 0$ for all $y \in \mathbb{R}^p$ and $\Delta t > 0$.

Proof. For induction assume the modified vector fields up to index N are divergence-free, that is

$$\nabla \cdot \tilde{f}^{(i)}_{\Delta t}(y) = 0, \qquad 0 \leq i \leq N,\ y \in \mathbb{R}^p,\ \Delta t > 0. \tag{4.57}$$

Since (2.1) preserves phase volume if and only if $\nabla \cdot f \equiv 0$, we note that the inductive hypothesis is true for $N = 0$. Now suppose the modified vector field of index $N + 1$ is not divergence-free for some $\Delta t > 0$. This implies that

$$\nabla \cdot q_{N+1}(u_0) \neq 0 \tag{4.58}$$

for some $u_0 \in \mathbb{R}^p$.

Since $\tilde{S}^{(N+1)}_{t,\Delta t}$ is close to the identity for t small and q_{N+1} is smooth, we deduce from (4.58) that there are numbers $\gamma, h, T > 0$ sufficiently small and a number $\delta > 0$ such that

$$|\nabla \cdot q_{N+1}(y)| \geq \delta, \quad \forall y \in \bigcup_{\substack{t\in[0,T]\\ \Delta t\in[0,h]}} \tilde{S}^{(N+1)}_{t,\Delta t}(B, t),$$

where B is the closed ball at u_0 of radius γ. Also, if we let

$$V^{(N+1)}(t) = \int_{\tilde{S}^{(N+1)}_{t,\Delta t}(B)} dv,$$

then

$$\frac{dV^{(N+1)}}{dt} = \int_{\tilde{S}^{(N+1)}_{t,\Delta t}(B)} \nabla \cdot \tilde{f}^{(N+1)}_{\Delta t}(v)\, dv = \int_{\tilde{S}^{(N+1)}_{t,\Delta t}(B)} \Delta t^{r+N} \nabla \cdot q_{N+1}(v)\, dv.$$

Assuming first that $\nabla \cdot q_{N+1}(u_0) > 0$, we get

$$\frac{dV^{(N+1)}}{dt} \geq \delta \Delta t^{r+N} V^{(N+1)}(t), \quad t \in [0,T],\ \Delta t \in [0,h].$$

This inequality leads to the bounds

$$V^{(N+1)}(t) \geq \exp(\delta \Delta t^{r+N} t) V^{(N+1)}(0) \geq (1 + \delta \Delta t^{r+N} t) V^{(N+1)}(0),$$

and hence

$$|V^{(N+1)}(t) - V^{(N+1)}(0)| \geq \delta \Delta t^{r+N} t V^{(N+1)}(0), \quad t \in [0,T],\ \Delta t \in [0,h]. \tag{4.59}$$

Alternatively, in the case where $\nabla \cdot q_{N+1}(u_0) < 0$, we have

$$\frac{dV^{(N+1)}}{dt} \leq -\delta \Delta t^{r+N} V^{(N+1)}(t), \quad t \in [0,T],\ \Delta t \in [0,h],$$

and

$$V^{(N+1)}(t) \leq \exp(-\delta \Delta t^{r+N} t) V^{(N+1)}(0) \leq (1 - \frac{\delta \Delta t^{r+N} t}{2}) V^{(N+1)}(0), \tag{4.60}$$

for all $t \in [0,T]$ and $\Delta t \in [0,h]$, after reducing T if necessary. Thus, regardless of the sign of $\nabla \cdot q_{N+1}(u_0)$, we deduce from (4.59) and (4.60) that

$$|V^{(N+1)}(t) - V^{(N+1)}(0)| \geq \frac{\delta \Delta t^{r+N} t V^{(N+1)}(0)}{2}, \quad t \in [0,T],\ \Delta t \in [0,h]. \tag{4.61}$$

To obtain a contradiction, let

$$V_{\Delta t}(n) = \int_{\bar{S}^n_{\Delta t}(B)} dv.$$

Then, since the numerical method conserves volume, we have $V_{\Delta t}(n) = V^{(N+1)}(0)$, and hence

$$|V^{(N+1)}(t) - V^{(N+1)}(0)| = |V^{(N+1)}(t) - V_{\Delta t}(n)| \tag{4.62}$$

for any $t = n\Delta t \in [0,T]$ with $\Delta t \in [0,h]$. Using the relations

$$V^{(N+1)}(t) = \int_{\tilde{S}^{(N+1)}_{t,\Delta t}(B)} dv = \int_B \det[d\tilde{S}^{(N+1)}_{t,\Delta t}(v)]\, dv$$

and

$$V_{\Delta t}(n) = \int_{\bar{S}^n_{\Delta t}(B)} dv = \int_B \det[d\bar{S}^n_{\Delta t}(v)]\, dv,$$

together with (4.62) and the derivative bounds in Theorem 3.2 , we find that there exists a constant C_{18} such that

$$|V^{(N+1)}(t) - V^{(N+1)}(0)| \leq C_{18} \Delta t^{r+N+1}$$

for any $t = n\Delta t \in [0,T]$ with $\Delta t \in [0,h]$. This contradicts (4.61) and thus we must have $\nabla \cdot \tilde{f}^{(N+1)}_{\Delta t}(y) = 0$ for all $y \in \mathbb{R}^p$ and $\Delta t > 0$. ■

4.6. *A Negative Example.*

The results presented in this article fit into a general framework that may be summarized as follows: *If the differential equations have a certain structural property and the numerical method has an analogous structural property for $\Delta t > 0$, then all modified equations share this property for $\Delta t > 0$.* However, the results proved are strongly tied to structural properties associated with certain subspaces of $\mathcal{V}(\mathbb{R}^p)$, the infinite-dimensional Lie algebra of smooth vector fields on $\mathbb{R}^p$, as outlined in the Introduction. Vector fields sharing a particular structural property do not necessarily form a subspace of $\mathcal{V}(\mathbb{R}^p)$, and it is natural to ask about the properties enjoyed by modified equations in this case. As the next example shows, modified equations do not generically inherit structural properties shared by the numerical method and underlying system.

Consider the scalar problem with $f(u) = -u^3$. This is an example of a gradient system with Lyapunov function $F(u) = u^4/4$. For this problem we have

$$\frac{d}{dt}F(S_t(u)) = -f(S_t(u))^2. \tag{4.63}$$

Hence, the Lyapunov function F decreases along every non-constant solution trajectory. It follows that

$$S_t(u) \to 0, \quad \text{as} \quad t \to \infty, \quad \forall u \in \mathbb{R}^p. \tag{4.64}$$

For the implicit Euler method $U_{n+1} = U_n + \Delta t f(U_{n+1})$ it is readily shown that, for $f(u) = -u^3$,

$$\frac{F(U_{n+1}) - F(U_n)}{\Delta t} \leq -\left(\frac{U_{n+1} - U_n}{\Delta t}\right)^2.$$

It follows that, independently of Δt,

$$\bar{S}^n_{\Delta t}(u) \to 0, \quad \text{as} \quad n \to \infty, \quad \forall u \in \mathbb{R}^p, \tag{4.65}$$

which is the discrete analogue of (4.64).

The first modified equation associated with the implicit Euler method on this problem has a vector field of the form

$$\tilde{f}^{(1)}_{\Delta t}(u) = -u^3 + \frac{3\Delta t}{2}u^5.$$

In this case,

$$\tilde{S}^{(1)}_{t,\Delta t}(u) \to 0, \quad \text{as} \quad t \to \infty, \quad \text{for} \quad |u| < \sqrt{\frac{2}{3\Delta t}},$$

and

$$|\tilde{S}^{(1)}_{t,\Delta t}(u)| \to \infty, \quad \text{as} \quad t \to \infty, \quad \text{for} \quad |u| > \sqrt{\frac{2}{3\Delta t}}.$$

Hence, the first modified equation does not have the property of global convergence to the origin, even though this property is shared by the original equation and the discretization.

5. References.

S.P. Auerbach & A. Friedman (1991) "Long-time Behaviour of Numerically Computed Orbits: Small and Intermediate Time-Step Analysis of One-Dimensional Systems", *J. Computational Physics,* **93**, 189–223.

G. Benettin & A. Giorgilli (1994) "On the Hamiltonian Interpolation of Near-to-the-Identity Symplectic Mappings with Application to Symplectic Integrators", *J. Stat. Phys.,* **74**, 1117–1143.

W.-J. Beyn (1991) "Numerical methods for dynamical systems", in *Advances in Numerical Analysis; Volume 1.* Clarendon Press, Oxford.

M.P. Calvo, A. Murua & J.M. Sanz-Serna (1994) "Modified Equations for ODEs", *Contemporary Mathematics* **172**, 63–74.

M.P. Calvo, A. Iserles & A. Zanna (1996) "Runge–Kutta methods for orthogonal and isospectral flows", Numerical Analysis Report NA08, DAMTP, University of Cambridge.

A.J. Dragt & J.M. Finn (1976) "Lie Series and Invariant Functions for Analytic Symplectic Maps", *J. Math. Phys.,* **17**, 2215–2227.

Feng Kang and Wang Dao-Liu (1994) "Dynamical systems and geometric construction of algorithms", in *Contemporary Mathematics 163,* editors Z. Shi and C. Yang.

Feng Kang and Shang Zai-Jiu (1995) "Volume-preserving difference schemes for source free dynamical systems", *Numer. Math.,* **71**, 451–463.

B. Fiedler and J. Scheurle (1996) "Discretization of Homoclinic orbits, rapid forcing and 'invisible' chaos", *Memoirs of the AMS*, Providence, RI.

G.H. Golub & C.F. Van Loan (1996) *Matrix Computations,* Third Ed., Johns Hopkins University Press.

D.F. Griffiths & J.M. Sanz-Serna (1986) "On the Scope of the Method of Modified Equations", *SIAM J. Sci. Stat. Comp.*, **7**, 994–1008.

E. Hairer (1994) "Backward Analysis of Numerical Integrators and Symplectic Methods", *Annals of Numerical Mathematics,* **1**, 107–132.

E. Hairer (1997) "Variable Time Step Integration with Symplectic Methods", Applied Numerical Mathematics 25, 219–227.

E. Hairer & Ch. Lubich (1997) "The Life-Span of Backward Error Analysis for Numerical Integrators", Numer. Math. 76, 441–462.

E. Hairer & D. Stoffer (1997) "Reversible Long-Term Integration with Variable Step-Sizes", *SIAM J. Sci. Comp.*, **18**, 257–269.

R.S. Mackay (1992) "Some Aspects of the Dynamics and Numerics of Hamiltonian Systems", *Proceedings of the IMA Conference on The Dynamics of Numerics and the Numerics of Dynamics, 1990*, editors D. Broomhead and A. Iserles, Cambridge University Press.

A.I. Neishtadt (1984) "The separation of motions in systems with rapidly rotating phase", *J. Appl. Math. Mech.*, **48**, 133–139.

G.R.W. Quispel (1995) "Volume preserving integrators", *Phys Lett.*, **206A**, 26–30.

G.W. Reddien (1995) "On the Stability of Numerical Methods of Hopf Points Using Backward Error Analysis", *Computing*, **55**, 163–180.

S. Reich (1993) "Numerical integration of the generalized Euler equations," Technical Report 93-20, Department of Computer Science, University of British Columbia.

S. Reich (1996) "Backward Error Analysis for Numerical Integrators", Preprint SC 96-21, Konrad-Zuse-Zentrum für Informationstechnik, Berlin.

J.M. Sanz-Serna (1992) "Symplectic Integrators for Hamiltonian Problems: An Overview", *Acta Numerica 1992*, 243–286.

J.M. Sanz-Serna & M.P. Calvo (1994) *Numerical Hamiltonian Problems,* Chapman and Hall.

J.M. Sanz-Serna & A. Murua (1997) *NSF-CBMS Lecture, Golden, Colorado, 1997* SIAM, to appear.

Shang Zai-Jiu (1994) "Construction of volume-preserving difference schemes for source free systems via generating functions", *J. Comp. Math.,* **12**, 265–272.

D. Stoffer (1995) "Variable Steps for Reversible Integration Methods", *Computing*, **55**, 1–22.

S.H. Strogatz (1994) *Nonlinear Dynamics and Chaos,* Addison Wesley, Reading MA.

A.M. Stuart & A.R. Humphries (1996) *Dynamical Systems and Numerical Analysis,* Cambridge University Press.

Y.B. Suris (1996) "Partioned Runge–Kutta methods as phase volume preserving integrators", *Phys. Lett.*, **220A**, 63–69.

R.F. Warming & B.J. Hyett (1974) "The Modified Equation Approach to the Stability and Accuracy Analysis of Finite Difference Methods", *J. Computational Physics*, **14**, 159–179.

H. Yoshida (1993) "Recent Progress in the Theory and Application of Symplectic Integrators," *Celestial Mechanics and Dynamic Astronomy,* **56**, 27–43.