On the Qualitative Properties of Modified Equations

O. Gonzalez [1] [2] and A.M. Stuart [3] [4]

Division of Applied Mechanics, Department of Mechanical Engineering, Stanford University, Stanford, CA 94305-4040, USA

An arbitrary consistent one-step approximation of an ordinary differential equation is studied. If the scheme is assumed to be accurate of $\mathcal{O}(\Delta t^r)$, then it may be shown to be $\mathcal{O}(\Delta t^{r+m})$ accurate as an approximation to a modified equation for any integer $m \geq 1$. A technique is introduced for proving that the modified equations inherit qualitative properties from the numerical method. For Hamiltonian problems the modified equation is shown to be Hamiltonian if the numerical method is symplectic, and for general problems the modified equation is shown to possess integrals shared between the numerical method and the underlying system. The technique for proving these results does not use the well-known theory for power series expansions of particular methods such as Runge-Kutta schemes, but instead uses a simple contradiction argument based on the approximation properties of the numerical method. Although the results presented are known, the method and generality of the proofs are new and may be of independent interest.

1. Introduction.

In this note 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.

The theoretical developments of this paper begin in Section 2 where 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 which 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. Standard error estimates for such methods show that the error over a finite time interval is

[1] Graduate fellow supported by the National Science Foundation

[2] Current Address: University of Maryland, Institute for Physical Science and Technology, CSS Bldg, College Park, MD 20742

[3] Work supported by the National Science Foundation under grant DMS-9201727 and by the Office of Naval Research under grant N00014-92-J-1876

[4] Program in Scientific Computing and Computational Mathematics

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 used to ensure conservation of invariants for Hamiltonian problems.

In Section 3 we discuss modified equations; in particular, for any 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. Note that the idea of modified equations is well-known and was first studied in detail in the paper Warming & Hyett [1974] within the context of partial differential equations. Therein the modified equation approach was found to be useful in interpreting the qualitative properties of errors introduced by numerical approximation; for example, numerical dissipation or dispersion for wave propagation problems can often be clearly understood by studying associated modified equations. For further results on the usefulness and applicability of the modified equation approach see Griffiths & Sanz-Serna [1986].

The main contribution of this paper 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 inherits a certain structural property from the underlying ordinary differential equation, then the family of associated modified equations also inherits this property. The specific structural properties that we consider are conservation of a scalar function for general problems, and conservation of the canonical symplectic two-form for Hamiltonian problems.

Results similar to those in Section 4 are already contained in the literature. However, the technique of proof that we employ is new and, since it is very straightforward to apply, may have some merit when compared to existing proofs which, although somewhat shorter and more elegant, require more sophisticated mathematical machinery.

A result, similar to our Theorem 4.1 concerning conservation of a scalar function, is proved in Reich [1996]. The result of Theorem 4.2, showing that a symplectic numerical method has a family of modified equations which are Hamiltonian, is known in a wide variety of cases – see Auerbach & Friedman [1991] and Yoshida [1993] for specific examples, and see Mackay [1992], Sanz-Serna [1992] and Sanz-Serna & Calvo [1994] for general discussions concerning the backward error interpretation of symplectic schemes. The first general result concerning symplectic numerical methods and their associated modified equations is due to Hairer [1994] (see also Hairer & Lubich [1995]), who proves that all symplectic partitioned Runge-Kutta methods have Hamiltonian modified equations. Despite the great generality of this result, the proof is tied in a fundamental way to the specific form of Runge-Kutta methods. In Benettin & Giorgilli [1994], it is proved that the modified equations of symplectic methods are Hamiltonian; their analysis is not restricted to Runge-Kutta methods either and, in addition, is considerably more elegant than the one we present here. However, it is tied in a very specific way to the use of Poisson brackets to elucidate the symplectic structure. In contrast, our proof is a simple example of a more general contradiction approach which will apply to a wide variety of other structure-preserving numerical methods, including the integral-conserving methods studied here.

In summary, although the results proved here are for the most part not new, the techniques used, and the generality of the framework used, may be of independent interest. We present full proofs only for selected results. Complete details may be found in Gonzalez & Stuart [1995].

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

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

is a solution to (2.1) with initial condition $u(0) = U$ and is 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 which depends smoothly on the parameter Δt. For any $u_0 \in \mathbb{R}^p$ we assume the numerical scheme generates an 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 $\bar{S} : \mathcal{B} \times [0,h] \to \mathbb{R}^p$ such that for any $U \in \mathcal{B}$ and $\Delta t \in [0,h]$ the sequence (U_n) generated by

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

satisfies (2.4) for all $n\Delta t \in [0, \mathcal{T}]$. Here $\bar{S}^n(U, \Delta t)$ denotes the n-fold composition of the map $\bar{S}_{\Delta t} : \mathcal{B} \to I\!\!R^p$.

Given any $u_0 \in I\!\!R^p$ 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

$$\frac{\partial^i}{\partial \tau^i}\bigg|_{\tau=0} \bar{S}(U, \tau) = \frac{\partial^i}{\partial \tau^i}\bigg|_{\tau=0} S(U, \tau), \qquad i = 1, \ldots, r \tag{2.6}$$

where $r \geq 1$ by consistency.

For any $n\Delta t \in [0, T]$ with $\Delta t \in [0, h]$ let $d\bar{S}^n_{\Delta t}(U) \in I\!\!R^{p\times p}$ denote the derivative of $\bar{S}^n_{\Delta t} : B \to I\!\!R^p$ at a point $U \in B$, and let $||\cdot||$ denote the standard Euclidean norm on $I\!\!R^p$. Then, by standard results from the numerical analysis of ordinary differential equations (see e.g. 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, for any $t = n\Delta t \in [0, T]$ with $\Delta t \in [0, h]$, the derivative of the mapping $\bar{S}^n_{\Delta t} : B \to I\!\!R^p$ 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* $\tilde{f}^{(N)}_{\Delta t} : I\!\!R^p \to I\!\!R^p$ 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 $q_i : I\!\!R^p \to I\!\!R^p$ $(i = 1, \ldots, N)$. With the appropriate choice of the functions q_i $(i = 1, \ldots, N)$ the numerical scheme (2.4) is an order $r+N$ approximation to (3.1) as we now show.

For any $v_0 \in \mathbb{R}^p$ 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

$$v_{\Delta t}(t) = \tilde{S}^{(N)}(V,t,\Delta t) = \tilde{S}^{(N)}_{\Delta t}(V,t) = \tilde{S}^{(N)}_{t,\Delta t}(V) \tag{3.3}$$

is a solution to (3.1) with initial condition $v_{\Delta t}(0) = V$ and 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 $t \in [0,\tilde{T}]$.

Consider the local evolution semigroups at v_0 generated by (2.1) and (2.4), and without loss of generality assume $\tilde{B} = B$, $\tilde{T} = T$, and $\tilde{h} = h$. For any $U \in B$, and for $t \in [0,T]$ and $\Delta t \in [0,h]$ sufficiently small, we may expand $S(U,t)$, $\bar{S}(U,\Delta t)$ and $\tilde{S}^{(N)}(U,t,\Delta t)$ in Taylor series about $t = 0$ and $\Delta t = 0$ as

$$S(U,t) = \sum_{j=0}^{k} \frac{t^j}{j!}\alpha_j(U) + \mathcal{O}\left(t^{k+1}\right) \tag{3.4}$$

$$\bar{S}(U,\Delta t) = \sum_{j=0}^{k} \frac{\Delta t^j}{j!}\beta_j(U) + \mathcal{O}\big(\Delta t^{k+1}\big) \tag{3.5}$$

$$\tilde{S}^{(N)}(U,t,\Delta t) = \sum_{j=0}^{k}\sum_{\ell=0}^{j} \binom{j}{\ell} \frac{t^{j-\ell}\Delta t^{\ell}}{j!}\tilde{\alpha}_{j\ell}(U) + \mathcal{O}\left((t+\Delta t)^{k+1}\right) \tag{3.6}$$

where $k \geq (r+N)$ is an integer and the coefficients are defined as

$$\alpha_j(U) = \frac{\partial^j}{\partial\tau^j}\Big|_{\tau=0} S(U,\tau) \tag{3.7}$$

$$\beta_j(U) = \frac{\partial^j}{\partial\tau^j}\Big|_{\tau=0} \bar{S}(U,\tau) \tag{3.8}$$

$$\tilde{\alpha}_{j\ell}(U) = \frac{\partial^j}{\partial\tau^{j-\ell}\partial s^{\ell}}\Big|_{\tau=0,s=0} \tilde{S}^{(N)}(U,\tau,s). \tag{3.9}$$

By definition of the semigroups we have

$$\alpha_0(U) = \beta_0(U) = \tilde{\alpha}_{00}(U) = U \tag{3.10}$$

and, since $\tilde{S}^{(N)}(U,0,\Delta t) = U$ for all $\Delta t \in [0,h]$, we have

$$\tilde{\alpha}_{jj}(U) = 0, \quad 1 \leq j \leq k. \tag{3.11}$$

Now, since (2.4) is an order r approximation to (2.1), we have

$$\alpha_j(U) = \beta_j(U), \quad 1 \leq j \leq r \tag{3.12}$$

for all $U \in B$. Furthermore, from (3.1), (3.2) and its relation with (2.1) we deduce that

$$\tilde{\alpha}_{j0}(U) = \alpha_j(U), \quad 1 \leq j \leq k \tag{3.13}$$

and that

$$\tilde{\alpha}_{j\ell}(U) = 0, \quad 1 \leq j \leq k, \quad 1 \leq \ell < \min\{r, j+1\}. \tag{3.14}$$

Hence the coefficients $\tilde{\alpha}_{j\ell}(U)$ for $0 \leq j \leq r$ and $0 \leq \ell \leq j$ are fully determined by properties of the underlying evolution semigroups. Our task now is to examine the remaining coefficients $\tilde{\alpha}_{j\ell}(U)$ for $r < j \leq k$ and $r \leq \ell \leq j$, and choose the functions q_i such that (2.4) is an order $r+N$ approximation to (3.1). From (3.1) and the definition of its local evolution semigroup we deduce that the functions q_i appear in the coefficients $\tilde{\alpha}_{j\ell}$ for $r+1 \leq j \leq r+N$ and $\ell = j-1$; in particular,

$$q_i(U) = \tilde{\alpha}_{(i+r)\,(i+r-1)}(U)/(i+r-1)!, \quad i = 1, \ldots, N. \tag{3.15}$$

This follows from (3.9) using the fact that

$$q_i(U) = \frac{1}{(i+r-1)!} \frac{\partial^{i+r-1}}{\partial s^{i+r-1}}\bigg|_{s=0} \tilde{f}_s^{(N)}(U) \tag{3.16}$$

and

$$\tilde{f}_s^{(N)}(U) = \frac{\partial}{\partial \tau}\bigg|_{\tau=0} \tilde{S}^{(N)}(U, \tau, s). \tag{3.17}$$

We now determine the functions q_i such that (2.4) approximates (3.1) to order $r+N$. Note that this order is achieved if over one time step, i.e. $t = \Delta t$, the Taylor expansions of $\tilde{S}^{(N)}(U, \Delta t, \Delta t)$ and $\bar{S}(U, \Delta t)$ agree through order $r+N$; that is, if

$$\sum_{\ell=0}^{j} \binom{j}{\ell} \tilde{\alpha}_{j\ell}(U) = \beta_j(U), \quad 0 \leq j \leq r+N. \tag{3.18}$$

In view of (3.10) through (3.14) we have this equality for $0 \leq j \leq r$, and we now choose the functions q_i $(i = 1, \ldots, N)$ so that this equality holds for $r+1 \leq j \leq r+N$. Replacing j in (3.18) by $i+r$ we get

$$\sum_{\ell=0}^{i+r} \binom{i+r}{\ell} \tilde{\alpha}_{i+r\,\ell}(U) = \beta_{i+r}(U), \quad 1 \leq i \leq N \tag{3.19}$$

and, since $i+r \geq 2$ for $1 \leq i \leq N$, we may write

$$\sum_{\ell=0}^{i+r-2} \binom{i+r}{\ell} \tilde{\alpha}_{i+r\,\ell}(U) + \binom{i+r}{i+r-1} \tilde{\alpha}_{(i+r)\,(i+r-1)}(U) + \binom{i+r}{i+r} \tilde{\alpha}_{(i+r)\,(i+r)}(U) = \beta_{i+r}(U), \quad 1 \leq i \leq N. \tag{3.20}$$

In view of (3.11) and (3.15) we obtain

$$q_i(U) = \frac{1}{(i+r)!} \left(\beta_{i+r}(U) - \sum_{\ell=0}^{i+r-2} \binom{i+r}{\ell} \tilde{\alpha}_{i+r\,\ell}(U) \right)$$

and hence, by (3.13) and (3.14), we find that

$$q_i(U) = \frac{1}{(i+r)!} \left(\beta_{i+r}(U) - \alpha_{i+r}(U) - \sum_{\ell=r}^{i+r-2} \binom{i+r}{\ell} \tilde{\alpha}_{i+r\,\ell}(U) \right). \tag{3.21}$$

Here we use the convention that a sum from $\ell = a$ to $\ell = b$ with $b < a$ is zero.

By direct calculation one can show that the coefficients $\{\tilde{\alpha}_{(i+r)\,\ell}\}_{\ell=r}^{i+r-2}$ only depend upon the functions $\{q_j\}_{j=1}^{i-1}$, and thus (3.21) provides a recursive definition for the functions q_i (see Gonzalez & Stuart [1995] for details). Also, since the evolution semigroups are by assumption C^∞ smooth, we note that the functions q_i are C^∞ smooth.

Remarks

1) The foregoing developments provide only *local* definitions of the functions q_i in a ball B about an arbitrary point $v_0 \in \mathbb{R}^p$. However, since the above constructions can be performed at any point, we can use these local definitions to construct mappings on all of $\mathbb{R}^p$. Note that these global mappings are well-defined since the local ones are equal on the intersection of their domains; in particular, this follows from the fact that the underlying evolution semigroups are equal on the intersection of their domains, i.e. uniqueness of solutions to the systems in (2.1), (2.4) and (3.1).

2) Arguments similar to those just given may also be found in Section 3 of Benettin and Giorgilli [1994]. ■

Since the semigroup for the modified equation (3.1) over a time interval of length Δt agrees with that of the numerical method over one time step to order $\mathcal{O}(\Delta t^{r+N})$, it follows by standard techniques (see e.g. Stuart & Humphries [1996, Theorem 6.2.1]) that the solution operator and its derivative with respect to initial data converge with order $r+N$. Furthermore, because the modified vector field is

$\mathcal{O}(\Delta t^r)$ close to the original vector field in the C^1 sense, it follows that any local evolution semigroup of the modified equation is $\mathcal{O}(\Delta t^r)$ close to the corresponding local evolution semigroup of the original equation in the C^1 sense. These observations are combined with the foregoing developments in the following

Theorem 3.1. *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, for each $U \in B$, there is a constant $C_8 = C_8(T, N, U) > 0$ such that*

$$|||d\tilde{S}^{(N)}_{t,\Delta t}(U) - d\bar{S}^n_{\Delta t}(U)||| + ||\tilde{S}^{(N)}_{t,\Delta t}(U) - \bar{S}^n_{\Delta t}(U)|| \leq C_8 \Delta t^{r+N} \tag{3.22}$$

and

$$|||d\tilde{S}^{(N)}_{t,\Delta t}(U) - dS_t(U)||| + ||\tilde{S}^{(N)}_{t,\Delta t}(U) - S_t(U)|| \leq C_8 \Delta t^{r} \tag{3.23}$$

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

4. Qualitative Properties of the Modified Equations.

In this section we will employ an induction on N to prove various properties of the modified equation (3.1). In view of Theorem 3.1 we see that, given any $u_0 \in \mathbb{R}^p$, the ball B at u_0 and the numbers $h, T > 0$ 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 finite for every fixed integer $N \geq 1$.

By virtue of Theorem 3.1 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 order $N + 1$. Thus, for $1 \leq m \leq N + 1$,

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

For simplicity we define the modified equation of order $N = 0$ to be the original unperturbed equation (2.1) itself. Thus

$$\tilde{S}^{(0)}_{\Delta t}(u, t) = S(u, t) \quad \text{and} \quad \tilde{f}^{(0)}_{\Delta t}(u) = f(u), \qquad \forall u \in \mathbb{R}^p. \tag{4.2}$$

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 and $\bar{S}$ 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 $n\Delta t \in [0, T]$. Given a modified equation for (2.1) and (2.4) the question arises as to whether or not $\mathcal{F}$ is an integral for the modified system. In this section we show that $\mathcal{F}$ is indeed an integral for the associated modified equation of index N for any integer $N \geq 1$. Precisely, we have the following

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* $N \geq 1$. *In particular, 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, with local semigroup denoted by $\tilde{S}^{(N)}$, 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 0 is the original equation (2.1) itself.

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

$$\mathcal{F}(\tilde{S}^{(N)}_{\Delta t}(U, t)) = \mathcal{F}(U), \tag{4.3}$$

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

$$\nabla \mathcal{F}(u) \cdot \tilde{f}^{(N)}_{\Delta t}(u) = 0, \qquad \forall u \in \mathcal{I}m\big(\tilde{S}^{(N)}_{\Delta t}\big) \tag{4.4}$$

where

$$\mathcal{I}m\big(\tilde{S}^{(N)}_{\Delta t}\big) = \{u \in \mathbb{R}^p \mid u = \tilde{S}^{(N)}_{\Delta t}(U, t), \quad U \in B, \quad t \in [0, T]\}. \tag{4.5}$$

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

$$\frac{dv}{dt} = \tilde{f}^{(N+1)}_{\Delta t}(v) = \tilde{f}^{(N)}_{\Delta t}(v) + \Delta t^{r+N} q_{_{N+1}}(v). \tag{4.6}$$

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

$$\nabla \mathcal{F}(u_0) \cdot q_{_{N+1}}(u_0) \neq 0. \tag{4.7}$$

Otherwise, $\nabla\mathcal{F}(u)\cdot\tilde{f}_{\Delta t}^{(N+1)}(u)=0$ 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 $C_9(u_0)>0$; 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

$$\nabla\mathcal{F}(U)\cdot q_{N+1}(U)\geq C_9>0, \qquad \forall U\in D. \tag{4.8}$$

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 $\tilde{S}_{\Delta t}^{(N+1)}(U,t)\in D$ for all $t\in[0,T]$ and $U_n=\bar{S}_{\Delta t}^{n}(U)\in D$ for all $n\Delta t\in[0,T]$. Then, for any $\Delta t\in[0,h]$ and $t\in[0,T]$, we have by (4.4) and (4.6)

$$\begin{aligned}\frac{\partial}{\partial\tau}\Big|_{\tau=t}\mathcal{F}(\tilde{S}_{\Delta t}^{(N+1)}(U,\tau)) &= \nabla\mathcal{F}(\tilde{S}_{\Delta t}^{(N+1)}(U,t))\cdot\tilde{f}_{\Delta t}^{(N+1)}(\tilde{S}_{\Delta t}^{(N+1)}(U,t))\\ &= \Delta t^{r+N}\nabla\mathcal{F}(\tilde{S}_{\Delta t}^{(N+1)}(U,t))\cdot q_{N+1}(\tilde{S}_{\Delta t}^{(N+1)}(U,t))\\ &\geq C_9\Delta t^{r+N},\end{aligned} \tag{4.9}$$

which implies

$$|\mathcal{F}(\tilde{S}_{\Delta t}^{(N+1)}(U,T))-\mathcal{F}(U)|\geq C_9T\Delta t^{r+N}, \tag{4.10}$$

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

$$|\mathcal{F}(U)-\mathcal{F}(V)|\leq C_{10}||U-V||, \quad \forall U,V\in D. \tag{4.11}$$

Furthermore, in view of (3.22), the modified equation of index $N+1$ and the numerical scheme (2.4) have solutions satisfying

$$||\tilde{S}_{\Delta t}^{(N+1)}(U,T)-\bar{S}_{\Delta t}^{n}(U)||\leq C_8\Delta t^{r+N+1}, \tag{4.12}$$

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 $\bar{S}$ we use (4.11) and (4.12) to write

$$\begin{aligned}|\mathcal{F}(\tilde{S}_{\Delta t}^{(N+1)}(U,T))-\mathcal{F}(U)| &= |\mathcal{F}(\tilde{S}_{\Delta t}^{(N+1)}(U,T))-\mathcal{F}(\bar{S}_{\Delta t}^{n}(U))|\\ &\leq C_{10}||\tilde{S}_{\Delta t}^{(N+1)}(U,T)-\bar{S}_{\Delta t}^{n}(U)||\\ &\leq C_8C_{10}\Delta t^{r+N+1},\end{aligned} \tag{4.13}$$

for all $\Delta t=T/n$ and $n\geq n^*$. This yields a contradiction, since for $\Delta t<TC_9/C_8C_{10}$ both (4.10) and (4.13) cannot hold. Hence $\mathcal{F}$ must be an integral for the modified equation of index $N+1$. Since $N=0$ gives the original equation (2.1), the result follows by induction. ■

Remark A more general result which includes the above as a special case has recently appeared in Reich [1996].

4.2. *Symplecticity of the Modified Semigroup.*

Suppose now that the underlying system (2.1) is Hamiltonian on $\mathbb{R}^p$ with the canonical symplectic structure (assuming p is even, say $p = 2m$) and local semigroup denoted by S. Furthermore, assume the numerical approximation scheme (2.4) generates a semigroup $\bar{S}$ which is symplectic. Given an associated modified equation for (2.1) and (2.4) the question now arises as to whether or not the semigroup for the modified system defines a symplectic map. In this section we show that the semigroup for the associated modified equation of index N is indeed symplectic for any integer $N \geq 1$. Before doing this we introduce some notation.

For simplicity, we assume that the vector field $f : \mathbb{R}^p \to \mathbb{R}^p$ is Hamiltonian with respect to the canonical symplectic structure, that is

$$f(u) = J\nabla H(u) \tag{4.14}$$

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

$$J = \begin{pmatrix} O_m & I_m \\ -I_m & O_m \end{pmatrix}. \tag{4.15}$$

Here O_m and I_m denote the zero and identity matrices in $\mathbb{R}^{m\times m}$, respectively.

Theorem 4.2. *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), (3.2) *has the form*

$$\frac{dv}{dt} = J\nabla[\, H(v) + \Delta t^r Q^{(N)}(v; \Delta t)\,]. \tag{4.16}$$

Proof. The result follows by a contradiction argument similar to that used in the proof of Theorem 4.1. The main idea is to show that $\tilde{f}_{\Delta t}^{(N+1)}$ is *infinitesimally symplectic* given that $\tilde{f}_{\Delta t}^{(N)}$ is. One then uses induction and the results of Dragt & Finn [1976] to establish the result. See Gonzalez & Stuart [1995] for details. ∎

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.

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

O. Gonzalez & A.M. Stuart (1995) "Remarks on the Qualitative Properties of Modified Equations," SC/CM Technical Report, Scientific Computing and Computational Mathematics Program, Stanford University.

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.

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.

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

E. Hairer and Ch. Lubich (1995) "The Life-Span of Backward Error Analysis for Numerical Integrators," preprint.

S. Reich (1996) "Backward Error Analysis for Numerical Integrators," preprint.

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.

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

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.