UNIVERSITÉ DE GENÈVE
Section de mathématiques

FACULTÉ DES SCIENCES
Prof. Ernst HAIRER

Symmetric Linear Multistep Methods for Hamiltonian Systems on Manifolds

Thèse

présentée à la Faculté des Sciences de l'Université de Genève

pour obtenir les grades de

Docteur dès Sciences de l'Université de Genève, mention Mathématiques

par

Paola CONSOLE

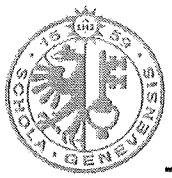
de

Taranto (Italie)

Thèse N° 4591

Atelier d'impression ReproMail de l'Université de Genève

2013



UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES

Doctorat ès sciences
Mention mathématiques

Thèse de *Madame Paola CONSOLE*

intitulée :

" Symmetric Linear Multistep Methods for Hamiltonian Systems on Manifolds "

La Faculté des sciences, sur le préavis de Messieurs E. HAIRER, professeur ordinaire et directeur de thèse (Section de mathématiques), M. GANDER, professeur ordinaire (Section de mathématiques), C. LUBICH, professeur (Institut für Mathematik, Universität Tübingen, Deutschland) et G. SÖDERLIND, professeur (Centre for Mathematical Sciences, Lund University, Sweden), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 4 septembre 2013

Thèse - 4591 -

Le Doyen, Jean-Marc TRISCONE

N.B.- La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

a Papà

a Francesco

Remerciements

Je voudrais tout d'abord remercier Ernst Hairer pour sa disponibilité, sa patience et son aide, et pour la confiance qu'il m'a montrée durant ces quatre années ; ce fut un grand honneur d'être une de ses élèves.

Je suis trés reconnaissante à Christian Lubich et Gustaf Söderlind d'avoir accepté d'être membres du jury de soutenance, et pour avoir montré lors de chacune de nos rencontres tant de gentillesse et d'intérêt pour mon travail de thèse.

Je tiens aussi à remercier toutes les personnes que j'ai rencontrées au cours de ces années à Genève ; un grand merci à Martin Gander et à l'ensemble du groupe d'analyse numérique : Soheil, Jérôme, Heiko, Hui, Bankim, Erwin. Une mention spéciale à Félix, pour son soutien constant et son aide (pour la thèse et non), et à Christophe, avec qui j'ai eu le plaisir de partager de nombreux moments au sein de la section de mathématiques et de nombreux voyages.

Merci aussi à Nicola Guglielmi et Victorita Dolean pour la gentillesse démontrée durant ces années, et à Raffaele D'Ambrosio pour son amitié et son soutien mathématique et moral lors des derniers, fondamentaux, mois.

Ces années n'auraient pas été les mêmes sans mes collègues doctorants et post-doctorants, passés et présents, de la Section de mathématiques ; merci à tous pour le temps que nous avons passé ensemble, malgré mon amour pour la célèbre cantine d'Uni-Mail. En particulier, merci à Aglaia pour tous les efforts faits pour être là lors de la soutenance et pour ses recettes (!!), merci à Anita pour avoir souffert du froid avec moi, et merci à Christian (et ce n'est pas le moins important) de m'avoir appris comment faire un bon mojito : aux deux merci pour l'aide avec la thèse et à vous tous un grand merci pour les merveilleux moments que nous avons passés ensemble. Un grand merci à Pierre- Alain Cherix, et une mention spéciale à Shaula, pour avoir été ma première vraie amie à Genève et pour m'avoir beaucoup aidée lors des premiers mois (et pas seulement).

Un grand merci aux secrétaires pour avoir supporté mon stress de la bureaucratie durant les derniers mois, et aux bibliothécaires pour leur sourire, leur convivialité et leur professionnalisme.

Merci à Valentina, Daniela, Graziana et Rossana pour la fraîcheur qu'elles ont apporté à mes années genevoises.

Une pensée ne peut pas ne pas aller aux trop lointains, mais très proches amis de Lecce, donc un grand merci à Fabio, Chiara, Laura, Alfredo, Michele, Manuela, Mary, Alessandro, Valentina, Rosy, Adele et Carlo ; une centaine de pages ne suffirait pas à en énumérer les raisons, mais vous les savez toutes. Un grand merci à Diego et Ilaria (et Newton), qui ont été durant ces dernières années mon annexe italienne en Suisse.

Un grand merci à Anna Maria Cherubini, pour avoir cru en moi même quand je ne l'ai pas fait, et pour ses précieux conseils.

Un grand merci à Sara pour avoir été mon amie depuis si longtemps.

Un gros bisou à ma famille ; merci à Beatrice pour avoir fait autant de kilomètres pour assister à ma soutenance.

Un merci particulier et un gros bisou à ma mère qui m'a toujours soutenue dans tous mes choix.

Dulcis in fundo, un bisou et un grand merci à Francesco : merci pour chaque jour passé proches ou lointains, et pour tout ce que tu es et que tu fais.

Contents

Chapter 0

Introduction and main results

The aim of the work described in this thesis is the study of symmetric linear multistep methods applied to Hamiltonian systems.

We show that this class of methods can have good properties of near preservation of the energy and momenta for long-time integrations of Hamiltonian systems; furthermore multistep methods are easy to construct, to program, and they can reach arbitrarily high order. All the analysis presented in this thesis follows the line of the study of symmetric linear multistep methods applied to second order Hamiltonian equations made in [HL04].

The thesis is divided into two main parts.

In the first part (Chapter 1 and 2) we study partitioned linear multistep methods applied to first order Hamiltonian equations: we focus mainly on the application of this class of methods to separable Hamiltonians. In this study it is shown how the use of symmetric partitioned multistep method can lead to near preservation of energy for a specific class of separable Hamiltonians.

In the second part (Chapter 3, 4 and 5) we study symmetric linear multistep methods applied to second order constrained Hamiltonian systems. In Chapter 3 and 4 we focus on the theoretical analysis of the excellent behaviour that this class of methods presents on this kind of problems; the analysis is supplemented with practical considerations on the construction of these methods, and with some numerical experiments. In Chapter 5 we study the optimization of the implementation of this class of methods.

Chapter 1 We investigate the application of partitioned linear multistep methods applied to Hamiltonian systems.

This theoretical study is based on backward error analysis and on the technique of modulated Fourier expansions, and it is mainly focused on separable Hamiltonians. For the smooth solution, we can construct modified equations but it is impossible to construct a first integral that is close to the Hamiltonian; we show that it is possible to improve the behavior of the smooth solution imposing some order conditions depending on the coefficients of the method. Next we present the analysis for the parasitic components for separable Hamiltonian systems, and we show that on the time intervals where they stay small and bounded, the numerical solution behaves like that of a symmetric one step method.

Chapter 2 Some complements to the analysis made in Chapter 1 are presented.

We show the optimization of the stability of symmetric partitioned multistep methods which satisfy the additional order condition described in Chapter 1. This work is made for the classes of methods of order 4 and 6, and it is supplemented with some numerical

experiments made on some separable and non-separable Hamiltonian systems, in which also the behavior of the parasitic components is shown.

Chapter 3 We present the theoretical study of symmetric linear multistep methods applied to second order constrained Hamiltonian systems.

For this class of methods it is possible to extend the techniques of [HL04], by constructing the modified equations and then the modified Hamiltonian. The analysis is completed by showing, under non-restrictive hypotesis, the boundedness of the parasitic components, which explains the excellent behaviour reported in the numerical experiments that supplement the theoretical analysis. The study is completed with the presentation of the construction of this class of methods.

Chapter 4 We present some complements to the analysis made in Chapter 3.

We study the interval of periodicity and the error constant of the classes of methods of order 4, 6 and 8 presented in Chapter 3, and this study is completed with some figures representing these quantities.

Chapter 5 The aim of this chapter is the optimization of the implementation of the methods studied in Chapter 3.

It is an important issue since, because of the high order of the methods, it is very easy to reach discretization errors of the size of the machine precision. If this happens the round-off error becomes the dominating source of error, and so it has to be optimized in order to avoid its linear growth due to deterministic errors. We present the techniques used for this purpose, and we show with some figures the effect that they have on the propagation of the round-off error.

0.1 Hamiltonian Systems

In this Section we give an overview on the Hamiltonian formalism, developed by Hamilton (1834) for modeling physical systems: this way of representing the equation of motions reveals some elegant intrinsic properties of some dynamical systems.

Hamiltonian equations are often used in astronomy for modeling planetary motions, in molecular dynamics or to represent the motions of rigid bodies.

0.1.1 Some notions on Lagrangian mechanics

Lagrange's formalism is the base for the construction of Hamiltonian formalism, that will be described in the next section.
Consider a mechanical system of d degrees of freedom: let $(q_1, \dots, q_d)$ be the coordinates representing the positions (the so-called *generalized coordinates*), and $(\dot{q}_1, \dots, \dot{q}_d)$ the respective velocities. Using these coordinates we write the functions

$$\begin{aligned} T &= T(q, \dot{q}) \\ U &= U(q), \end{aligned}$$

which represent the kinetic and the potential energy of the mechanical system respectively. If we denote by

$$L = T - U$$

the *Lagrangian* of the system, then it is possible to obtain the following *Lagrange equation*

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) = \frac{\partial L}{\partial q}\,: \tag{0.1}$$

this is a second order system of d equations, whose solution describes the motion of the mechanical system.

Example 0.1.1 (Pendulum)**.** Using Lagrange equations we describe the motion of a mathematical pendulum with mass $m = 1$ and length $l = 1$.

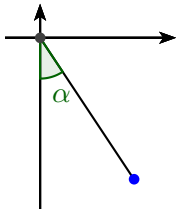


Figure 1 – Simple Pendulum.

We use as generalized coordinate $q = \alpha$, where α is the angle shown in Figure 1. Denoting the gravity with g, the kinetic and potential energies are given by $T = \frac{\dot{q}^2}{2}$ and $U = -g\cos q$. The Lagrangian is thus given by $L = \frac{\dot{q}^2}{2} + g\cos q$, and so the Lagrange equation is $\ddot{q} + g\sin q = 0$.

0.1.2 Hamiltonian equations

In 1834 Hamilton introduced a new formalism which simplified the structure of the Lagrange equations.

We introduce a new set of variables called *momenta*, which are defined as

$$p_k = \frac{\partial L}{\partial \dot{q}_k}(q, \dot{q}) \quad \text{for } k = 1, \ldots, d, \tag{0.2}$$

and define the *Hamiltonian function*

$$H(p, q) := p^{\mathsf{T}}\dot{q} - L(q, \dot{q}). \tag{0.3}$$

$\dot{q}$ is obtained as a function of p and q by using (0.2) if it is invertible (that locally means $\det \frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j} \neq 0$). This transform is called *Legendre transform.*

One can prove that Lagrange equations are equivalent to the system of equations

$$\begin{cases} \dot{p} &= -\frac{\partial H}{\partial q}(p, q) \\ \dot{q} &= \frac{\partial H}{\partial p}(p, q) \end{cases}\,: \tag{0.4}$$

in this way we have a first order system of $2d$ equations instead of a second order system of d equations. The equations (0.4) are called *Hamiltonian* or *canonical equations.*

This system of equations can be written in matrix form as

$$\dot{y} = J\nabla_y H(y),$$

where $y = (q\ ,\ p)^{\mathsf{T}}$ and J is the so-called *canonical structure matrix* that is defined as

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

Example 0.1.2 (Pendulum as an Hamiltonian system)**.** Consider the pendulum from Example 0.1.1: for (0.2) we obtain $p = \dot{q}$, so the Hamiltonian is $H(p, q) = \frac{p^2}{2} - g\cos q$, and its corresponding canonical equations are

$$\dot{p} = -g\sin q,\ \ \dot{q} = p.$$

Let us prove that the Hamiltonian remains constant along the solutions of (0.4), i.e. that it is a *first integral* of the motion. In fact, by defining $H_p = \frac{\partial}{\partial p}H(p,q)$ and $H_q = \frac{\partial}{\partial q}H(p,q)$, it results

$$\frac{d}{dt}H(p,q) = H_p^\mathsf{T}\dot{p} + H_q^\mathsf{T}\dot{q} = -H_p^\mathsf{T}H_q + H_p^\mathsf{T}H_q = 0.$$

If a Hamiltonian has the form $H(p,q) = T(p) + U(q)$ it is called *separable*.

Another important concept is the *flow* of a Hamiltonian system (it can be defined for every system of differential equations).

Definition 0.1.3. The *flow* of a Hamiltonian system is a map $\varphi : U \to \mathbb{R}^{2d}$ (U is an open set of $\mathbb{R}^{2d}$), which associates to any $y_0 = (q_0, p_0) \in U$ the evaluation at the time t of the solution of the system (0.4) with the initial values $(q(0), p(0)) = y_0$, i.e.

$$\varphi_t(p_0, q_0) = (p(t, p_0, q_0), q(t, p_0, q_0)) .$$

It is possible to show that the Hamiltonian flow satisfies a remarkable property, that is a characteristic property of Hamiltonian systems: we will present this property following the line of [HLW06].

Definition 0.1.4. A differentiable map $g : U \to \mathbb{R}^{2d}$ (with U open set of $\mathbb{R}^{2d}$) is *symplectic*, if its Jacobian matrix $g'(p,q)$ is everywhere symplectic, that is

$$g'(p,q)^\mathsf{T} J g'(p,q) = J \tag{0.5}$$

The following remarkable results for the Hamiltonian flow are proved in [HLW06].

Theorem 0.1.5 (Poincaré, 1899)**.** *If $H(p,q)$ is twice continuously differentiable in $U \subset \mathbb{R}^{2d}$, then for every fixed t the flow φ_t is a symplectic transformation wherever it is defined.*

The other implication is true only locally:

Theorem 0.1.6. *Let $f : U \to \mathbb{R}^{2d}$ be continuously differentiable. Then $\dot{y} = f(y)$ is locally Hamiltonian if and only if its flow $\varphi_t(y)$ is symplectic for all $y \in U$ and for all sufficiently small t.*

Another property of symplectic transformations is that they preserve the Hamiltonian character of the equations. It is in fact possible to prove the following result (shown in [HLW06])

Theorem 0.1.7. *Let $\psi : U \to V$ be a change of coordinates such that ψ and ψ^{-1} are continuously differentiable functions. If ψ is symplectic, the Hamiltonian system $\dot{y} = J^{-1}\nabla H(y)$ becomes in the new variables $z = \psi(y)$*

$$\dot{z} = J^{-1}\nabla K(z) \quad \textit{with} \quad K(z) = H(y). \tag{0.6}$$

Conversely, if ψ transform every Hamiltonian system to another Hamiltonian system via (0.6), then ψ is symplectic.

An important result concerning the Hamiltonian systems is *Noether's theorem*. It connects the existence of symmetries in the Lagrangian to the existence of first integrals of motion; the theorem is proved in [HLW06].

Theorem 0.1.8 (Noether's theorem)**.** *Consider a system with Hamiltonian $H(p,q)$ and Lagrangian $L(q,\dot{q})$. Suppose $\{g_s : s \in \mathbb{R}\}$ is a one-parameter group of transformations ($g_s \circ g_r = g_{s+r}$) which leaves the Lagrangian invariant, i. e.*

$$L(g_s(q), g_s'(q)\dot{q}) = L(q,\dot{q}) \quad \text{for all } s \text{ and all } (q,\dot{q}).$$

Let $a(q) = (d/ds)|_{s=0} g_s(q)$ be defined as the vector field with flow $g_s(q)$. Then

$$I(p,q) = p^{\mathsf{T}} a(q)$$

is a first integral of the Hamiltonian system.

0.2 Linear symmetric multistep methods for Hamiltonian systems

Linear multistep methods constitute one of the two big classes of numerical integrators for ordinary differential equations: they have the advantage to be easy to construct and to implement.

The classical theory of linear multistep methods has been developed by Dahlquist. The first results on the application of linear multistep methods to Hamiltonian systems is due to Quinlan and Tremaine in [QT90], where they show the excellent behavior obtained by applying symmetric linear multistep methods to the outer solar system. A complete explanation of this excellent behavior is provided in [HL04].

In this Section we give the basic definition and properties of linear multistep methods, and we explain briefly the techniques and results described in [HL04].

0.2.1 Linear multistep methods for first order equations

We consider a first order system of differential equations $\dot{y} = f(y)$: a linear multistep method is defined by

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(y_{n+j});$$

where α_j, β_j are real numbers, $\alpha_k \neq 0$ and $|\alpha_0| + |\beta_0| > 0$. It is possible to associate to a multistep method for first order equations the *generating polynomials*

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j \quad \text{and} \quad \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j :$$

the method is *explicit* if $\sigma(\zeta)$ has degree smaller than k. In this thesis we will consider *symmetric* linear multistep methods, whose polynomials satisfy the properties

$$\alpha_{k-j} = -\alpha_j \qquad \beta_{k-j} = \beta_j \qquad \text{for all } j :$$

another way to define symmetric methods is that $\zeta^k \rho(\zeta^{-1}) = -\rho(\zeta)$ and $\zeta^k \sigma(\zeta^{-1}) = \sigma(\zeta)$.

Some remarkable properties of multistep methods are:

Order A multistep method has *order* r if

$$\rho(e^h) - h\sigma(e^h) = \mathcal{O}(h^{r+1}) \qquad \text{for } h \to 0.$$

Stability A multistep method for first order equations is *stable* if all the roots of $\rho(\zeta)$ satisfy $|\zeta| \leq 1$, and those on the unit circle are simple. We observe that for symmetric methods this requirement implies that all the roots of $\rho(\zeta)$ are simple and lie on the unit circle.

Covergence A multistep method for first order equations is *convergent* if it is stable and it has order $r \geq 1$.

0.2.2 Linear multistep methods for second order equations

A linear multistep method for second order differential equations $\ddot{y} = f(y)$ is defined as

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h^2 \sum_{j=0}^{k} \beta_j f(y_{n+j}); \tag{0.7}$$

as for the methods for first order equations we can associate the *generating polynomials*

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j \quad \text{and} \quad \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j.$$

A linear multistep method for second order equations is *symmetric* if

$$\alpha_{k-j} = \alpha_j \qquad \beta_{k-j} = \beta_j \qquad \text{for all } j :$$

in this case it means that $\zeta^k \rho(\zeta^{-1}) = \rho(\zeta)$ and $\zeta^k \sigma(\zeta^{-1}) = \sigma(\zeta)$.

As for the first order equations, we recall the fundamental properties.

Order A linear multistep method for second order equations has *order* r if

$$\rho(e^h) - h^2 \sigma(e^h) = \mathcal{O}(h^{r+2}) \qquad \text{for } h \to 0.$$

Stability A linear multistep method for second order equations is *stable* if all the zeros of $\rho(\zeta)$ statisfy $|\zeta| \leq 1$, and those on the unit circle are at most double zeros, and *strictly stable* if all the zeros are inside the unit circle except $\zeta = 1$.

Convergence A linear multistep method for second order equations is convergent if it is stable and has order $r \geq 1$.

0.2.3 Symmetric multistep methods for Hamiltonian systems

In this section we describe briefly the techniques used and the results found in [HL04], which is the starting point for the work described in this thesis.

This paper describes and explains the long time behavior of symmetric linear multistep methods (0.7) when applied to second order Hamiltonian equations of the form

$$\ddot{q} = -M^{-1} \nabla U(q) :$$

the Hamiltonian associated to this equation is of the form $H(p, q) = \frac{1}{2} p^{\mathsf{T}} M^{-1} p + U(q)$, with $\dot{p} = M^{-1} \dot{q}$. The linear multistep method is chosen symmetric and *s-stable*, i.e. all the zeros of $\rho(\zeta)$, except for the double root at $\zeta = 1$, are simple and of modulus one.

The study is based on *backward error analysis* for the smooth numerical solution, and on the technique of *modulated Fourier expansions* for the parasitic components.

If we denote by $\zeta_{\pm l}$, $l = 1, \dots k/2 - 1$ the simple roots different from 1, and enumerate them such that $\bar{\zeta_l} = \zeta_{-l}$, the numerical solution can be written as

$$q_n = y(nh) + \sum_{1 \le |l| \le k/2} \zeta_l^n z_l(nh) : \tag{0.8}$$

the functions $z_l(t)$ are called *parasitic components*[1].

The idea of the paper is to prove that $y(t)$ is solution of a modified differential equation, which has a first integral close to the Hamiltonian, and that the parasitic components stay small and bounded for long times: this explains the excellent behavior of the error on the energy shown in the numerical experiments.

One of the results of [HL04] is the following.

Theorem 0.2.1. *The total energy is conserved up to $\mathcal{O}(h^p)$ over times $\mathcal{O}(h^{-r-2})$ along numerical solutions obtained by the s-stable symmetric multistep method of order r:*

$$H(p_n, q_n) = H(p_0, q_0) + \mathcal{O}(h^r) \quad \text{for} \quad nh \le h^{-r-2},$$

where the constant symbolized by $\mathcal{O}$ is independent of n, h with $nh \le h^{-r-2}$.

In [HL04] a similar result is found for quadratic first integral of the form $L(q,p) = p^{\mathsf{T}} A q$, with A such that AM^{-1} is skew-symmetric.

0.3 New Results

In this Section we briefly describe the new results obtained in this thesis.

Chapter 1-2: Long-Term Stability of Symmetric Partitioned Linar Multistep Methods

We extend the results of [HL04], studying the long time behaviour of a symmetric partitioned linear multistep method $\rho_p(\zeta)$, $\sigma_p(\zeta)$, $\rho_q(\zeta)$, $\sigma_q(\zeta)$ when applied to separable Hamiltonians of the form $H(p,q) = T(p) + U(q)$: we assume that the roots of $\rho_p(\zeta)$ and $\rho_q(\zeta)$ are all simple and of modulus 1, and that $\rho_p(\zeta)$ and $\rho_q(\zeta)$ do not have roots in common except for $\zeta = 1$.

We notice from numerical examples that this class of methods seems to work as a symmetric non-symplectic one-step method on relatively short intervals, but we remark that for non reversible and chaotic problems, parasitic components become dominant on longer intervals.

This is explained by the theoretical analysis, which follows the line of [HL04]: the aim is to express the numerical solution as the sum of a smooth part, which corresponds to the main root $\zeta = 1$, and a parasitic part, which corresponds to all the other roots of $\rho_p(\zeta)$ and $\rho_q(\zeta)$ and to their products.

The analysis of the smooth part shows that we can construct a modified differential equation but, differently from [HL04], in general we cannot construct a first integral of the modified equation that is close to the Hamiltonian. Nevertheless, we show that it is possible to improve the behaviour of the smooth solution by imposing additional order conditions, which are functions of parameters that we use to construct the methods. The analysis of the parasitic components shows that they can be computed as the solutions

[1] Here we present a simplified description: in [HL04] the sum in (0.8) is made on the product of the powers of the roots of $\rho(\zeta)$.

of a system of differential algebraic equations, but they can be bounded only under very restrictive hypothesis on the form of the Hamiltonian.

We show a similar analysis also for the near-preservation of quadratic first integrals.

We present as well the construction of partitioned linear multistep methods, starting by the construction of linear multistep methods for first order equations. We then show that it is possible to construct methods of order 4 and 6 such that the root condition and that some additional order conditions are satisfied: then, among the parameters that satisfy these conditions we choose those that optimize the stability of the method.

Chapter 3-4: Symmetric multistep methods for constrained Hamiltonian Systems

We extend the results of [HL04], studying the performance of a symmetric linear multistep method when applied to a constrained Hamiltonian system of the form

$$\begin{aligned} M\ddot{q} &= -\nabla U(q) - G(q)^{\mathsf{T}}\lambda \\ 0 &= g(q). \end{aligned}$$

The algorithm that we study is an extension of SHAKE [RCB77], to explicit linear multistep method for the integration of the positions; the momenta are computed a posteriori with a symmetric central finite differences formula.

We assume, as in [HL04], that the multistep method is s-stable and symmetric, and our aim is to prove that this class of methods has good properties of near preservation of the energy and momenta for long time integrations.

The analysis of the linear problem immediately suggests that it is necessary to add extra assumptions about the roots of $\sigma(\zeta)$, and then we need also that all non zero roots of $\sigma(\zeta)$ are simple of modulus one. We show that it is possible to construct symmetric methods of order 4, 6 and 8 that satisfy these requirements.

As in [HL04], to prove the near preservation of the energy we show that the smooth part $y(t)$ is the solution of a modified differential equation who has a first integral that is close to the Hamiltonian, and that the parasitic components $z_l(t)$ stay small and bounded for long times.

This shows that it is possible to prove a result similar to Theorem 0.2.1, i. e. near-preservation of the energy on long-time integration, even if on a slightly smaller interval with respect to the unconstrained case.

We present a similar analysis for the preservation of quadratic first integrals of the form $L(q,p) = p^{\mathsf{T}}Ap$, with MA skew symmetric.

We study then the error constant and the interval of periodicity of this class of methods.

Chapter 5: Implementation and round-off error optimization

We know that with a high order algorithm it is very easy to obtain a discretization error of the same size of the machine precision: in this case the round-off error becomes the dominating source of error. This motivates the necessity to eliminate all the deterministic error in the implementation of multistep methods described in Chapter 3, in order to obtain an optimized round-off error: in this way we eliminate every source of linear growth and the round-off error will behave like a random walk.

This study has been made analyzing the following points.

- *"SHAKE-like" vs "RATTLE-like" implementation:* we compare two different formulation of the algorithm, observing how the one that is close to RATTLE [And83] leads to a smaller round-off error than the one obtained with SHAKE [RCB77].

- *Influence of initial approximations:* we compare the numerical solutions obtained by using different initial approximations.
- *Manipulation of the coefficients:* we compare the performances of the algorithm using integer coefficients instead of rationals.
- *Compensated summations:* we adapt the well-known compensated summations (Gill (1951), Kahan (1965)) to multistep methods. This technique improves the performance of multistep methods for unconstrained systems but it is not sufficient for constrained systems.
- *Accurate constraint:* we observe how an accurate formulation of the constraint can avoid cancellation errors in the computation of the Lagrange multipliers. It is shown how the accurate formulation can be obtained by using compensated summations: using both the techinques we obtain for constrained system the same improvement that we obtain for unconstrained systems with compensated summations.
- *Iteration until convergence:* we compare the long-time behavior of the error by using a standard stopping criterion and a machine independent stopping criterion for the Newton iteration for the computation of Lagrange multipliers. With this technique we eliminate the linear growth of the round-off error observed in long-time integrations, and we achieve the optimized random walk behavior.

Chapter 1

Long-Term Stability of Symmetric Partitioned Linear Multistep Methods

Note: This chapter is identical to the article [CH13a] in collaboration with E. Hairer. All the computations represented in figures have been made with a different compiler than in [CH13a].

1.1 Introduction

Linear multistep methods are an important alternative to Runge–Kutta one-step methods for the numerical solution of ordinary differential equations. Adams-type methods are frequently used for the integration of nonstiff differential equations, and BDF schemes have excellent properties for the solution of stiff differential equations. In the context of 'geometric numerical integration', where structure-preservation and long-time integration are important, there has been a remarkable publication [QT90], where certain symmetric multistep methods for second order differential equations have been successfully applied to the integration of planetary motion. A theoretical explanation of the observed excellent long-time behavior has been given in [HL04]. It is based on a backward error analysis, and rigorous estimates for the parasitic solution components are obtained, when the system is Hamiltonian of the form $\ddot{q} = -\nabla U(q)$, and derivative approximations are obtained locally by finite differences.

The main aim of the present contribution is to study to which extend this excellent behavior and its theoretical explanation is valid also in more general situations – separable Hamiltonians $H(p,q) = T(p) + U(q)$ with general functions $T(p)$ and $U(q)$, and problems with position dependent kinetic energy. The presentation of the results is in three parts. In the first part we briefly recall the classical theory of partitioned linear multistep methods (order, zero-stability, convergence) and known results on the long-time behavior of symmetric multistep methods for second order Hamiltonian systems. We also present numerical experiments illustrating an excellent long-time behavior in interesting situations. The theoretical explanation of the long-time behavior is based on a backward error analysis for partitioned multistep methods. Part 2 is devoted to the study of the underlying one-step method. This method is symmetric, and we investigate conditions on the coefficients of the method to achieve good conservation of the Hamiltonian. When using multistep methods one is necessarily confronted with parasitic solution components,

because the order of the difference equation is higher than the order of the differential equation. These parasitic terms will be studied in Part 3. On time intervals, where the parasitic terms remain bounded and small, the multistep method essentially behaves like a symmetric one-step method.

1.1.1 Classical Theory of Partitioned Linear Multistep Methods

Hamiltonian systems are partitioned ordinary differential equations of the form

$$\begin{aligned} \dot{p} &= f(p,q), \qquad p(0) = p_0, \\ \dot{q} &= g(p,q), \qquad q(0) = q_0, \end{aligned} \tag{1.1}$$

where $f(p,q) = -\nabla_q H(p,q)$, $g(p,q) = \nabla_p H(p,q)$, and $H(p,q)$ is a smooth scalar energy function. For their numerical solution we consider partitioned linear multistep methods

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j^p p_{n+j} &= h \sum_{j=0}^{k} \beta_j^p f(p_{n+j}, q_{n+j}) \\ \sum_{j=0}^{k} \alpha_j^q q_{n+j} &= h \sum_{j=0}^{k} \beta_j^q g(p_{n+j}, q_{n+j}), \end{aligned} \tag{1.2}$$

where the p and q components are discretized by different multistep methods. Following the seminal thesis of Dahlquist, we denote the generating polynomials of the coefficients α_j, β_j of a multistep method by

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j, \qquad \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j.$$

The generating polynomials of the method (1.2) are thus $\rho_p(\zeta)$, $\sigma_p(\zeta)$ and $\rho_q(\zeta)$, $\sigma_q(\zeta)$, respectively. In the following we collect some basic properties of linear multistep methods (see e.g., [HNW93]).

Zero-stability. A linear multistep method is called stable, if the polynomial $\rho(\zeta)$ satisfies the so-called *root condition*, i.e., all zeros of the equation $\rho(\zeta) = 0$ satisfy $|\zeta| \le 1$, and those on the unit circle are simple.

Order of consistency. A linear multistep method has order r if

$$\frac{\rho(\zeta)}{\log \zeta} - \sigma(\zeta) = \mathcal{O}\big((\zeta - 1)^r\big) \quad \text{for} \quad \zeta \to 1.$$

For a given polynomial $\rho(\zeta)$ of degree k satisfying $\rho(1) = 0$, there exists a unique $\sigma(\zeta)$ of degree k such that the order of the method is at least $k+1$; and there exists a unique $\sigma(\zeta)$ of degree $k-1$ (which yields an explicit method) such that the order of the method is at least k.

Convergence. If both methods of (1.2) are stable and of order r, then we have convergence of order r. This means that for sufficiently accurate starting approximations and for $t_n = nh \le T$ we have

$$\|p_n - p(t_n)\| + \|q_n - q(t_n)\| \le C(T)\, h^r \quad \text{for} \quad h \to 0. \tag{1.3}$$

The constant $C(T)$ is independent of n and h. It typically increases exponentially as a function of T.

Symmetry. A multistep method is symmetric if the coefficients satisfy $\alpha_j = -\alpha_{k-j}$ and $\beta_j = \beta_{k-j}$ for all j. In terms of the generating polynomials this reads

$$\rho(\zeta) = -\zeta^k \rho(1/\zeta), \qquad \sigma(\zeta) = \zeta^k \sigma(1/\zeta). \tag{1.4}$$

If $\alpha_0 = 0$, the number k has to be reduced in this definition. Symmetry together with zero-stabilty imply that all zeros of $\rho(\zeta)$ have modulus one and are simple.

Remark 1.1.1. The idea to use different discretizations for different parts of the differential equation is not new. Already Dahlquist [Dah59, Chapter 7] considers stable combinations of two multistep schemes for the solution of second order differential equations. Often, the vector field is split into a sum of two vector fields (stiff and nonstiff), cf. [ACM99]. In the context of differential-algebraic equations, the differential and algebraic parts can be treated by different methods, cf. [AS95]. An essential difference of these approaches to the present work is the use of symmetric methods with the aim of preserving a qualitatively correct long-time behavior of the numerical approximation.

1.1.2 Known results about the long-time behavior

Classical convergence estimates are usually of the form (1.3), where $C(T) = e^{LT}$ and L is proportional to a Lipschitz constant of the differential equation. They give information only on intervals of length $\mathcal{O}(1)$. Different techniques, usually based on a kind of backward error analysis, are required to get insight into the long-time behavior (e.g., energy-preservation or error growth for nearly integrable systems) of the numerical solution.

From one-step methods it is known that symplecticity and/or symmetry of the numerical integrator play an important role in the long-time behavior of numerical approximations for Hamiltonian systems. This motivates the consideration of symmetric multistep methods. However, already Dahlquist [Dah59, p. 52] pointed out the danger of applying symmetric multistep methods for long-time integration, when he writes[1] "then the unavoidable weak instability arising from the root $\zeta = -1$ of $\rho(\zeta)$ may make [such methods] inferior to methods with a lower value of p in integrations over a long range". Also the analysis of [Hai99] indicates that symmetric multistep methods (applied to the whole differential system) are usually not reliable for integrations over long times. This is the reason why we are mainly interested in partitioned multistep methods, where the characteristic polynomials $\rho_p(\zeta)$ and $\rho_q(\zeta)$ do not have common zeros with the exception of $\zeta = 1$.

For separable Hamiltonian systems with

$$H(p,q) = \frac{1}{2} p^{\mathsf{T}} M^{-1} p + U(q), \tag{1.5}$$

where M is a constant, symmetric, positive definite matrix, the long-time behavior of linear multistep methods is well understood. In this case the differential equation reduces to the second order problem $\ddot{q} = -M^{-1}\nabla U(q)$. Also in the partitioned multistep method the presence of the momenta p_n can be eliminated, which yields

$$\sum_{j=0}^{2k} \alpha_j^{(2)} q_{n+j} = -h^2 \sum_{j=0}^{2k} \beta_j^{(2)} M^{-1} \nabla U(q_{n+j}), \tag{1.6}$$

where the generating polynomial $\rho_2(\zeta), \sigma_2(\zeta)$ of the coefficients $\alpha_j^{(2)}, \beta_j^{(2)}$ are related to those of (1.2) by

$$\rho_2(\zeta) = \rho_p(\zeta)\rho_q(\zeta), \qquad \sigma_2(\zeta) = \sigma_p(\zeta)\sigma_q(\zeta).$$

[1] We thank Gustaf Söderlind for drawing our attention to this part of Dahlquist's thesis.

Formula (1.6) permits the computation of $\{q_n\}$ independent of velocity and momenta. They can be computed a posteriori by a finite difference formula of the form

$$p_n = \frac{1}{h} \sum_{j=-l}^{l} \delta_j \, M \, q_{n+j}. \tag{1.7}$$

This is a purely local approach, which does not influence the propagation of the numerical solution, and therefore has no effect on its long-time behavior.

We now present a few interesting results from the publication [HL04] about the long-time behavior of numerical solutions. This article considers linear multistep methods (1.6), which do not necessarily originate from a partitioned method (1.2), together with local approximations of the momenta. Assumptions on the method (1.6) are the following:

(A1) it is of order r, i.e., $\rho_2(\zeta)/(\log \zeta)^2 - \sigma_2(\zeta) = \mathcal{O}((\zeta-1)^r)$ for $\zeta \to 1$,

(A2) it is symmetric, i.e., $\rho_2(\zeta) = \zeta^k \rho_2(1/\zeta)$ and $\sigma_2(\zeta) = \zeta^k \sigma_2(1/\zeta)$,

(A3) it is s-stable, i.e., apart from the double zero at 1, all zeros of $\rho_2(\zeta)$ are simple and of modulus one.

Under these assumptions we have the following results on the long-time behavior:

- the total energy (1.5) is preserved up to $\mathcal{O}(h^r)$ over times $\mathcal{O}(h^{-r-2})$, i.e.,

$$H(p_n, q_n) = H(p_0, q_0) + \mathcal{O}(h^r) \quad \text{for} \quad nh \le h^{-r-2},$$

- quadratic first integrals of the form $L(p,q) = p^{\mathsf{T}} A \, q$ are nearly preserved:

$$L(p_n, q_n) = L(p_0, q_0) + \mathcal{O}(h^r) \quad \text{for} \quad nh \le h^{-r-2},$$

- for integrable reversible systems (under suitable assumptions, see [HL04]) we have for the angle variable $\Theta(p,q)$ and the action variable $I(p,q)$ the estimates

$$\begin{aligned} \Theta(p_n, q_n) &= \Theta(p_0, q_0) + \mathcal{O}(t\, h^r) \\ I(p_n, q_n) &= I(p_0, q_0) + \mathcal{O}(h^r) \end{aligned} \qquad \text{for} \quad 0 \le t = nh \le h^{-r}.$$

The constants symbolized by $\mathcal{O}$ are independent of n and h.

1.1.3 Numerical experiments

For systems with Hamiltonian (1.5), partitioned linear multistep methods of the form (1.2) have the same long-time behavior as linear multistep methods for second order problems (Section 1.1.2) even if the derivative approximation is not given locally by a finite difference formula as in (1.7). The aim of this section is to get some insight into the long-time behavior of partitioned linear multistep methods (1.2) applied to Hamiltonian systems that are more general than (1.5).

Separable Hamiltonian systems. Let us first consider separable polynomial Hamiltonians $H(p,q) = T(p) + U(q)$, where

$$T(p) = \sum_{2 \le j+k \le 3} a_{jk} \, p_1^j \, p_2^k + \left(p_1^4 + p_2^4\right), \quad U(q) = \sum_{2 \le j+k \le 3} b_{jk} \, q_1^j \, q_2^k + \left(q_1^4 + q_2^4\right).$$

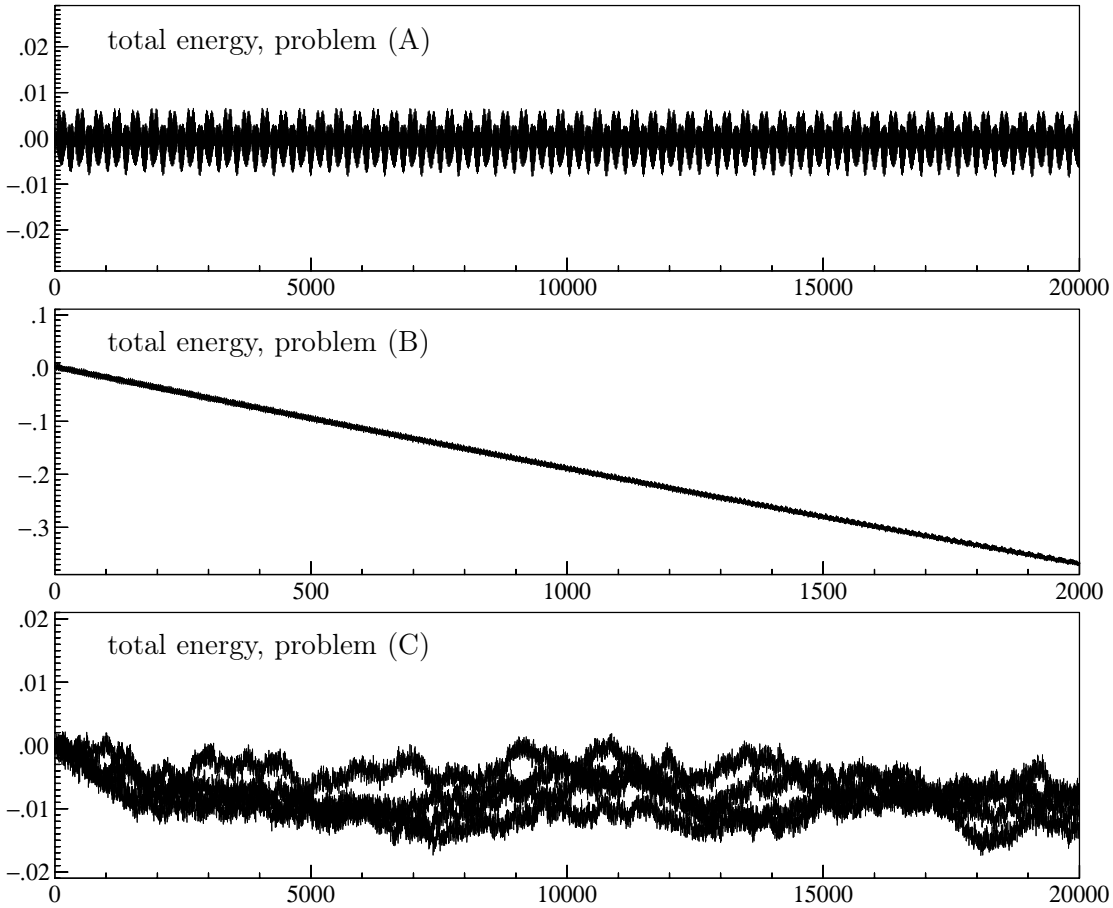


Figure 1.1 – Numerical Hamiltonian of method 'plmm2' applied with step size $h = 0.005$ for problems (A) and (B), and with $h = 001$ for problem (C); initial values $q_1(0) = 1$, $q_2(0) = -1.2$, $p_1(0) = 0.2$, $p_2(0) = -0.9$. Starting approximations are computed with high precision.

The positive definite quartic terms imply that solutions remain in a compact set. We consider the following three situations:

(A) Non-vanishing coefficients are $a_{02} = 1$, $a_{20} = 1$, and $b_{02} = 2$, $b_{20} = 1$, $b_{03} = 1$. Since $T(-p) = T(p)$, the system is reversible with respect to $p \leftrightarrow -p$. Moreover, it is separated into two systems with one degree of freedom.

(B) Non-vanishing coefficients are $a_{02} = 1$, $a_{20} = 1$, $a_{03} = 1$, $a_{30} = -0.5$, and $b_{02} = 2$, $b_{20} = 1$, $b_{03} = 1$. The system is not reversible, but still equivalent to two systems with one degree of freedom.

(C) Non-vanishing coefficients are $a_{02} = 1$, $a_{20} = 1$, and $b_{02} = 2$, $b_{20} = 1$, $b_{12} = -1$, $b_{21} = 2$. The system is reversible, and it is a coupled system with two degrees of freedom.

We consider the following partitioned linear multistep methods:

plmm2	$\rho_p(\zeta) = (\zeta - 1)(\zeta + 1)$	$\sigma_p(\zeta) = 2\,\zeta$
	$\rho_q(\zeta) = (\zeta - 1)(\zeta^2 + 1)$	$\sigma_q(\zeta) = \zeta^2 + \zeta$
plmm4	$\rho_p(\zeta) = \zeta^4 - 1$	$\sigma_p(\zeta) = \frac{4}{3}\big(2\zeta^3 - \zeta^2 + 2\zeta\big)$
	$\rho_q(\zeta) = \zeta^5 - 1$	$\sigma_q(\zeta) = \frac{5}{24}\big(11\zeta^4 + \zeta^3 + \zeta^2 + 11\zeta\big)$

Table 1.1 – Numerical energy behavior on intervals of length $\mathcal{O}(h^{-2})$; t is time, h the step size.

method	problem (A)	problem (B)	problem (C)
plmm2, order 2	$\mathcal{O}(h^2)$	$\mathcal{O}(th^2)$	$\mathcal{O}(\sqrt{t}\,h^2)$
plmm4, order 4	$\mathcal{O}(h^4)$	$\mathcal{O}(th^4)$	$\mathcal{O}(h^4)$
plmm4c, order 4	$\mathcal{O}(h^4)$	$\mathcal{O}(h^4 + th^6)$	$\mathcal{O}(h^4)$

1.1 shows the numerical Hamiltonian for the second order method 'plmm2', and Table 1.1 presents the qualitative behavior in dependence of time and step size. Looking at Figure 1.1, we notice that this partitioned multistep method behaves very similar to (non-symplectic) symmetric one-step methods, as can be seen from the experiments of [HMS09]. For non-reversible problems without any symmetry we have a linear growth in the energy, for reversible problems we observe boundedness for integrable systems and for problems with one degree of freedom, and we observe a random walk behavior of the numerical energy for chaotic solutions. This is illustrated by plotting the numerical Hamiltonian of 4 trajectories with randomly perturbed initial values (perturbation of size $\approx 10^{-15}$) for problem (C).

The intervals considered in the experiments of Figure 1.1 are relatively short. What happens on longer time intervals? For problem (A), the numerical energy of the method 'plmm2' shows the same regular, bounded, $\mathcal{O}(h^2)$ behavior on intervals as long as 10^7. No secular terms and no influence of parasitic components can be observed. For problem (B) the linear error growth in the energy as $\mathcal{O}(th^2)$ can be observed on intervals of length $\mathcal{O}(h^{-2})$. The behavior for problem (C) is shown in Figure 1.2. We observe that after a time that is proportional to h^{-2} (halving the step size increases the length of the interval by a factor four) an exponential error growth is superposed to the random walk behavior of Figure 1.1. Such a behavior is not possible for symmetric one-step methods. It will be explained by the presence of parasitic solution components.

We have repeated all experiments with the fourth order partitioned linear multistep method 'plmm4' with characteristic polynomials given at the beginning of this section. Table 1.1 shows the behavior on intervals of length $\mathcal{O}(h^{-2})$. Whereas the behavior for problems (A) and (B) is expected, we cannot observe a random walk behavior for problem (C). On very long time intervals, the energy error remains nicely bounded of size $\mathcal{O}(h^4)$ for the problem (A). For the problems (B) and (C), however, an exponential error growth like $\delta \exp(ch^2 t)$ with small δ is superposed, which becomes visible after an interval of length $\mathcal{O}(h^{-2})$. , the exponent two in the length of the interval is not related to the order of the method.

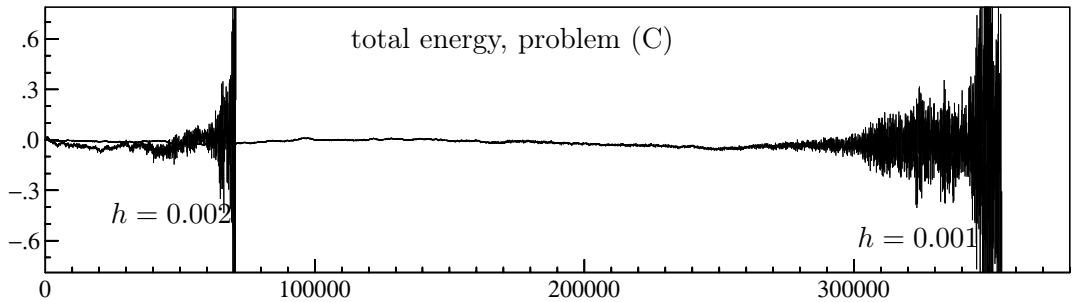


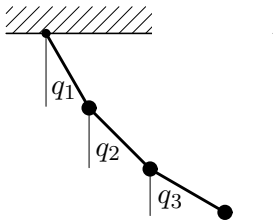
Figure 1.2 – Numerical Hamiltonian of method 'plmm2' for problem (C); data as in Figure 1.1, but on a longer time interval.

Triple pendulum. For non-separable Hamiltonians, symmetric and/or symplectic one-step methods are in general implicit. It is therefore of interest to study the behavior of explicit symmetric multistep methods applied to such systems. We consider the motion of a triple pendulum, which leads to a Hamiltonian system with

$$H(p,q) = \tfrac{1}{2}\, p^{\mathsf{T}} M(q)^{-1} p + U(q),$$

where $U(q) = -3\cos q_1 - 2\cos q_2 - \cos q_3$ and

$$M(q) = \begin{pmatrix} 3 & 2\cos(q_2-q_1) & \cos(q_3-q_1) \\ 2\cos(q_2-q_1) & 2 & \cos(q_3-q_2) \\ \cos(q_3-q_1) & \cos(q_3-q_2) & 1 \end{pmatrix}.$$



This matrix is positive definite with $\det M(q) = 4 - 2\cos^2(q_2-q_1) - \cos^2(q_3-q_2)$. We have experimented with both partitioned multistep methods (order 2 and order 4) and we observed that the methods give excellent results when the angles are not to large, and the motion is not too chaotic.

For example, if we take initial values $q_1(0) = \pi/12$, $q_2(0) = \pi/6$, for $q_3(0)$ a value between 0 and $5\pi/12$, and zero initial values for the velocities, then the error in the Hamiltonian is of size $\mathcal{O}(h^2)$ (for 'plmm2') and $\mathcal{O}(h^4)$ (for 'plmm4') without any drift. This has been verified numerically on an interval $[0, 10^7]$. Changing the initial value for $q_3(0)$ to $-\pi/12$ shows an exponential increase of the error after $t \approx 4\cdot 10^6$, and a change to $6\pi/12$ shows such a behavior already at $t \approx 4\,000$.

Ablowitz–Ladik discrete nonlinear Schrödinger equation. As an example of a completely integrable lattice equation we consider the Ablowitz–Ladik discrete nonlinear Schrödinger equation (see [AL76])

$$i\,\dot{u}_k + \frac{1}{\Delta x^2}\big(u_{k+1} - 2u_k + u_{k-1}\big) + |u_k|^2(u_{k+1} + u_{k-1}) = 0,$$

under periodic boundary conditions $u_{k+N} = u_k$, where $\Delta x = L/N$. Separating real and imaginary parts in the solution $u_k = p_k + i\,q_k$, the equation becomes

$$\begin{aligned} \dot{p}_k &= -\frac{1}{\Delta x^2}\big(q_{k+1} - 2q_k + q_{k-1}\big) - (p_k^2 + q_k^2)(q_{k+1} + q_{k-1}) \\ \dot{q}_k &= \frac{1}{\Delta x^2}\big(p_{k+1} - 2p_k + p_{k-1}\big) + (p_k^2 + q_k^2)(p_{k+1} + p_{k-1}) \end{aligned} \tag{1.8}$$

with boundary conditions $p_{k+N} = p_k$ and $q_{k+N} = q_k$. This system can be written in the non-canonical Hamiltonian form

$$\dot{p} = -D(p,q)\,\nabla_q H(p,q), \qquad \dot{q} = D(p,q)\,\nabla_p H(p,q),$$

where $D(p,q)$ is the diagonal matrix with entries $d_k(p,q) = \frac{1}{\Delta x}(1 + \Delta x^2(p_k^2 + q_k^2))$, and the Hamiltonian is given by

$$H(p,q) = \frac{1}{\Delta x}\sum_{k=1}^{N}\big(p_k p_{k-1} + q_k q_{k-1}\big) - \frac{1}{\Delta x^3}\sum_{k=1}^{N}\ln\big(1 + \Delta x^2(p_k^2 + q_k^2)\big). \tag{1.9}$$

Furthermore, the expression

$$I(p,q) = \frac{1}{\Delta x}\sum_{k=1}^{N}\big(p_k p_{k-1} + q_k q_{k-1}\big) \tag{1.10}$$

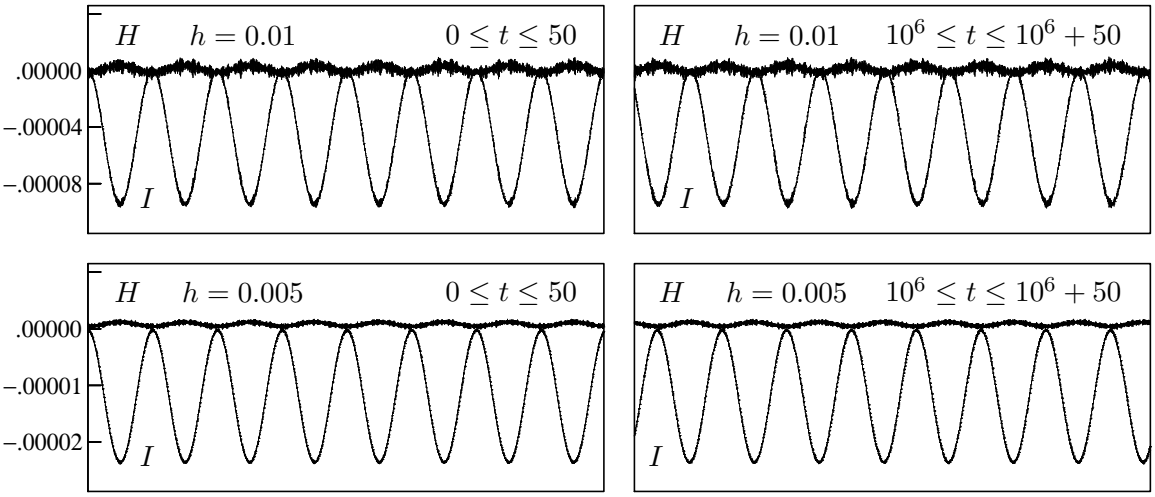


Figure 1.3 – Numerical preservation of the invariants H and I, defined in (1.9) and (1.10), with the method 'plmm2' applied with step sizes $h = 0.01$ and $h = 0.005$; initial data are that of (1.11).

is a first integral of the system (1.8). Since the system is completely integrable, there are in addition $N - 2$ other independent first integrals.

Since we are confronted with a Poisson system with non-separable Hamiltonian, there do not exist symplectic and/or symmetric integrators that are explicit. It is therefore of high interest to study the performance of explicit partitioned linear multistep methods, when applied to the system (1.8). Notice that the system is reversible with respect to the symmetries $p \leftrightarrow -p$ and $q \leftrightarrow -q$. Following [Sch99, HS00], we consider initial values

$$p_k(0) = \tfrac{1}{2}\big(1 - \varepsilon \cos(bx_k)\big), \qquad q_k(0) = 0, \tag{1.11}$$

where $x_k = -L/2 + (k-1)\Delta x$, $\Delta x = L/N$, $b = 2\pi/L$ with $L = 2\pi\sqrt{2}$, and $\varepsilon = 0.01$. We apply the second order method 'plmm2' to the system with $N = 16$, and we use various time step sizes for integrations over long time intervals. Figure 1.3 shows the error in both first integrals, H and I, to the left on the first subinterval of length 50, and to the right on the final subinterval starting at $t = 10^6$. We observe that halving the step size decreases the error by a factor of $4 = 2^2$, which is in accordance with a second order integrator. Similar to an integration with a symplectic scheme, the partitioned multistep method behaves very well over long times and no drift in the invariants can be seen. Comparing the results for different step sizes at the final interval, we notice a time shift in the numerical solution, but amplitude and shape of the oscillations are not affected. We also observe that the errors are a superposition of a slowly varying function scaled with h^2, and of high oscillations that decrease faster than with a factor 4, when the step size is halved.

The same qualitative behavior can be observed with the 4th order, explicit, partitioned multistep method 'plmm4' for step sizes smaller than $h = 0.005$. As expected, the error decreases by a factor of $16 = 2^4$ when having the step size. For larger values of ε, say $\varepsilon \geq 0.05$ the behavior of the partitioned multistep method is less regular.

Further numerical experiments can be found in [Hai99]. Excellent long-time behavior of partitioned linear multistep methods is reported for the Kepler problem and for a test problem in molecular dynamics simulation (frozen Argon crystal). Exponentially fitted partitioned linear multistep methods are considered in [VS04] for the long-term integration of N-body problems.

1.2 Long-time analysis of the underlying one-step method

For one-step methods, the long-time behavior of numerical approximations is easier to analyze than for multistep methods. Whereas the notions of symplecticity and energy preservation are straightforward for one-step methods, this is not the case for multistep methods. It has been shown by Kirchgraber [Kir86] that the numerical solution of strictly stable[2] linear multistep methods essentially behaves like that of a one-step method, which we call *underlying one-step method*. For a fixed step size h and a differential equation $\dot y = f(y)$, it is defined as the mapping $\Phi_h(y)$, such that the sequence defined by $y_{n+1} = \Phi_h(y_n)$ satisfies the multistep formula. This means that for starting approximations given by $y_j = \Phi_h^j(y_0)$ for $j = 0, 1, \dots, k-1$, the numerical approximations obtained by the multistep formula coincides with that of the underlying one-step method (neglecting round-off effects).

For symmetric linear multistep methods, which cannot be strictly stable, such an underlying one-step method exists as a formal series in powers of h (see [Fen95, page 274] and [HLW06, Sect. XV.2.2]). Despite its non-convergence, it can give much insight into the long-time behavior of the method.

1.2.1 Analysis for the harmonic oscillator

Consider a harmonic oscillator, written as a first order Hamiltonian system,

$$\begin{aligned} \dot p &= -\omega\, q, & p(0) &= p_0, \\ \dot q &= \omega\, p, & q(0) &= q_0. \end{aligned}$$

Applying the partitioned linear multistep method (1.2) to this system yields the difference equations

$$\rho_p(E)\, p_n = -\omega h\, \sigma_p(E)\, q_n, \qquad \rho_q(E)\, q_n = \omega h\, \sigma_q(E)\, p_n, \tag{1.12}$$

where we have made use of the shift operator $Ey_n = y_{n+1}$. Looking for solutions of the form $p_n = a\zeta^n$, $q_n = b\zeta^n$ we are led to the 2-dimensional linear system

$$R(\omega h, \zeta) \begin{pmatrix} a \\ b \end{pmatrix} = 0 \qquad \text{with} \qquad R(\omega h, \zeta) = \begin{pmatrix} \rho_p(\zeta) & \omega h\, \sigma_p(\zeta) \\ -\omega h\, \sigma_q(\zeta) & \rho_q(\zeta) \end{pmatrix}. \tag{1.13}$$

It has a nontrivial solution if and only if $\det R(\omega h, \zeta) = 0$. For small values of ωh the roots of this equation are close to the zeros of the polynomials $\rho_p(\zeta)$ and $\rho_q(\zeta)$. By consistency we have two roots close to 1, they are conjugate to each other, and they satisfy $\zeta_0 = \zeta_0(\omega h) = 1 + i\omega h + \mathcal{O}(h^2)$ and $\overline{\zeta}_0 = \overline{\zeta}_0(\omega h) = 1 - i\omega h + \mathcal{O}(h^2)$ (principal roots). They lead to approximations to the exact solution, which is a linear combination of $e^{i\omega t}$ and $e^{-i\omega t}$. The other roots lead to parasitic terms in the numerical approximations. The general solution $(p_n, q_n)^{\mathsf T}$ of the difference equation (1.12) is in fact a linear combination of $\zeta^n (a, b)^{\mathsf T}$, where ζ is a root of $\det R(\omega h, \zeta) = 0$, and the vector $(a, b)^{\mathsf T}$ satisfies the linear system (1.13).

Underlying one-step method. We consider a numerical solution of (1.12) that is built only on linear combinations of ζ_0^n and $\overline{\zeta}_0^n$. It has to be of the form

$$\begin{pmatrix} p_n \\ q_n \end{pmatrix} = \Phi_n \begin{pmatrix} p_0 \\ q_0 \end{pmatrix}, \qquad \Phi_n = \tfrac{1}{2}\big(\zeta_0^n + \overline{\zeta}_0^n\big)\, I + \tfrac{1}{2i}\big(\zeta_0^n - \overline{\zeta}_0^n\big)\, C, \tag{1.14}$$

[2] A linear multistep is called strictly stable, if $\zeta_1 = 1$ is a simple zero of the ρ polynomial, and all other zeros have modulus strictly smaller than one.

where the matrix C satisfies $R_0(I - i\,C) = 0$ and $\overline{R}_0(I + i\,C) = 0$, so that the vectors multiplying ζ_0^n and $\overline{\zeta}_0^n$ satisfy the relation (1.13) with $R_0 = R(\omega h, \zeta_0)$. It follows from the consistency of the method that for small but nonzero ωh the real and imaginary parts of the matrix R_0 are invertible. This permits us to compute the real matrix $C = -i\,(R_0 + \overline{R}_0)^{-1}(R_0 - \overline{R}_0)$. As a consequence of $R_0 = \frac{1}{2}(R_0 + \overline{R}_0)(I + i\,C)$ and $\det R_0 = 0$ we have $\det C = 1$ and $\operatorname{trace} C = 0$, which implies $C^2 = -I$. The matrix Φ_n of (1.14) thus satisfies $\Phi_{n+1} = \Phi_n \Phi_1$, and consequently $\Phi_n = \Phi_1^n$, so that the underlying one-step method is seen to be given by

$$\begin{pmatrix} p_{n+1} \\ q_{n+1} \end{pmatrix} = \Phi(\omega h) \begin{pmatrix} p_n \\ q_n \end{pmatrix}, \qquad \Phi(\omega h) = \tfrac{1}{2}\big(\zeta_0 + \overline{\zeta}_0\big)\, I + \tfrac{1}{2\,i}\big(\zeta_0 - \overline{\zeta}_0\big)\, C. \tag{1.15}$$

Notice that $\Phi(\omega h)$ is not an analytic function of ωh.

Properties of the underlying one-step method. The above derivation is valid for all partitioned multistep methods. If the method is symmetric, also the coefficients of the polynomial $\det R(h\omega, \zeta)$ are symmetric, so that with $\zeta_0 = \zeta_0(\omega h)$ also its inverse is a solution of $\det R(h\omega, \zeta) = 0$. This implies $\zeta_0^{-1} = \overline{\zeta}_0$, and hence also $|\zeta_0| = 1$. Similarly, the symmetry of the methods (ρ_p, σ_p) and (ρ_q, σ_q) imply that $C(-\omega h) = C(\omega h)$. Consequently, we have $\Phi(-\omega h)\Phi(\omega h) = I$, which proves the *symmetry* of the underlying one-step method.

Furthermore, the mapping defined by the matrix $\Phi(\omega h)$ is *symplectic*:

$$\Phi(\omega h)^{\mathsf{T}} J\, \Phi(\omega h) = J \qquad \text{with} \qquad J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{1.16}$$

This follows from the relations $C^{\mathsf{T}} J + J\,C = 0$ and $C^{\mathsf{T}} J\, C = J$, which are a consequence of $\det C = 1$ and $\operatorname{trace} C = 0$.

Since the eigenvalues of C are $\pm i$, we have

$$T\,C\,T^{-1} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \qquad \text{with} \qquad T = \begin{pmatrix} 1 & 0 \\ a & b \end{pmatrix},$$

where (a, b) is the first row of the matrix C. Notice that we have $a = \mathcal{O}((\omega h)^2)$ and $b = 1 + \mathcal{O}((\omega h)^2)$. This transformation implies that $T\Phi(\omega h)T^{-1}$ is an orthogonal matrix, so that

$$\frac{\omega}{2} \left\| T \begin{pmatrix} p_n \\ q_n \end{pmatrix} \right\|^2 = \frac{\omega}{2} \left(p_n^2 + (a p_n + b q_n)^2 \right)$$

is a conserved quantity that is $\mathcal{O}(h^2)$ close to the true Hamiltonian.

Parasitic solution components. The complete solution of the difference equation (1.12) is given by

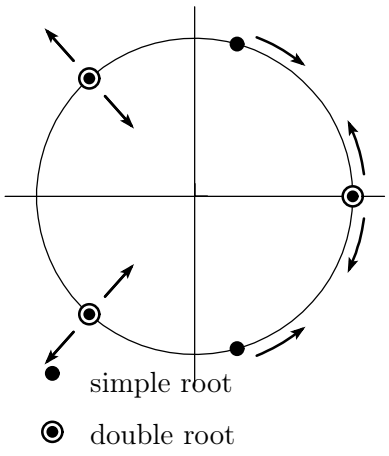
$$\begin{pmatrix} p_n \\ q_n \end{pmatrix} = \Phi_1(\omega h)^n \begin{pmatrix} a \\ b \end{pmatrix} + \sum_{l=1}^{2k-2} \zeta_l(\omega h)^n \begin{pmatrix} a_l \\ b_l \end{pmatrix},$$

where $\zeta_l(\omega h)$ are the roots of $\det R(\omega h, \zeta) = 0$ which are different from the principal roots $\zeta_0(\omega h)$ and $\overline{\zeta}_0(\omega h)$. They are called parasitic roots of the method. Initial approximations (p_j, q_j) for $j = 0, 1, \ldots, k-1$ uniquely determine the vectors (a, b) and (a_l, b_l), recalling that (a_l, b_l) has to satisfy the relation (1.13).

If the starting values (p_j, q_j) approximate for $j = 0, 1, \ldots, k-1$ the exact solution $(p(t_0 + jh), q(t_0 + jh))$ up to an error of size $\mathcal{O}(h^{\nu+1})$ with $\nu \le r$, then we have

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} p_0 \\ q_0 \end{pmatrix} + \mathcal{O}(h^{\nu+1}), \qquad \begin{pmatrix} a_l \\ b_l \end{pmatrix} = \mathcal{O}(h^{\nu+1}) \qquad \text{for all } l.$$

For zero-stable multistep methods, all roots of $\det R(\omega h, \zeta) = 0$ can be bounded by $|\zeta_l(\omega h)| \le 1 + \gamma\omega h$ (here $\gamma > 0$ and $\omega > 0$). This implies that $|\zeta_l(\omega h)^n| \le e^{\gamma\omega T}$ for $nh \le T$, and the parasitic solution components remain small of size $\mathcal{O}(h^{\nu+1})$ on intervals of fixed length. To have a similar estimate on arbitrarily long intervals, the roots $\zeta_l(\omega h)$ have to be bounded by 1.
In general, we do not have a control on the modulus of ζ_l. However, for symmetric methods we know that with ζ_l not only the complex conjugate $\overline{\zeta}_l$, but also the inverse ζ_l^{-1} are roots of $\det R(\omega h, \zeta) = 0$. Furthermore, the roots $\zeta_l(\omega h)$ depend continuously on its argument.



If $\zeta_l(0)$ is a double root of $\det R(0, \zeta) = 0$, then it is possible that it splits for $\omega h > 0$ into a pair of roots, one of which has modulus larger than 1, and one smaller than 1 (see the figure). If $\zeta_l(0)$ is a simple root, then we must have $\overline{\zeta}_l(\omega h) = \zeta_l(\omega h)^{-1}$, implying $|\zeta_l(\omega h)| = 1$ for sufficiently small $\omega h > 0$.

Consequently, if apart from the double root at 1, all roots of $\det R(0, \zeta) = 0$ are simple (i.e., with the exception of 1, all zeros of $\rho_p(\zeta)$ are different from those of $\rho_q(\zeta)$), the parasitic solution components remain bounded of size $\mathcal{O}(h^{\nu+1})$ independent of the length of the integration interval.

Linear change of coordinates. Partitioned linear multistep methods are invariant with respect to linear transformations of the form $\tilde{p} = T_p\, p$, $\tilde{q} = T_q\, q$. However, care has to be taken when p and q components are mixed. Suppose, for example, that after such a transformation the harmonic oscillator reduces to a Hamiltonian system with (we put $\omega = 1$ for convenience)

$$H(p, q) = \tfrac{1}{2}\big(p^2 + 2\,\varepsilon\, p\, q + q^2\big),$$

where $\varepsilon \neq 0$ is a small parameter. An application of the partitioned multistep method yields the difference equation

$$\begin{aligned} \rho_p(E)\, p_n &= -h\,\big(\varepsilon\, \sigma_p(E)\, p_n + \sigma_p(E)\, q_n\big), \\ \rho_q(E)\, q_n &= h\,\big(\sigma_q(E)\, p_n + \varepsilon\, \sigma_q(E)\, q_n\big). \end{aligned} \tag{1.17}$$

Instead to (1.13) we are led this time to the system

$$R(h, \zeta)\begin{pmatrix} a \\ b \end{pmatrix} = 0 \quad \text{with} \quad R(h, \zeta) = \begin{pmatrix} \rho_p(\zeta) + \varepsilon h \sigma_p(\zeta) & h\,\sigma_p(\zeta) \\ -h\,\sigma_q(\zeta) & \rho_q(\zeta) - \varepsilon h \sigma_q(\zeta) \end{pmatrix}.$$

Even if we only consider symmetric partitioned linear multistep methods, the coefficients of the polynomial $\det R(h, \zeta)$ are no longer symmetric, so that the modulus of its zeros is in general not equal to one. A straightforward computation shows that for simple roots of $R(0, \zeta) = 0$ (for example if we have $\rho_p(\zeta_l) = 0$ but $\rho_q(\zeta_l) \neq 0$), the continuous continuation satisfies

$$\zeta_l(h) = \zeta_l\big(1 - \mu_l \varepsilon h + \mathcal{O}(h^2)\big), \qquad \mu_l = \frac{\sigma_p(\zeta_l)}{\zeta_l\, \rho_p'(\zeta_l)}.$$

From the symmetry of the method it follows that μ_l is a real number. It is called growth parameter. We conclude from this asymptotic formula that $|\zeta_l(h)| > 1$ for small h, if the

product $\mu_l\varepsilon$ is negative. In such a situation parasitic solution components grow exponentially with time, and the numerical solution becomes meaningless on integration intervals whose length T is such that $h^{\nu+1}e^{-\mu_l\varepsilon T} \geq 1$.

1.2.2 Backward error analysis (smooth numerical solution)

An important tool for the study of the long-time behavior of numerical approximations is 'backward error analysis'. The idea is to interpret the numerical solution of a one-step method as the exact solution of a modified differential equation (for details see Chapter IX of [HLW06]). For linear multistep methods, it is in principle possible to construct the underlying one-step method as a formal series in powers of the step size h, and then to apply the well-established techniques. Here, we follow the approach of [Hai99, HL04], where the modified differential equation is directly obtained from the multistep schemes without passing explicitly through the underlying one-step method.

Theorem 1.2.1 (modified differential equation)**.** *Consider a consistent, partitioned linear multistep method (1.2), applied to a partitioned system (1.1). There then exist h-independent functions $f_j(p,q)$, $g_j(p,q)$, such that for every truncation index N every solution $p_h(t), q_h(t)$ of the system*

$$\begin{aligned} \dot p &= f(p,q) + hf_1(p,q) + \ldots + h^{N-1}f_{N-1}(p,q) \\ \dot q &= g(p,q) + hg_1(p,q) + \ldots + h^{N-1}g_{N-1}(p,q) \end{aligned} \tag{1.18}$$

satisfies the multistep formula up to a defect of size $\mathcal{O}(h^{N+1})$, i.e.,

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j^p p_h(t+jh) &= h\sum_{j=0}^{k} \beta_j^p f\big(p_h(t+jh), q_h(t+jh)\big) + \mathcal{O}(h^{N+1}) \\ \sum_{j=0}^{k} \alpha_j^q q_h(t+jh) &= h\sum_{j=0}^{k} \beta_j^q g\big(p_h(t+jh), q_h(t+jh)\big) + \mathcal{O}(h^{N+1}). \end{aligned} \tag{1.19}$$

The constant symbolized by $\mathcal{O}$ is independent of h, but depends on the truncation index N. It also depends smoothly on t. If the method is of order r, then we have $f_j(p,q) = g_j(p,q) = 0$ for $1 \leq j < r$.

Proof. We closely follow the proof for second order equations in [HL04]. Denoting time differentiation by D, the Taylor series expansion of a function can be written as $y(t+h) = e^{hD}y(t)$. The equations (1.19) thus become

$$\begin{aligned} \rho_p(e^{hD})p_h(t) &= h\,\sigma_p(e^{hD})f\big(p_h(t), q_h(t)\big) + \mathcal{O}(h^{N+1}) \\ \rho_q(e^{hD})q_h(t) &= h\,\sigma_q(e^{hD})g\big(p_h(t), q_h(t)\big) + \mathcal{O}(h^{N+1}). \end{aligned} \tag{1.20}$$

With the coefficients of the expansions

$$\frac{x\,\sigma_p(e^x)}{\rho_p(e^x)} = 1 + \mu_1^p x + \mu_2^p x^2 + \ldots, \qquad \frac{x\,\sigma_q(e^x)}{\rho_q(e^x)} = 1 + \mu_1^q x + \mu_2^q x^2 + \ldots, \tag{1.21}$$

this becomes equivalent to (omitting the argument t)

$$\begin{aligned} \dot p_h &= \big(1 + \mu_1^p hD + \mu_2^p h^2 D^2 + \ldots\big) f(p_h, q_h) + \mathcal{O}(h^N) \\ \dot q_h &= \big(1 + \mu_1^q hD + \mu_2^q h^2 D^2 + \ldots\big) g(p_h, q_h) + \mathcal{O}(h^N). \end{aligned} \tag{1.22}$$

For a function $\Psi(p,q)$, we have

$$D\Psi(p_h,q_h) = \partial_p\Psi(p_h,q_h)f_h(p_h,q_h) + \partial_q\Psi(p_h,q_h)g_h(p_h,q_h),$$

where the functions $f_h(p,q)$ and $g_h(p,q)$ are an abbreviation for the right-hand side of (1.18). Applying this formula iteratively to the expressions in (1.22) and collecting equal powers of h, a comparison of the equations (1.18) and (1.22) determines recursively the functions $f_j(p,q)$ and $g_j(p,q)$. □

The flow of the modified differential equation (1.18) depends on the parameter h. If we denote this flow by $\varphi_t^{[h]}(p,q)$, then the underlying one-step method of the partitioned linear multistep method is given by $\Phi_h(p,q) = \varphi_h^{[h]}(p,q)$ up to an error of size $\mathcal{O}(h^{N+1})$.

Corollary 1.2.2. *Assume that the partitioned linear multistep method is symmetric, i.e., both multistep schemes satisfy the symmetry relations (1.4). We then have:*

a) The expansion of the vector field of the modified differential equation (1.18) is in even powers of h.

b) If the differential equation (1.1) is reversible, i.e., $f(-p,q) = f(p,q)$ and $g(-p,q) = -g(p,q)$, then the modified differential equation (1.18) is also reversible.

Proof. The symmetry relations (1.4) imply that the expressions of (1.21) are even functions of x. This proves statement (a).

If $\big(f_h(p,q), g_h(p,q)\big)$ is a reversible vector field, then the function $D^2\Psi(p,q)$ has the same parity in p as the function $\Psi(p,q)$. As a consequence of the recursive construction of the modified differential equation, and of the fact that only even powers of D appear in (1.22), this observation proves the statement (b). □

Theorem 1.2.1 tells us that the solution of the truncated modified differential equation (1.18) satisfies the multistep formulas up to a defect of size $\mathcal{O}(h^{N+1})$. Consequently, the classical analysis shows that on intervals of length $T = \mathcal{O}(1)$,

$$\|p_n - p_h(nh)\| + \|q_n - q_h(nh)\| \le C(T)\, h^N.$$

1.2.3 Near energy preservation

Whereas the analysis of the previous Section 1.2.2 is valid for general partitioned differential equations, we assume here that the vector field is Hamiltonian and given by

$$f(p,q) = -\nabla_q H(p,q), \qquad g(p,q) = \nabla_p H(p,q). \tag{1.23}$$

In this situation the exact solution satisfies $H\big(p(t),q(t)\big) = \mathit{Const}$, and it is of interest to study whether numerical approximations of partitioned linear multistep methods (nearly) preserve the energy $H(p,q)$ over long times. Recall that in this chapter we consider only 'smooth' numerical solutions, which are given by the flow of the modified differential equation (1.18) up to an arbitrarily small error of size $\mathcal{O}(h^N)$. We therefore have to investigate the near preservation of $H\big(p_h(t),q_h(t)\big)$.

The solution of the truncated modified equation satisfies (1.20). Instead of dividing by the ρ polynomial, which led us to the construction of the modified differential equation, we divide the relation by the σ polynomial. This leads to

$$\begin{aligned} \big(1+\lambda_1^p hD+\lambda_2^p h^2D^2+\ldots\big)\dot p_h &= -\nabla_q H(p_h,q_h)+\mathcal{O}(h^N) \\ \big(1+\lambda_1^q hD+\lambda_2^q h^2D^2+\ldots\big)\dot q_h &= \nabla_p H(p_h,q_h)+\mathcal{O}(h^N), \end{aligned} \tag{1.24}$$

where the coefficients in the expansion are given by

$$\frac{\rho_p(e^x)}{x\,\sigma_p(e^x)} = 1+\lambda_1^p x+\lambda_2^p x^2+\ldots, \qquad \frac{\rho_q(e^x)}{x\,\sigma_q(e^x)} = 1+\lambda_1^q x+\lambda_2^q x^2+\ldots\ . \tag{1.25}$$

For symmetric methods, we are concerned with even functions of x, so that the expansions in (1.24) are in even powers of h. In this situation we multiply the first relation of (1.24) with $\dot q_h$, the second one with $\dot p_h$, and we subtract both so that the right-hand side becomes a total differential. This yields

$$\dot q_h^{\mathsf T}\big(1+\lambda_2^p h^2D^2+\ldots\big)\dot p_h - \dot p_h^{\mathsf T}\big(1+\lambda_2^q h^2D^2+\ldots\big)\dot q_h + \frac{d}{dt}H\big(p_h,q_h\big) = \mathcal{O}(h^N). \tag{1.26}$$

The main ingredient for a further simplification is the fact that

$$\dot q_h^{\mathsf T} p_h^{(2j+1)} - \dot p_h^{\mathsf T} q_h^{(2j+1)} = \frac{d}{dt}\Big(\sum_{l=1}^{2j}(-1)^{l+1} q_h^{(l)T} p_h^{(2j+1-l)}\Big) \tag{1.27}$$

is also a total differential. We now distinguish the following situations:

Case A: both multistep methods are identical. This case has been treated in Section XV.4.3 of [HLW06]. We have $\lambda_j^p = \lambda_j^q$ for all j, and it follows from (1.27) that the entire left-hand side of (1.26) is a total differential. Using the modified differential equation (1.18), first and higher derivatives of p_h and q_h can be substituted with expressions depending only on p_h and q_h. This proves the existence of functions $H_{2j}(p,q)$, such that after integration of (1.26)

$$H(p_h,q_h)+h^2H_2(p_h,q_h)+h^4H_4(p_h,q_h)+\ldots = \mathit{Const}+\mathcal{O}(th^N). \tag{1.28}$$

As long as the solution of the modified differential equation (i.e., the numerical solution) remains in a compact set, we thus have $H(p_h,q_h) = \mathit{Const}+\mathcal{O}(h^r)+\mathcal{O}(th^N)$, where r is the order of the method and N can be chosen arbitrarily large.

This is a nice result, but of limited interest. If the p and q components are discretized by the same multistep method, parasitic components are usually not under control and they destroy the long-time behavior of the underlying one-step method.

Case B: separable Hamiltonian with quadratic kinetic energy. This situation is treated in [HL04]. For a Hamiltonian of the form $H(p,q) = \frac{1}{2}p^{\mathsf T}M^{-1}p+U(q)$ (without loss of generality we assume $M = I =$ identity) we have $\nabla_p H(p,q) = p$. The second relation of (1.24) therefore permits to express p_h as a linear combination of odd derivatives of q_h. Inserted into (1.26), this gives rise to a linear combination of terms $q_h^{(m)T} q_h^{(2j+1-m)}$, which all can be written as total differentials because of

$$2\,q_h^{(m)T} q_h^{(2j+1-m)} = \frac{d}{dt}\Big(\sum_{l=m}^{2j-m}(-1)^{l-m} q_h^{(l)T} q_h^{(2j-l)}\Big). \tag{1.29}$$

Without any assumptions on the coefficients λ_j^p and λ_j^q, a modified Hamiltonian satisfying (1.28) can be obtained as in Case (A). This is an important result, because the parasitic components can be shown to remain bounded and small (see [HL04] and Chapter 1.3 below).

Case C: additional order conditions. If both multistep schemes are of order r, then $\lambda_j^p = \lambda_j^q = 0$ holds for $1 \le j < r$. Can we construct schemes, where the polynomials $\rho_p(\zeta)$ and $\rho_q(\zeta)$ have no common zeros other than $\zeta = 1$, such that $\lambda_j^p = \lambda_j^q$ also for $j = r$ (and possibly also for larger j)?

The class of explicit, symmetric 3-step methods of order $r = 2$ is given by

$$\rho(\zeta) = (\zeta - 1)(\zeta^2 + 2a\zeta + 1), \qquad \sigma(\zeta) = (a + 1)(\zeta^2 + \zeta),$$

where $|a| < 1$ by stability (for $a = 1$ it is reducible and equivalent to the 2-step explicit midpoint rule). The coefficient λ_2 in the expansion (1.25) is $\lambda_2 = \frac{1}{2}\left(\frac{1}{a+1} - \frac{1}{6}\right)$, and it is not possible to have the same λ_2 for different values of a.

Symmetric 5-step methods of order $r = 4$ are given by

$$\rho(\zeta) = (\zeta - 1)(\zeta^2 + 2a_1\zeta + 1)(\zeta^2 + 2a_2\zeta + 1),$$

where $|a_1| < 1$ and $|a_2| < 1$ (one of these coefficients is allowed to be equal to 1, but then the method reduces to a 4-stage method). The polynomial $\sigma(\zeta)$ is uniquely determined by assuming the method to be explicit and of order 4. In this case, the coefficient

$$\lambda_4 = \frac{131 - 19\,(a_1 + a_2) + 11\,a_1 a_2}{720\,(1 + a_1)(1 + a_2)}$$

in (1.25) depends on two parameters, and it is possible to construct different methods with the same value of λ_4. This happens, for example, when the coefficients $a_j^p = \cos\theta_j^p$ and $a_j^q = \cos\theta_j^q$ for the two ρ polynomials are given by

$$\text{plmm4c} \qquad \begin{array}{lll} \rho_p(\zeta): & \theta_1^p = \pi/8 & \theta_2^p = 3\pi/4 \\ \rho_q(\zeta): & \theta_1^q = 3\pi/8 & \theta_2^q \approx 0.68\,\pi \end{array}$$

(here, θ_1^p, θ_2^p, θ_1^q are arbitrarily fixed, and θ_2^q is computed to satisfy $\lambda_4^p = \lambda_4^q$). We apply this method to the three problems with separable Hamiltonian of Section 1.1.3. For problem (A) there is no difference to the behavior of methods plmm2 and plmm4. The error in the Hamiltonian is of size $\mathcal{O}(h^4)$ and no drift can be observed. Numerical results for problems (B) and (C) are presented in Figure 1.4. For problem (B) we expect that the dominant error term in the Hamiltonian remains bounded. In fact, experiments with many different values of the step size h indicate that the error in the Hamiltonian is bounded by $\mathcal{O}(h^4) + \mathcal{O}(th^6)$ on intervals of length $\mathcal{O}(h^{-2})$. Similarly, also for problem (C) the dominant error term remains bounded. In this case we expect the error to behave like $\mathcal{O}(h^4) + \mathcal{O}(\sqrt{t}h^6)$. The second term is invisible on intervals of length $\mathcal{O}(h^{-2})$, see also Table 1.1. Beyond such an interval, Figure 1.4 shows that for both problems, (B) and (C), the error behaves like $\delta\exp(ch^2t)$ with a small constant δ. This undesirable exponential error growth will be explained by studying parasitic solution components in Chapter 1.3.

1.2.4 Near preservation of quadratic first integrals

We again consider general differential equations (1.1) and we assume the existence of a quadratic first integral of the form $L(p, q) = p^{\mathsf{T}} E\, q$, i.e.,

$$f(p, q)^{\mathsf{T}} E\, q + p^{\mathsf{T}} E\, g(p, q) = 0 \qquad \text{for all } p \text{ and } q. \tag{1.30}$$

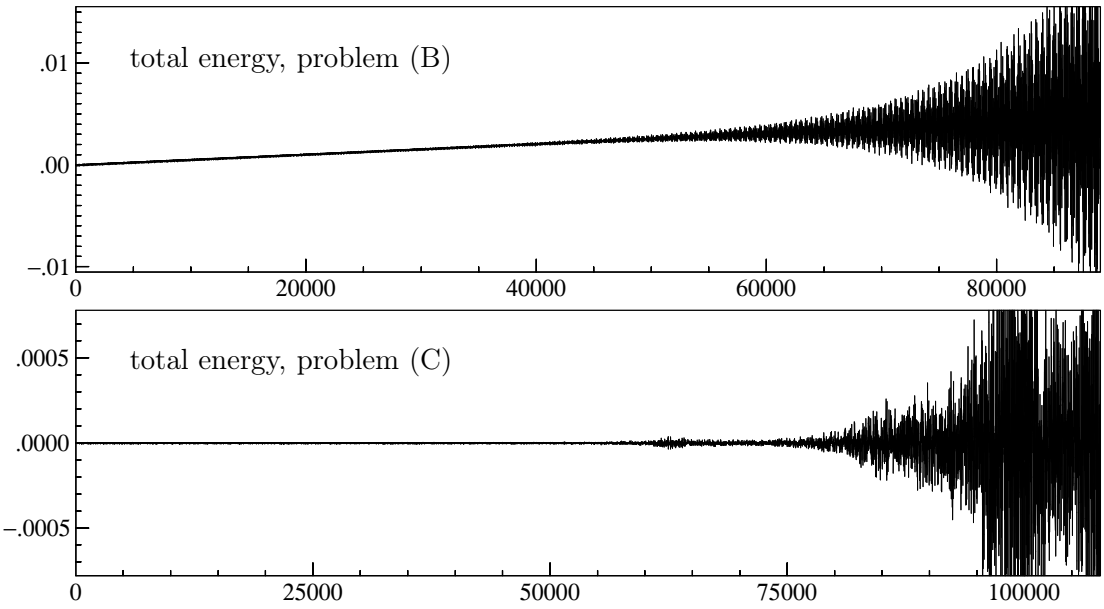


Figure 1.4 – Numerical Hamiltonian of method 'plmm4c' applied with step size $h = 0.005$ for problem (B), and with $h = 0.001$ for problem (C); initial values and starting approximations as in Figure 1.1.

The exact solution satisfies $L\big(p(t), q(t)\big) = \mathit{Const}$, and we are interested to know if the numerical approximation can mimic this behavior. As in the previous section we consider only smooth numerical approximations, which are formally equal to the values of the solution $\big(p_h(t), q_h(t)\big)$ at $t = nh$ of the modified differential equation. We therefore have to study the evolution of $L\big(p_h(t), q_h(t)\big)$.

Dividing the relations in (1.20) by h times the σ polynomials, the solution of the modified differential equation is seen to verify

$$\begin{aligned} \big(1 + \lambda_1^p hD + \lambda_2^p h^2 D^2 + \ldots\big)\dot{p}_h &= f(p_h, q_h) + \mathcal{O}(h^N) \\ \big(1 + \lambda_1^q hD + \lambda_2^q h^2 D^2 + \ldots\big)\dot{q}_h &= g(p_h, q_h) + \mathcal{O}(h^N), \end{aligned} \tag{1.31}$$

where the coefficients λ_j^p and λ_j^q are given by (1.25). We restrict our considerations to symmetric methods, so that the series are in even powers of h. We multiply the transposed first relation of (1.31) with Eq_h from the right, and the second one with $p_h^{\mathsf{T}}E$ from the left, and we add both so that by (1.30) the right-hand side becomes an expression of size $\mathcal{O}(h^N)$. We thus obtain

$$\big((1 + \lambda_2^p h^2 D^2 + \ldots)\dot{p}_h\big)^{\mathsf{T}} E\, q_h + p_h^{\mathsf{T}} E (1 + \lambda_2^q h^2 D^2 + \ldots)\, \dot{q}_h = \mathcal{O}(h^N). \tag{1.32}$$

An important simplification can be achieved by using the identity

$$(p_h^{2j+1})^{\mathsf{T}} E\, q_h + p_h^{\mathsf{T}} E\, q_h^{(2j+1)} = \frac{d}{dt}\Big(\sum_{l=0}^{2j} (-1)^l (p_h^{(2j-l)})^{\mathsf{T}} E\, q_h^{(l)} \Big) \tag{1.33}$$

As in the previous section we now distinguish the following situations:

Case A: both multistep methods are identical. This is the case considered in Section XV.4.4 of [HLW06]. We have $\lambda_j^p = \lambda_j^q$ for all j, and it follows from (1.33) that the expression in (1.32) is a total differential. As in Section 1.2.3, first and higher derivatives

of p_h and q_h can be substituted with expressions depending only on p_h and q_h. Hence, there exist functions $L_{2j}(p,q)$ with $L_0(p,q) = L(p,q) = p^{\mathsf{T}} E\, q$, such that after integration of (1.32)

$$L(p_h,q_h) + h^2 L_2(p_h,q_h) + h^4 L_4(p_h,q_h) + \ldots = \; Const + \mathcal{O}(th^N). \tag{1.34}$$

As long as the solution of the modified differential equation (i.e., the numerical solution) remains in a compact set, we thus have $L(p_h,q_h) = Const + \mathcal{O}(h^r) + \mathcal{O}(th^N)$, where r is the order of the method and N can be chosen arbitrarily large.

Note that such a result is not true in general for symmetric one-step methods. However, it is of limited interest, because parasitic components are usually not under control for the situation, where both multistep methods are identical.

Case B: special form of the differential equation. We consider problems of the form

$$\dot p = f(q), \qquad \dot q = M^{-1} p,$$

which are equivalent to second order differential equations $\ddot q = M^{-1} f(q)$. This corresponds to the situation treated in [HL04]. Without loss of generality we assume in the following that $M = I =$ identity). For such special differential equations the condition (1.30) splits into two conditions

$$f(q)^{\mathsf{T}} E\, q = 0, \qquad p^{\mathsf{T}} E\, p = 0 \qquad \text{for all } p \text{ and } q,$$

which implies that E is a skew-symmetric matrix. Moreover, because of $g(p,q) = p$, the second relation of (1.31) permits to express p_h as a linear combination of odd derivatives of q_h. Inserted into (1.32), this gives rise to a linear combination of terms $q_h^{(2m+1)T} E\, q_h^{(2j-2m+1)}$, which can be written as total differentials because

$$q_h^{(2m+1)T} E\, q_h^{(2j-2m+1)} = \frac{d}{dt}\Big(\sum_{l=2m+1}^{j} (-1)^{l-1} q_h^{(l)T} E\, q_h^{(2j-l+1)} \Big).$$

Without any assumptions on the coefficients λ_j^p and λ_j^q, a formal first integral of the form (1.34) is obtained that is $\mathcal{O}(h^r)$-close to the invariant $L(p,q) = p^{\mathsf{T}} E\; q$ of the differential equation. This result is important, because the parasitic components will be shown to remain bounded and small (see also [HL04]).

Case C: additional order conditions. If the partitioned multistep method is of order r, we have $\lambda_j^p = \lambda_j^q = 0$ for $1 \le j < r$. If the coefficients of the method are constructed such that $\lambda_j^p = \lambda_j^q$ also for $j = r$, we can apply the computation of case (A) to the leading error term. In this way an improved near conservation of quadratic first integrals can be achieved, similar to the near energy conservation in the previous section.

1.2.5 Symplecticity and conjugate symplecticity

In the numerical solution of Hamiltonian systems it is unavoidable to speak also about symplecticity. Together with the differential equation

$$\begin{aligned} \dot p &= -\nabla_q H(p,q), \\ \dot q &= \nabla_p H(p,q), \end{aligned} \tag{1.35}$$

whose flow we denote by $\varphi_t(p_0, q_0)$, we consider the variational differential equation

$$\begin{aligned} \dot{P} &= -\nabla_{qp} H(p,q)\, P - \nabla_{qq} H(p,q)\, Q, \\ \dot{Q} &= \nabla_{pp} H(p,q)\, P + \nabla_{pq} H(p,q)\, Q, \end{aligned} \tag{1.36}$$

where we use the notation $\nabla_{qp} H(p,q) = \big(\frac{\partial^2 H}{\partial q_i \partial p_j}\big)$. Here, $P(t)$ and $Q(t)$ are the derivatives with respect to initial values,

$$P(t) = \Big(\frac{\partial p(t)}{\partial p_0}, \frac{\partial p(t)}{\partial q_0}\Big), \quad Q(t) = \Big(\frac{\partial q(t)}{\partial p_0}, \frac{\partial q(t)}{\partial q_0}\Big) \quad \text{and} \quad \varphi'_t(p_0,q_0) = \begin{pmatrix} P(t) \\ Q(t) \end{pmatrix}.$$

The flow map $\varphi_t(p_0, q_0)$ of (1.35) is a symplectic transformation, see e.g., [HLW06, VI.2]. This means, by definition, that its Jacobian matrix satisfies

$$\varphi'_t(p_0,q_0)^{\mathsf{T}} J\, \varphi'_t(p_0,q_0) = J \quad \text{or equivalently} \quad P(t)^{\mathsf{T}} Q(t) - Q(t)^{\mathsf{T}} P(t) = J,$$

where J is the canonical structure matrix already encountered in (1.16). The important observation is that symplecticity just means that $P^{\mathsf{T}}Q - Q^{\mathsf{T}}P$ is a quadratic first integral of the combined system (1.35)-(1.36).

The smooth numerical solution of a partitioned multistep method is formally equal to the exact solution of the modified differential equation of Theorem 1.2.1. We therefore call the multistep method *symplectic*, if the derivative $(P_h(t), Q_h(t))$ (with respect to initial values) of the solution $(p_h(t), q_h(t))$ of the modified differential equation (1.18) satisfies

$$P_h(t)^{\mathsf{T}} Q_h(t) - Q_h(t)^{\mathsf{T}} P_h(t) = J.$$

Unfortunately, this is never satisfied unless for some trivial exceptions (implicit mid-point rule, symplectic Euler method, and the Störmer–Verlet scheme) which are partitioned linear multistep methods and one-step methods at the same time. Intuitively this is clear from the considerations of Section 1.2.4, because we did not encounter any result on the exact preservation of quadratic first integrals. A rigorous proof of this negative result has first been given by Tang [Tan93] (see also [HLW06, Sect. XV.4]).

In view of this negative result, it is natural to consider a weaker property than symplecticity, which nevertheless retains the same qualitative long-time behavior. We call a matrix-valued mapping $\Phi_h : (p,q) \mapsto (P,Q)$ *conjugate symplectic*, if there exists a global change of coordinates $(\widehat{p}, \widehat{q}) = \chi_h(p,q)$ that is $\mathcal{O}(h^r)$-close to the identity, such that the mapping is symplectic in the new coordinates, i.e., the mapping $\widehat{\Phi}_h = \chi_h \circ \Phi_h \circ \chi_h^{-1}$ is a symplectic transformation. Since

$$\widehat{\Phi}'_h(\widehat{p},\widehat{q}) = \Phi'_h(p,q) + h^r K_r(p,q) + h^{r+1} K_{r+1}(p,q) + \ldots,$$

the symplecticity of $\widehat{\Phi}_h$ yields the existence of functions $L_j(p,q)$ such that

$$\Phi'_h(p,q)^{\mathsf{T}} J\, \Phi'_h(p,q) + h^r L_r(p,q) + h^{r+1} L_{r+1}(p,q) + \ldots = J. \tag{1.37}$$

This means that for a method that is conjugate symplectic, there exists a modified first integral (as a formal series in powers of h) of the modified differential equation which is $\mathcal{O}(h^r)$-close to $P_h^{\mathsf{T}} Q_h - Q_h^{\mathsf{T}} P_h = (\Phi'_h)^{\mathsf{T}} J\, \Phi'_h$.

If Φ_h represents the underlying one-step method of a partitioned multistep method, we know from Section 1.2.4 that under suitable assumptions there exist functions $L_j(p,q)$ such that (1.37) holds. Does this imply that the method Φ_h is conjugate symplectic? That this is indeed the case follows from results of Chartier, Faou, and Murua [CFM06], see also [HLW06, Section XV.4.4]. We do not pursue this question in the present work.

1.3 Long-term stability of parasitic solution components

We consider the partitioned linear multistep method (1.2) applied to the differential equation (1.1). We assume that both multistep methods are symmetric and stable, so that the zeros of the polynomials $\rho_p(\zeta)$ and $\rho_q(\zeta)$ are all on the unit circle. We denote these zeros by $\zeta_0 = 1$, and ζ_j, $\zeta_{-j} = \overline{\zeta}_j$ for $j = 1, \dots, \kappa$ (if -1 is such a zero, we let $\zeta_{-\kappa} = \zeta_\kappa = -1$). Furthermore, we consider finite products of the zeros of the ρ-polynomials, which we again denote by ζ_j and $\zeta_{-j} = \overline{\zeta}_j$. The resulting index set is denoted by $\mathcal{I}$, so that

$$\{\zeta_l\}_{l\in\mathcal{I}} = \{\zeta = \zeta_1^{m_1} \cdot \ldots \cdot \zeta_\kappa^{m_\kappa}\,;\ m_j \geq 0\}.$$

The index set can be finite (if all zeros of the ρ-polynomials are roots of unity) or it can be infinite. It is convenient to denote $\mathcal{I}^* = \mathcal{I} \setminus \{0\}$.

Our aim is to write the numerical solution of (1.2) in the form

$$\begin{pmatrix} p_n \\ q_n \end{pmatrix} = \begin{pmatrix} p(t_n) \\ q(t_n) \end{pmatrix} + \sum_{l\in\mathcal{I}^*} \zeta_l^n \begin{pmatrix} u_l(t_n) \\ v_l(t_n) \end{pmatrix}, \tag{1.38}$$

where $t_n = nh$. Here, $(p(t), q(t))$ is an h-dependent approximation to the exact solution of (1.1), called *principal solution component*. To avoid any confusion, we denote in this chapter the exact solution of (1.1) as $(p_{exact}(t), q_{exact}(t))$. The functions $(u_l(t), v_l(t))$ also depend on the step size h, and they are called *parasitic solution components*. This chapter is devoted to get bounds on these parasitic solution components and to investigate the length of time intervals, where the parasitic components do not significantly perturb the principal solution component.

A similar representation of the numerical solution has been encountered when discussing the numerical solution for the harmonic oscillator in Section 1.2.1. There, only zeros of the ρ-polynomials are present in the sum. The appearance of products of such zeros in (1.38) is due to the nonlinearity of the vector field in (1.1).

1.3.1 Modified differential equation (full system)

We first study the existence of the coefficient functions in the representation (1.38). This is an extension of the backward error analysis of the smooth numerical solution as discussed in Section 1.2.2. It follows closely the presentation of [HLW06, Sect. XV.3.2]. In the following we use the notations $y(t) = (p(t), q(t))$, $z_l(t) = (u_l(t), v_l(t))$, and we collect in the vector $\mathbf{z}(t)$ the components $u_l(t)$ $(l \neq 0)$ for which $\rho_p(\zeta_l) = 0$ and the components $v_l(t)$ $(l \neq 0)$ for which $\rho_q(\zeta_l) = 0$.

Theorem 1.3.1. *Consider a consistent, symmetric, partitioned linear multistep method (1.2), applied to the differential equation (1.1). Then, there exist h-independent functions $f_j(p,q,\mathbf{z})$, $g_j(p,q,\mathbf{z})$, and $f_{l,j}(p,q,\mathbf{z})$, $g_{l,j}(p,q,\mathbf{z})$, such that for an arbitrarily chosen truncation index N and for every solution $p(t), q(t), u_l(t), v_l(t)$ of the system*

$$\begin{aligned}
\dot p &= f(p,q) + hf_1(p,q,\mathbf{z}) + \ldots + h^{N-1} f_{N-1}(p,q,\mathbf{z}) \\
\dot q &= g(p,q) + hg_1(p,q,\mathbf{z}) + \ldots + h^{N-1} g_{N-1}(p,q,\mathbf{z}) \\
\dot u_l &= f_{l,0}(p,q,\mathbf{z}) + hf_{l,1}(p,q,\mathbf{z}) + \ldots + h^{N-1} f_{l,N-1}(p,q,\mathbf{z}) \qquad \text{if} \quad \rho_p(\zeta_l) = 0 \\
\dot v_l &= g_{l,0}(p,q,\mathbf{z}) + hg_{l,1}(p,q,\mathbf{z}) + \ldots + h^{N-1} g_{l,N-1}(p,q,\mathbf{z}) \qquad \text{if} \quad \rho_q(\zeta_l) = 0 \\
u_l &= hf_{l,1}(p,q,\mathbf{z}) + \ldots + h^N f_{l,N}(p,q,\mathbf{z}) \qquad \text{if} \quad \rho_p(\zeta_l) \neq 0 \\
v_l &= hg_{l,1}(p,q,\mathbf{z}) + \ldots + h^N g_{l,N}(p,q,\mathbf{z}) \qquad \text{if} \quad \rho_q(\zeta_l) \neq 0 \\
u_l &= 0, \quad v_l = 0 \qquad \text{if} \quad \zeta_l \neq \zeta_1^{m_1} \cdot \ldots \cdot \zeta_\kappa^{m_\kappa} \quad \text{with} \quad m_1 + \ldots + m_\kappa < N,
\end{aligned} \tag{1.39}$$

with initial values $\mathbf{z}(0) = \mathcal{O}(h)$*, the function (with* $n = t/h$*)*

$$\begin{pmatrix} p_h(t) \\ q_h(t) \end{pmatrix} = \begin{pmatrix} p(t) \\ q(t) \end{pmatrix} + \sum_{l \in \mathcal{I}^*} \zeta_l^n \begin{pmatrix} u_l(t) \\ v_l(t) \end{pmatrix}, \tag{1.40}$$

satisfies the multistep formula up to a defect of size $\mathcal{O}(h^{N+1})$*, i.e.,*

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j^p p_h(t+jh) &= h \sum_{j=0}^{k} \beta_j^p f\big(p_h(t+jh), q_h(t+jh)\big) + \mathcal{O}(h^{N+1}) \\ \sum_{j=0}^{k} \alpha_j^q q_h(t+jh) &= h \sum_{j=0}^{k} \beta_j^q g\big(p_h(t+jh), q_h(t+jh)\big) + \mathcal{O}(h^{N+1}) \end{aligned} \tag{1.41}$$

as long as $(p(t), q(t))$ *remain in a compact set, and* $\|\mathbf{z}(t)\| \le Ch$*. The constant symbolized by* $\mathcal{O}$ *is independent of* h*, but depends on the truncation index* N*. It also depends smoothly on* t*. If the partitioned multistep method is of order* r*, then we have* $f_l(p,q) = g_l(p,q) = 0$ *for* $1 \le l < r$*.*

Remark 1.3.2. Because of the last line in (1.39), the sum in (1.40) is always finite. Substituting $\mathbf{z} = 0$ in the upper two equations of (1.39) yields the modified differential equation (1.18) of Section 1.2.2. The solution of the system (1.39) satisfies $u_{-l}(t) = \overline{u}_l(t)$, $v_{-l}(t) = \overline{v}_l(t)$, whenever these relations hold for the initial values.

Proof. The proof is very similar to that of Theorem 1.2.1, and we highlight here only the main differences. We insert the finite sum (1.40) into (1.41), we expand the nonlinearities around $(p(t), q(t))$, which we also denote by $(u_0(t), v_0(t))$, and we compare the coefficients of ζ_j^n. This yields, recalling that $y(t) = (p(t), q(t)) = (u_0(t), v_0(t))$ and $z_l(t) = (u_l(t), v_l(t))$, and omitting the argument t,

$$\begin{aligned} \rho_p(\zeta_l e^{hD}) u_l &= h\, \sigma_p(\zeta_l e^{hD}) \sum_{m \ge 0} \frac{1}{m!} \sum_{\zeta_{l_1} \cdots \zeta_{l_m} = \zeta_l} f^{(m)}\big(y\big)\big(z_{l_1}, \ldots, z_{l_m}\big) + \mathcal{O}(h^{N+1}), \\ \rho_q(\zeta_l e^{hD}) v_l &= h\, \sigma_q(\zeta_l e^{hD}) \sum_{m \ge 0} \frac{1}{m!} \sum_{\zeta_{l_1} \cdots \zeta_{l_m} = \zeta_l} g^{(m)}\big(y\big)\big(z_{l_1}, \ldots, z_{l_m}\big) + \mathcal{O}(h^{N+1}), \end{aligned} \tag{1.42}$$

where the second sum is over indices $l_1 \ne 0, \ldots, l_m \ne 0$. The summand for $m = 0$, which is $f(y(t))$, resp. $g(y(t))$, is present only for $l = 0$, i.e., for $\zeta_l = 1$. Notice further that for $l = 0$ the summand for $m = 1$ vanishes, because we always have $\zeta_{l_1} \ne \zeta_0$. In view of an inversion of the operators $\rho_p(\zeta_l e^{hD})$ and $\rho_q(\zeta_l e^{hD})$ we introduce the coefficients of the expansions (cf. equation (1.21) for $\zeta_0 = 1$)

$$\frac{x\, \sigma_p(\zeta_l\, e^x)}{\rho_p(\zeta_l\, e^x)} = \mu_{l0}^p + \mu_{l1}^p x + \mu_{l2}^p x^2 + \ldots, \qquad \frac{x\, \sigma_q(\zeta_l\, e^x)}{\rho_q(\zeta_l\, e^x)} = \mu_{l0}^q + \mu_{l1}^q x + \mu_{l2}^q x^2 + \ldots\,. \tag{1.43}$$

If $\rho_p(\zeta_l) \ne 0$, we have $\mu_{l0}^p = 0$. If $\rho_p(\zeta_l) = 0$, the expansion exists because ζ_l is a simple zero, and we have $\mu_{l0}^p \ne 0$ because $\sigma_p(\zeta_l) \ne 0$ as a consequence of the irreducibility of the method. The same statements hold for the second method. We therefore obtain the differential equations

$$\begin{aligned} \dot{u}_l &= \big(\mu_{l0}^p + \mu_{l1}^p hD + \ldots\big) \sum_{m \ge 0} \frac{1}{m!} \sum_{\zeta_{l_1} \cdots \zeta_{l_m} = \zeta_l} f^{(m)}\big(y\big)\big(z_{l_1}, \ldots, z_{l_m}\big) + \mathcal{O}(h^N), \\ &\qquad \text{if } \ \rho_p(\zeta_l) = 0, \\ \dot{v}_l &= \big(\mu_{l0}^q + \mu_{l1}^q hD + \ldots\big) \sum_{m \ge 0} \frac{1}{m!} \sum_{\zeta_{l_1} \cdots \zeta_{l_m} = \zeta_l} g^{(m)}\big(y\big)\big(z_{l_1}, \ldots, z_{l_m}\big) + \mathcal{O}(h^N), \\ &\qquad \text{if } \ \rho_q(\zeta_l) = 0, \end{aligned} \tag{1.44}$$

and the algebraic relations

$$\begin{aligned} u_l &= \big(\mu^p_{l1} hD + \mu^p_{l2} h^2 D^2 + \ldots\big) \sum_{m\ge 1} \frac{1}{m!} \sum_{\zeta_{l_1}\cdots\zeta_{l_m}=\zeta_l} f^{(m)}\big(y\big)\big(z_{l_1},\ldots,z_{l_m}\big) + \mathcal{O}(h^{N+1}), \\ &\qquad \text{if } \ \rho_p(\zeta_l)\neq 0, \\ v_l &= \big(\mu^q_{l1} hD + \mu^q_{l2} h^2 D^2 + \ldots\big) \sum_{m\ge 1} \frac{1}{m!} \sum_{\zeta_{l_1}\cdots\zeta_{l_m}=\zeta_l} g^{(m)}\big(y\big)\big(z_{l_1},\ldots,z_{l_m}\big) + \mathcal{O}(h^{N+1}), \\ &\qquad \text{if } \ \rho_q(\zeta_l)\neq 0. \end{aligned} \tag{1.45}$$

As in the proof of Theorem 1.2.1 we use (1.44) to recursively eliminate first and higher derivatives of u_l if $\rho_p(\zeta_l)=0$ and of v_l if $\rho_q(\zeta_l)=0$. Similarly, we use (1.45) to recursively eliminate u_l and its derivatives if $\rho_p(\zeta_l)\neq 0$ and of v_l and its derivatives if $\rho_q(\zeta_l)\neq 0$. Collecting equal powers of h yields the functions $f_j(p,q,\mathbf{z})$, $g_j(p,q,\mathbf{z})$, and $f_{l,j}(p,q,\mathbf{z})$, $g_{l,j}(p,q,\mathbf{z})$.

If $\zeta_l \neq \zeta_1^{m_1}\cdots\zeta_\kappa^{m_\kappa}$ with $m_1+\ldots+m_\kappa < N$, the right-hand side of (1.45) contains at least N factors of components of $\mathbf{z}$. By our assumption $\|\mathbf{z}(t)\| \le Ch$, this implies $u_l = \mathcal{O}(h^{N+1})$ and $v_l = \mathcal{O}(h^{N+1})$, so that these functions can be included in the remainder term. This justifies the last line of (1.39) and concludes the proof of the theorem. □

Initial values for the system (1.39). For an application of the multistep formula (1.2), starting approximations (p_j, q_j) for $j=0,\ldots,k-1$ have to be provided. We assume that they satisfy (with $0\le\nu\le r$)

$$p_j - p_{exact}(jh) = \mathcal{O}(h^{\nu+1}), \quad q_j - q_{exact}(jh) = \mathcal{O}(h^{\nu+1}), \quad j=0,\ldots,k-1. \tag{1.46}$$

Initial values for the differential equation (1.39) have to be such that

$$\begin{pmatrix} p_j \\ q_j \end{pmatrix} = \begin{pmatrix} p(jh) \\ q(jh) \end{pmatrix} + \sum_{l\in\mathcal{I}^*} \zeta_l^j \begin{pmatrix} u_l(jh) \\ v_l(jh) \end{pmatrix}, \qquad j=0,\ldots,k-1. \tag{1.47}$$

The solution of (1.39) is uniquely determined by the initial values $y(0), \mathbf{z}(0)$ (for the notation of y and $\mathbf{z}$ see the beginning of Section 1.3.1), so that the system (1.47) can be written as $F\big(y(0),\mathbf{z}(0),h\big)=0$. For $h=0$, it represents a linear Vandermonde system for $y(0),\mathbf{z}(0)$, which gives a unique solution. The Implicit Function Theorem thus proves the local existence of a solution of $F\big(y(0),\mathbf{z}(0),h\big)=0$ for sufficiently small step sizes h. Note that the initial values depend smoothly on h. Under the assumption (1.46) we have $p(0)=p_{exact}(0)+\mathcal{O}(h^{\nu+1})$, $q(0)=q_{exact}(0)+\mathcal{O}(h^{\nu+1})$, and $\mathbf{z}(0)=\mathcal{O}(h^{\nu+1})$.

1.3.2 Growth parameters

Before attacking the question of bounding rigorously the parasitic solution components, we try to get a feeling of the solution of the system (1.39). This system is equivalent to the equations (1.44) and (1.45). Our aim is to have small parasitic solution components. We therefore neglect all terms that are at least quadratic in $\mathbf{z}$.

The equations (1.44) for $l=0$ (principal solution components) become equivalent to the modified equation already studied in Chapter 1.2. If we consider only the leading (h-independent) term in the expansion (1.45), we get zero functions. All that remains are the equations (1.44) with $l\neq 0$ which, for $h=0$, are as follows:

- if ζ_l is a common zero of $\rho_p(\zeta)$ and $\rho_q(\zeta)$, we have

$$\begin{aligned} \dot{u}_l &= \mu_{l0}^p \big(f_p(p(t),q(t))\, u_l + f_q(p(t),q(t))\, v_l \big) \\ \dot{v}_l &= \mu_{l0}^q \big(g_p(p(t),q(t))\, u_l + g_q(p(t),q(t))\, v_l \big), \end{aligned} \tag{1.48}$$

- if ζ_l is a zero of $\rho_p(\zeta)$, but $\rho_q(\zeta_l) \neq 0$, we have

$$\dot{u}_l = \mu_{l0}^p\, f_p(p(t),q(t))\, u_l, \tag{1.49}$$

- if ζ_l is a zero of $\rho_q(\zeta)$, but $\rho_p(\zeta_l) \neq 0$, we have

$$\dot{v}_l = \mu_{l0}^q\, g_q(p(t),q(t))\, v_l. \tag{1.50}$$

The coefficient $\mu_l = \mu_{l0}$ is called *growth parameter* of a multistep method with generating polynomials $\rho(\zeta)$ and $\sigma(\zeta)$. It is defined by (1.43) for the limit $x \to 0$, and can be computed from

$$\mu_l = \frac{\sigma(\zeta_l)}{\zeta_l\, \rho'(\zeta_l)}.$$

We remark that for a symmetric linear multistep method the growth parameter is always real. This follows from $\sigma(1/\zeta_l) = \zeta_l^k \sigma(\zeta_l)$ and $-\zeta_l^{-2}\rho'(\zeta_l) = \zeta_l^k \rho'(\zeta_l)$, which is obtained by differentiation of the relation $\rho(1/\zeta) = \zeta^k \rho(\zeta)$.

Already when we use for $(p(t), q(t))$ the exact solution of the original problem, the equations (1.48)-(1.50) give much insight into the behavior of the multistep method. For example, if we consider the harmonic oscillator, for which $f(p,q) = -q$, $g(p,q) = p$, the differential equation (1.48) gives bounded solutions only if the product of the growth parameters of both methods satisfy $\mu_l^p \mu_l^q > 0$ for all l. For nonlinear problems, the differential equation (1.48) has bounded solutions only in very exceptional cases.

If the polynomials $\rho_p(\zeta)$ and $\rho_q(\zeta)$ do not have common zeros with the exception of $\zeta_0 = 1$, the situation with equation (1.48) cannot arise. Therefore, only the equations (1.49) and (1.50) are relevant. There are many interesting situations, where the solutions of these equations are bounded, e.g., if $f(p,q)$ only depends on q and $g(p,q)$ only depends on p, what is the case for Hamiltonian systems with separable Hamiltonian.

1.3.3 Bounds for the parasitic solution components

We study the system (1.39) of modified differential equations. We continue to use the notation $y = (p,q)$ and, as in Section 1.3.1, we denote by $\mathbf{z}(t)$ the vector whose components are $u_l(t)$ $(l \neq 0)$ for which $\rho_p(\zeta_l) = 0$ and $v_l(t)$ $(l \neq 0)$ for which $\rho_q(\zeta_l) = 0$. The system (1.39) can then be written in compact notations as

$$\begin{aligned} \dot{y} &= F_{h,N}(y) + G_{h,N}(y,\mathbf{z}) \\ \dot{\mathbf{z}} &= A_{h,N}(y)\,\mathbf{z} + B_{h,N}(y,\mathbf{z}), \end{aligned} \tag{1.51}$$

where $G_{h,N}(y,\mathbf{z})$ and $B_{h,N}(y,\mathbf{z})$ collect those terms that are quadratic or of higher order in $\mathbf{z}$. Note that, by the construction via the system (1.44), the differential equation for y does not contain any linear term in $\mathbf{z}$.

We consider a compact subset K_0 of the $y = (p,q)$ phase space, and for a small positive parameter δ we define

$$K = \big\{ (y,\mathbf{z})\,;\; y \in K_0, \|\mathbf{z}\| \leq \delta \big\}. \tag{1.52}$$

Regularity of the (original) differential equation implies that there exists a constant L such that

$$\|G_{h,N}(y,\mathbf{z})\| \le L\,\|\mathbf{z}\|^2, \qquad \|B_{h,N}(y,\mathbf{z})\| \le L\,\|\mathbf{z}\|^2 \qquad \text{for} \quad (y,\mathbf{z}) \in K. \tag{1.53}$$

Our aim is to get bounds on the parasitic solution components $\mathbf{z}(t)$, which then allow to get information on the long-time behavior of partitioned linear multistep methods. To this end, we consider the simplified system

$$\begin{aligned} \dot y &= F_{h,N}(y), \\ \dot{\mathbf{z}} &= A_{h,N}(y)\,\mathbf{z}, \end{aligned} \tag{1.54}$$

where quadratic and higher order terms of $\mathbf{z}$ have been removed from (1.51). The differential equation for y is precisely the modified differential equation for the smooth numerical solution (Section 1.2.2). The differential equation for $\mathbf{z}$ is linear with coefficients depending on time t through the solution $y(t)$. Its dominant h-independent term is the differential equation studied in Section 1.3.2.

In the case of linear multistep methods for second order Hamiltonian systems, a formal invariant of the full system (1.51) has been found that is close to $\|\mathbf{z}\|$ (see [HL04] or [HLW06, Sect. XV.5.3]; the ideas are closely connected to the study of adiabatic invariants in highly oscillatory differential equations [HL01]). This was the key for getting bounds of the parasitic solution components on time intervals that are much longer than the natural time scale of the system (1.54). Here, we include the existence of such a formal invariant in an assumption ('S' for stability and 'I' for invariant), and we later discuss situations, where it is satisfied.

Stability assumption (SI). *We say that a partitioned linear multistep method (1.2) applied to a partitioned differential equation (1.1) satisfies the stability assumption (SI), if there exists a smooth function $I_{h,N}(y,\mathbf{z})$ such that, for $0 < h \le h_0$,*

- *the invariance property*

$$I_{h,N}\big(y(h),\mathbf{z}(h)\big) = I_{h,N}\big(y(0),\mathbf{z}(0)\big) + \mathcal{O}(h^{M+1}\|z(0)\|^2)$$

  holds for solutions of the differential equation (1.54), for which $(y(t),\mathbf{z}(t)) \in K$ for t in the interval $0 \le t \le h$;

- *there exists a constant $C \ge 1$, such that*

$$I_{h,N}(y,\mathbf{z}) \le \|\mathbf{z}\|^2 \le C\,I_{h,N}(y,\mathbf{z}) \quad \text{for} \quad (y,\mathbf{z}) \in K.$$

We are interested in situations, where the stability assumption (SI) is satisfied with $M > 0$, and we obviously focus on situations which admit a large M.

Lemma 1.3.3. *Under the stability assumption (SI) we have, for $0 < h \le h_0$,*

$$I_{h,N}\big(y(h),\mathbf{z}(h)\big) = I_{h,N}\big(y(0),\mathbf{z}(0)\big) + \mathcal{O}(h^{M+1}\|\mathbf{z}(0)\|^2) + \mathcal{O}(h\,\delta\|\mathbf{z}(0)\|^2)$$

along solutions of the complete system (1.51) of modified differential equations, provided that they stay in the compact set K for $0 \le t \le h$.

Proof. The defect of the solution $(y(t),\mathbf{z}(t))$ of (1.51), when inserted into (1.54), is bounded by $\mathcal{O}(\|\mathbf{z}(0)\|^2)$. An application of the Gronwall Lemma therefore proves that the difference of the solutions of the two systems with identical initial values is bounded by $\mathcal{O}(h\|\mathbf{z}(0)\|^2)$. The statement then follows from the mean value theorem applied to the function $I_{h,N}(y,\mathbf{z})$ and from the fact that the derivative still contains a factor of $\mathbf{z}$. □

We are now able to state and prove the main result of this chapter. It tells us the length of the integration interval, on which the parasitic solution components do not destroy the long-time behavior of the underlying one-step method.

Theorem 1.3.4. *In addition to the stability assumption (SI) we require that*

(A1) *the partitioned linear multistep method (1.2) is symmetric, of order r, and the generating polynomials $\rho_p(\zeta)$ and $\rho_q(\zeta)$ do not have common zeros with the exception of $\zeta = 1$;*

(A2) *the vector field of (1.1) is defined and analytic in an open neighborhood of a compact set K_1;*

(A3) *the numerical solution $y_n = (p_n, q_n)$ stays for all n with $0 \le nh \le T_0$ in a compact set $K_0 \subset K_1$ which has positive distance from the boundary of K_1;*

(A4) *the starting approximations $(p_j, q_j), j = 0, \dots, k-1$ are such that the initial values for the full modified differential equation (1.51) satisfy $y(0) \in K_0$, and $\|\mathbf{z}(0)\| \le \delta/\sqrt{2eC}$ with C from the stability assumption (SI) and $\delta = \mathcal{O}(h)$.*

For sufficiently small h and δ and for a fixed truncation index N, chosen large enough such that $h^N \le \max(h^M\delta, \delta^2)$, there exist constants c_1, c_2 and functions $y(t), z_l(t)$ on an interval of length

$$T = \min\big(T_0,\, c_1\,\delta^{-1},\, c_2\,h^{-M}\big), \tag{1.55}$$

such that

- *the numerical solution satisfies $y_n = y(nh) + \sum_{l\in\mathcal{I}^*} \zeta_l^n\, z_l(nh)$ for $0 \le nh \le T$;*
- *on every subinterval $[mh, (m+1)h)$, the functions $y(t), z_l(t)$ are a solution of the system (1.51);*
- *at the time instants $t_m = mh$ the functions $y(t), z_l(t)$ have jump discontinuities of size $O(h^{N+1})$;*
- *the parasitic solution components are bounded: $\|\mathbf{z}(t)\| \le \delta$ for $0 \le nh \le T$.*

Proof. The proof closely follows that of Theorem 8 in the publication [HL04], see also [HLW06, Sect. XV.5.3]. We separate the integration interval into subintervals of length h. On a subinterval $[mh, (m+1)h)$ we define the functions $y(t) = (p(t), q(t))$ and $z_l(t) = (u_l(t), v_l(t))$ as the solution of the system (1.39) with initial values such that (1.47) holds with $j = m, m+1, \dots, m+k-1$. It follows from Theorem 1.3.1, formula (1.41), that $y_{m+k} - y(t_{m+k}) = \mathcal{O}(h^{N+1})$. Consequently, the construction of initial values for the next subinterval $[(m+1)h, (m+2)h)$ yields for the functions $y(t)$ and $z_l(t)$ a jump discontinuity at t_{m+1} that is bounded by $\mathcal{O}(h^{N+1})$.

We now study how well the expression $I_{h,N}(y(t), \mathbf{z}(t))$ is preserved on long time intervals. Lemma 1.3.3 gives a bound on the maximal deviation within a subinterval of length h. Together with the $\mathcal{O}(h^{N+1})$ bound on the jump discontinuities at t_m this proves for $I_m = I_{h,N}(y(t_m), \mathbf{z}(t_m))$ the estimate

$$I_{m+1} = I_m\big(1 + C_1 h^{M+1} + C_2 h\,\delta\big) + C_3 h^{N+1}\delta$$

as long as $(y(t), \mathbf{z}(t))$ remains in K. With $\gamma = C_1 h^M + C_2\delta$ the discrete Gronwall Lemma thus yields

$$I_m = I_0(1+\gamma\,h)^m + \frac{(1+\gamma\,h)^m - 1}{\gamma\,h}\, C_3 h^{N+1}\delta,$$

which, for $\gamma\, t_m \le 1$, gives the estimate $I_m \le I_0\, e + C_3(e-1)h^N \delta\, t_m$. This implies

$$\|\mathbf{z}(t)\|^2 \le C\, e\, \|\mathbf{z}(0)\|^2 + C_4 h^N \delta\, t,$$

so that $\|\mathbf{z}(t)\| \le \delta$ for times t subject to $\gamma\, t \le 1$, if the truncation index N is chosen sufficiently large. □

It is straight-forward to construct partitioned linear multistep methods of high order satisfying (A1). The assumption (A2) is satisfied for many important differential equations. The assumption (A3) can be checked a posteriori. If the method is of order r and if the starting approximations are computed with very high precision, then assumption (A4) is fulfilled with $\delta = \mathcal{O}(h^{r+1})$. This follows from the construction of the initial values for the system (1.39) as explained in the end of Section 1.3.1. The difficult task is the verification of the stability assumption (SI).

1.3.4 Near energy conservation

Combining our results on the long-time behavior of smooth numerical solutions with the bounded-ness of parasitic solution components we obtain the desired statements on the preservation of energy and of quadratic first integrals.

The near energy preservation has been studied analytically in Section 1.2.3 for smooth numerical solutions of symmetric partitioned multistep methods. We consider methods which, when applied to Hamiltonian systems, have a modified energy

$$H_h(p,q) = H(p,q) + h^r H_r(p,q) + \ldots + h^{N-1} H_{N-1}(p,q), \tag{1.56}$$

where r is the order of the method and $N > r$, such that

$$H_h(p_h, q_h) = \mathit{Const} + \mathcal{O}(t\, h^N) \tag{1.57}$$

along solutions of the modified differential equation (1.18). There are situations (cases (A) and (B) of Section 1.2.3), where N is arbitrarily large. This is the best behavior we can hope for. In the case (C) of Section 1.2.3 we achieve $N = r + 2$. The worst behavior is when $N = r$, in which case a linear drift for the numerical Hamiltonian is present from the beginning. This behavior of smooth numerical solutions carries over to the general situations as follows:

Theorem 1.3.5. *Consider a partitioned linear multistep method (1.2) of order r, applied to a Hamiltonian system (1.23). Assume that there exists a modified energy (1.56) such that (1.57) holds for smooth numerical solutions.*

Under the assumptions of Theorem 1.3.4 with $\delta = \mathcal{O}(h^r)$, the numerical solution satisfies

$$H(p_n, q_n) = \mathit{Const} + \mathcal{O}(h^r) \qquad \text{for} \qquad nh \le T,$$

where the length of the time interval T is limited by (1.55) and by $T \le \mathcal{O}(h^{r-N})$.

Proof. Let $y(t) = (p(t), q(t))$ and $z_l(t)$ (for $t_m \le t \le t_{m+1}$, $t_m = mh$) be a solution of the complete system (1.51) as in the statement of Theorem 1.3.4. Applying the proof of Lemma 1.3.3 to the near invariant $H_h(p,q)$ yields

$$H_h\big(p(t_{m+1}), q(t_{m+1})\big) = H_h\big(p(t_m), q(t_m)\big) + \mathcal{O}(h\delta^2) + \mathcal{O}(h^{N+1}).$$

Since the jump discontinuities at the grid points t_m can be neglected, we obtain by following the proof of Theorem 1.3.4 that

$$H_h(p_n, q_n) = H_h(p_0, q_0) + \mathcal{O}(t_n \delta^2) + \mathcal{O}(t_n h^N),$$

so that the statement follows from (1.56) and the requirement $\delta = \mathcal{O}(h^r)$. □

Analogous statements are obtained for the near conservation of quadratic first integrals. In this case the results of Section 1.2.4 have to be combined with the bounded-ness of the parasitic solution components (Theorem 1.3.4).

1.3.5 Verification of the stability assumption (SI)

It remains to study the stability assumption (SI), and to investigate how large the number M in the invariance property can be. The nice feature is that we only have to consider the simplified system (1.54), where the subsystem for the principle solution component y is separated from the parasitic solution components. Therefore, the differential equation for $\mathbf{z}$ is a linear differential equation with coefficients depending on t via the principle solution $y(t)$. Another nice feature is that we are concerned only with a local result (estimates on an interval of length h which is the step size of the integrator).

The linear system $\dot{\mathbf{z}} = A_{h,N}(y(t))\,\mathbf{z}$ is obtained from (1.42), where terms are neglected that are either at least quadratic in $\mathbf{z}$ or contain a sufficiently high power of h. We consider $\zeta_l \neq 1$ satisfying $\rho_p(\zeta_l) = 0$ and $\rho_q(\zeta_l) \neq 0$. By irreducibility of the method we then have $\sigma_p(\zeta_l) \neq 0$. For ease of presentation, we assume[3] that also $\sigma_q(\zeta_l) \neq 0$. We then can apply the inverse of the operators $\sigma_p(\zeta_l e^{hD})$ and $\sigma_q(\zeta_l e^{hD})$ to both sides of (1.42) and thus obtain

$$\begin{aligned} \Big(\frac{\rho_p}{\sigma_p}\Big)(\zeta_l e^{hD})u_l &= h\sum_{m\geq 1}\frac{1}{m!}\sum_{\zeta_{l_1}\cdots\zeta_{l_m}=\zeta_l} f^{(m)}\big(y\big)\big(z_{l_1},\ldots,z_{l_m}\big) + \mathcal{O}(h^{N+1}), \\ \Big(\frac{\rho_q}{\sigma_q}\Big)(\zeta_l e^{hD})v_l &= h\sum_{m\geq 1}\frac{1}{m!}\sum_{\zeta_{l_1}\cdots\zeta_{l_m}=\zeta_l} g^{(m)}\big(y\big)\big(z_{l_1},\ldots,z_{l_m}\big) + \mathcal{O}(h^{N+1}). \end{aligned} \tag{1.58}$$

Expanding the left-hand side into powers of h leads to the consideration of the series

$$i\,\frac{\rho_p(\zeta_l\, e^{ix})}{\sigma_p(\zeta_l\, e^{ix})} = \lambda^p_{l0} + \lambda^p_{l1}x + \lambda^p_{l2}x^2 + \ldots, \quad i\,\frac{\rho_q(\zeta_l\, e^{ix})}{\sigma_q(\zeta_l\, e^{ix})} = \lambda^q_{l0} + \lambda^q_{l1}x + \lambda^q_{l2}x^2 + \ldots$$

(note that $\lambda^p_{l0} = 0$ if $\rho_p(\zeta_l) = 0$). The symmetry of the methods implies that the coefficients λ^p_{lj} and λ^q_{lj} are real. For the conjugate root $\zeta_{-l} = \overline{\zeta_l}$ we have

$$\lambda^p_{-l,j} = (-1)^{j+1}\lambda^p_{l,j}, \qquad \lambda^q_{-l,j} = (-1)^{j+1}\lambda^q_{l,j}. \tag{1.59}$$

Removing in (1.58) the terms with $m \geq 2$, we thus obtain

$$\begin{aligned} \ldots + \lambda^p_{l2}\,(-ih)^2\,\ddot{u}_l + \lambda^p_{l1}\,(-ih)\,\dot{u}_l &= ih\,\big(f_p(p,q)\,u_l + f_q(p,q)\,v_l\big) \\ \ldots + \lambda^q_{l2}\,(-ih)^2\,\ddot{v}_l + \lambda^q_{l1}\,(-ih)\,\dot{v}_l + \lambda^q_{l0}\,v_l &= ih\,\big(g_p(p,q)\,u_l + g_q(p,q)\,v_l\big) \end{aligned} \tag{1.60}$$

and the same relations for l replaced by $-l$. An important ingredient for a further study is the fact that

$$\begin{aligned} \mathrm{Re}\,\Big(\overline{z}^{\mathsf{T}} z^{(2m+1)}\Big) &= \frac{1}{2}\frac{d}{dt}\Big(\sum_{j=0}^{2m}(-1)^j(\overline{z}^{(j)})^{\mathsf{T}} z^{(2m-j)}\Big) \\ \mathrm{Im}\,\Big(\overline{z}^{\mathsf{T}} z^{(2m)}\Big) &= \frac{1}{2i}\frac{d}{dt}\Big(\sum_{j=0}^{2m-1}(-1)^j(\overline{z}^{(j)})^{\mathsf{T}} z^{(2m-j-1)}\Big) \end{aligned} \tag{1.61}$$

are total differentials. We first put the main result of [HL04] on the long-time behavior of parasitic solution components into the context of the present investigation.

[3] The case $\sigma_q(\zeta_l) = 0$ needs special attention, see the end of Section 1.3.5 or [HL04] for the special case of second order differential equations.

Second order Hamiltonian systems. We consider partitioned systems

$$\dot p = -\nabla U(q), \qquad \dot q = p,$$

which are equivalent to second order differential equations $\ddot q = -\nabla U(q)$. In this case we have $g_q(p,q) = 0$ and $g_p(p,q) = I$, so that from the lower line of (1.60) the expression ihu_l is seen to be a linear combination of derivatives of v_l. Inserted into the upper relation of (1.60) this gives

$$\ldots - \lambda_{l3}\,(-ih)^2\, v_l^{(3)} - \lambda_{l2}\,(-ih)\,\ddot v_l - \lambda_{l1}\,\dot v_l = -ih\,\nabla^2 U(q)\,v_l, \tag{1.62}$$

where $\lambda_{l1} = \lambda^p_{l1}\lambda^q_{l0}$, $\lambda_{l2} = \lambda^p_{l2}\lambda^q_{l0} + \lambda^p_{l1}\lambda^q_{l1}$, etc. are real coefficients. It follows from the symmetry of the Hessian matrix $\nabla^2 U(q)$ that $\mathrm{Im}\,(\overline{v}_l^{\mathsf T}\nabla^2 U(q)\,v_l) = 0$. Taking the scalar product of (1.62) with $\overline{v}_l^{\mathsf T}$ and considering its real part, we thus obtain

$$\ldots + h^2\lambda_{l3}\,\mathrm{Re}\,(\overline{v}_l^{\mathsf T} v_l^{(3)}) - h\lambda_{l2}\,\mathrm{Im}\,(\overline{v}_l^{\mathsf T}\ddot v_l) - \lambda_{l1}\,\mathrm{Re}\,(\overline{v}_l^{\mathsf T}\dot v_l) = 0.$$

The magic formulas (1.61) show that the left-hand expression is a total differential. Its dominant term is the derivative of $-\lambda_{l1}\frac{1}{2}\|v_l\|^2$. The other terms are the derivative expressions containing higher derivatives of v_l. These can be eliminated with the help of the simplified modified differential equation. Because of $\lambda_{l1} \neq 0$, we thus get a formal invariant (a near invariant if the series is truncated) of the system (1.60), which is of the form

$$\ldots + h^2 I_{l2}(y,\mathbf{z}) + h\,I_{l1}(y,\mathbf{z}) + \|v_l\|^2 = I_l(y,\mathbf{z}).$$

Since all functions $I_{lj}(y,\mathbf{z})$ are bounded by a constant times $\|\mathbf{z}\|^2$ and since we obtain such a formal invariant for all components of $\mathbf{z}$, the stability assumption (SI) is proved with $C = 1 + \mathcal{O}(h)$ and for arbitrarily large M.

Remark 1.3.6. This derivation of a near invariant that is close to $\|v_l\|^2$ essentially relies on the fact that the polynomials $\rho_p(\zeta)$ and $\rho_q(\zeta)$ do not have common roots other than $\zeta = 1$. If, in addition to $\rho_p(\zeta_l) = 0$, also $\rho_q(\zeta_l) = 0$ would be satisfied, then the coefficient λ^q_{l0} would be zero. This would imply $\lambda_{l1} = 0$, so that the formal invariant does not contain the term $\|v_l\|^2$.

Separable Hamiltonian systems. We next consider a Hamiltonian system with

$$H(p,q) = T(p) + U(q).$$

We still consider partitioned linear multistep methods (1.2), where the ρ-polynomials do not have common zeros with the exeption of $\zeta = 1$. In the situation of (1.60) the vector v_l contains a factor h. Since $f_p(p,q) = 0$ for a separable Hamiltonian system, the differential equation for u_l contains an additional factor h. Consequently, the differential equation (1.54) for $\mathbf{z}$ is in fact of the form $\dot{\mathbf{z}} = h\,A^0_{h,N}(y)\mathbf{z}$. Therefore we have $\|\mathbf{z}(h)\| \le \|\mathbf{z}(0)\|(1+\mathcal{O}(h^2))$, so that the stability assumption (SI) is satisfied with $M = 1$.

Discussion of the examples of Section 1.1.3. In the numerical experiments of Section 1.1.3 we have seen situations, where the parasitic solution components remain bounded on intervals of length $\mathcal{O}(h^{-2})$. According to our Theorem 1.3.4 this requires the stability assumption to be satisfied for $M = 2$. The system (1.60) is of the form

$$\begin{aligned} \lambda^p_{l1}\dot u_l &= \nabla^2 U(q)v_l + \mathcal{O}(h^2\|\mathbf{z}\|) \\ \lambda^q_{l0} v_l &= ih\nabla^2 T(p)u_l + \mathcal{O}(h^2\|\mathbf{z}\|) \end{aligned} \tag{1.63}$$

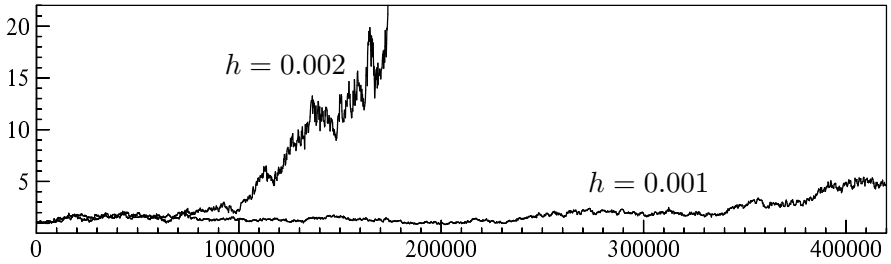


Figure 1.5 – Euclidean norm of the parasitic solution component v_1; data for the Hamiltonian system are as in Figure 1.1, problem (C); initial data for the parasitic component are normalized to $\|v_1(0)\| = 1$.

which yields the differential equation

$$\dot{u}_l = i\,h\,\lambda\,\nabla^2 U(q)\nabla^2 T(p)\,u_l + \mathcal{O}(h^2\|\mathbf{z}\|)$$

with $\lambda = \lambda^q_{l0}/\lambda^p_{l1}$. If the product of the two Hessian matrices is symmetric or, equivalently, if their commutator vanishes, i.e.,

$$\big[\nabla^2 U(q), \nabla^2 T(p)\big] = \nabla^2 U(q)\nabla^2 T(p) - \nabla^2 T(p)\nabla^2 U(q) = 0, \tag{1.64}$$

we can multiply the differential equation with $\overline{u}_l^{\mathsf{T}}$ and we obtain

$$\|u_l(h)\|^2 = \|u_l(0)\|^2 + \mathcal{O}(h^3\|\mathbf{z}(0)\|^2)$$

as a consequence of $\mathrm{Im}\,(\overline{u}_l^{\mathsf{T}}\nabla^2 U(q)\nabla^2 T(p)\,u_l) = 0$. This prove the validity of the stability assumption (SI) with $M = 2$. Unfortunately, the commutativity of the two Hessian matrices is a strong requirement and not often satisfied.

The examples (A) and (B) of Section 1.1.3 are separable Hamiltonian equations, which split into independent subsystems having one degree of freedom. The condition (1.64) is therefore trivially satisfied.

For the example (C) the condition (1.64) is not satisfied, so that we do not have better than $M = 1$ in the stability assumption (SI). Let us explain the behavior observed in Figure 1.2. The parasitic roots of method 'plmm2' are $\zeta_1 = i$, $\zeta_{-1} = -i$, and $\zeta_2 = -1$.

We have $\sigma_q(\zeta_2) = 0$, so that the division by $\sigma_q(\zeta_2 e^{hD})$ is not permitted in (1.58). We thus go back to formula (1.42), which shows that for $\rho_q(\zeta_l) \neq 0$ and $\sigma_q(\zeta_l) = 0$ the vector v_l is an expression multiplied by h^2. Inserted into the first equation of (1.63) we see that the right-hand side of the differential equation for u_2 contains the factor h^2, so that $\|u_2(h)\|^2 = \|u_2(0)\|^2 + \mathcal{O}(h^3\|\mathbf{z}(0)\|^2)$.

For the root $\zeta_1 = i$ we study numerically the dominant term of the parasitic solution component. We have $\lambda^p_{l0} = -1$ and $\lambda^q_{l1} = 2$ for the method 'plmm2', so that the differential equation for v_l becomes

$$\dot{v}_1 = -\frac{i\,h}{2}\,\nabla^2 T(p)\nabla^2 U(q)\,v_1 + \mathcal{O}(h^2\|\mathbf{z}\|).$$

We neglect the $\mathcal{O}(h^2\|\mathbf{z}\|)$ term and solve the linear differential equation for v_1 numerically with the code DOPRI5 of [HNW93]. Since the problem is chaotic, care has to be taken about the credibility of the results. We therefore solve the problem with a high accuracy requirement of $tol = 10^{-12}$ and with many different initial values of norm $\|v_1(0)\| = 1$. The result is qualitatively the same for all runs, and we plot in Figure 1.5 one such parasitic solution.

If the starting approximations for the partitioned multistep method are computed with high accuracy (what is the case for all our numerical experiments), the initial values of the parasitic solution components are of size $\mathcal{O}(h^{r+1})$ (where r denotes the order of the method). Consequently, the functions shown in Figure 1.5 have to be scaled with a factor $\mathcal{O}(h^{r+1})$. A comparison with Figure 1.2 shows that this solution, where we have removed quadratic and higher order terms in $\mathbf{z}$ as well a linear terms in $\mathbf{z}$ with a a factor of at least h^2, cannot be the reason of the exponential divergence in Figure 1.2. It must be a consequence of the next term having a factor h^2. This nicely explains why the parasitic solution components remain small and bounded on intervals of length $\mathcal{O}(h^{-2})$.

Conclusion

We have studied the long-time behavior of partitioned linear multistep methods applied to Hamiltonian systems. These are methods, where the momenta p and the positions q of the system are treated by two different multistep formula. It turns out that the following two properties are essential for a qualitative correct simulation over long times:

- both multistep schemes have to be symmetric;
- the generating polynomials $\rho_p(\zeta)$ and $\rho_q(\zeta)$ of the two methods are not allowed to have common zeros with the exception of $\zeta = 1$.

The study is motivated by the analysis of [HL04] for special multistep methods and Hamiltonian systems of the form $\ddot{q} = -\nabla U(q)$. We have extended the techniques of proof to a more general situation.

The positive insight of our investigation is that for problems having symmetries and a regular solution behavior, the numerical results concerning long-time preservation of energy and quadratic first integrals are excellent. This is remarkable, because the considered methods are explicit, of arbitrarily high order, and can be implemented very efficiently. We expect that this excellent long-time behavior is typical for all nearly integrable systems. A more thorough investigation of this question is outside the scope of the present work.

For separable Hamiltonian systems with chaotic solution, we observed that the 'smooth' numerical solution behaves exactly like a symmetric (non-symplectic) one-step method. The parasitic solution components are typically bounded on a time interval of length $\mathcal{O}(h^{-2})$, but usually not on longer time intervals. This observation is independent of the order of the method.

Recently we have extended our numerical experiments and also the theoretical investigations to constrained Hamiltonian systems, which are differential-algebraic equations of index 3. Preliminary results are very encouraging and we expect to obtain a new efficient class of methods for such problems.

Chapter 2

Complements on symmetric partitioned LMM for Hamiltonian systems

2.1 Introduction

In this Chapter we describe some complements to the theoretical analysis of linear partitioned multistep methods applied to Hamiltonian systems of Chapter 1.

The Chapter is divided in two parts.

In the first part we recall some known techniques for the construction of multistep methods of arbitrarily high order, and then we construct and optimize the stability partitioned multistep methods of order 4 and 6 that satisfy the additional order conditions described in Section 1.2.3. This analysis is supplemented with numerical methods to compare the performance of different methods of the same class.

The second part is dedicated to linear partitioned multistep methods applied to non-separable Hamiltonians. We test different linear partitioned multistep methods (including the methods discussed in the first part) on a few non-separable Hamiltonians, showing the error on the energy and the numerical solution of the leading term of the parasitic equations.

2.2 Construction of the method and stability optimization

As explained in Section 1.2.3, it is possible to construct explicit partitioned multistep methods $\rho_p(\zeta)$, $\sigma_p(\zeta)$, $\rho_q(\zeta)$, $\sigma_q(\zeta)$of order r, such that the $\rho(\zeta)$ polynomials have no roots in common but $\zeta = 1$, and the expansions

$$\frac{\rho(e^x)}{x\sigma(e^x)} = 1 + \lambda_r x^r + \lambda_{r+2} x^{r+2} + \dots \tag{2.1}$$

satisfy

$$\lambda_i^p = \lambda_i^q \text{ for some } i \geq r. \tag{2.2}$$

In this Section, we construct classes of methods of order up to 6 that will depend on some parameters: we show then that these parameters can be chosen such that (2.2) is satisfied. Furthermore we discuss how to optimize the stability of these methods by maximizing the distances of the roots of the $\rho(\zeta)$ polynomials.

This analysis is supplemented with some numerical experiments to compare the performance of the optimized method with that of a non-optimized method of the same class.

2.2.1 Construction of multistep methods

We recall the definition of order of consistency: a linear multistep method has order r if the following condition is satisfied:

$$\frac{\rho(\zeta)}{\log\zeta} - \sigma(\zeta) = \mathcal{O}\big((\zeta-1)^r\big) \quad \text{for} \quad \zeta \to 1. \tag{2.3}$$

We can use this condition to derive a high order method starting from a polynomial $\rho(\zeta)$ of degree $k+1$ with k even: we consider $\rho(\zeta)$ of the form

$$\rho(\zeta) = (\zeta-1)\prod_{j=1}^{k/2}(\zeta^2 + 2a_j\zeta + 1) \quad \text{with } |\alpha_j| < 1$$

and distinct a_j.

We choose a polynomial of this form so that all the zeros are simple and on the unit circle; as remarked in Section 1.1, for a fixed $\rho(\zeta)$ there exist a unique $\sigma(\zeta)$ which yields an explicit method of order at least k.

We observe that polynomial of degree k obtained by expanding $\rho(\zeta)/\log\zeta$ around $\zeta = 1$ is not symmetric. To obtain a symmetric $\sigma(\zeta)$ of degree k, first of all we have to compute the symmetric polynomial $\bar\sigma(\zeta)$ of degree $k+1$ obtained by truncating the expansion of $\rho(\zeta)/\log\zeta$. We consider then the expression $\bar\sigma(\zeta) + c(\zeta-1)^k(\zeta+1)$, which is a symmetric polynomial of degree $k+1$: to make it of degree k, we compute c in order to have the coefficient of the term of degree $k+1$ equal to zero.

In this way, we obtain classes of symmetric multistep methods for first order equations of an arbitrarily high order: in Appendix B we show the Maple codes for computing of the classes of methods of order 4 and 6.

2.2.2 Stability optimization: partitioned LMM with 5 steps and order 4

As showed in Section 1.2.3, the class of 5-step methods with order 4 can be described by the polynomials

$$\rho(\zeta) = (\zeta-1)\left(\zeta^2 + 2a_1\zeta + 1\right)\left(\zeta^2 + 2a_2\zeta + 1\right)$$

and

$$\sigma(\zeta) = (-s_2 + 5s_1 + 11)\left(\zeta^4 + \zeta\right)/6 + (13s_2 + 7s_1 + 1)\left(\zeta^3 + \zeta^2\right)/6,$$

where $|a_1|, |a_2| < 1$, $s_1 = a_1 + a_2$ and $s_2 = a_1a_2$. The corresponding λ_4 is given by

$$\lambda_4 = \frac{131 - 19s_1 + 11s_2}{720(1 + s_1 + s_2)} :$$

this parameter depends on two different coefficients, so it is possible to construct an order 4 partitioned multistep method $\rho_p(\zeta)$, $\sigma_p(\zeta)$, $\rho_q(\zeta)$, $\sigma_q(\zeta)$ that satisfies

$$\lambda_4^p = \lambda_4^q. \tag{2.4}$$

In Section 1.2.3, we show a choice of values a_1^p, a_2^p and a_1^q, a_2^q such that (2.4) is satisfied (plmm4c); here we are interested in optimizing the stability of this method, i.e. choosing a_1^p, a_2^p and a_1^q, a_2^q such that (2.4) is satisfied, and such that the distances between the roots

of $\rho_p(\zeta)$ and $\rho_q(\zeta)$ on the unit circle are maximized.
The problem has been solved numerically, with a code that computes the distances of the roots of the polynomials $\rho_p(\zeta)$ and $\rho_q(\zeta)$: the parameters a_1^p, a_2^p and a_1^q vary on grids of equidistant points between -1 and 1, and the fourth parameter a_2^q is thus computed using (2.4).

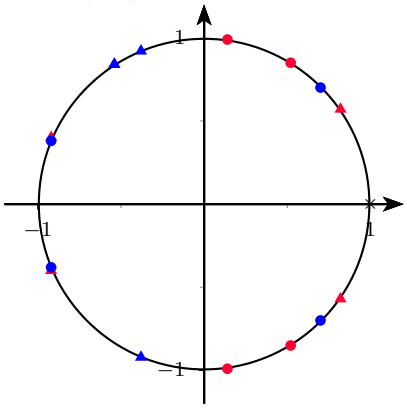


Figure 2.1 – Zeros of the optimized order 4 method (red) and of the non optimized order 4 method (blue).

In this way we found that the optimized configuration is given by the following values of the angles (in radians)

$$\theta_1^p \approx 0.4084, \quad \theta_2^p \approx 2.5133 \qquad (2.5)$$
$$\theta_1^q \approx 1.6860, \quad \theta_2^q \approx 2.1040.$$

This configuration is represented in red in Figure 2.1: the roots of $\rho_p(\zeta)$ and $\rho_q(\zeta)$ are respectively represented by dots and triangles; the same is done in blue for plmm4c described in 1.2.3.

We now want to compare the method corresponding to the optimized configuration (2.5) with plmm4c: we use the following separable Hamiltonian $H(p,q) = T(p) + U(q)$ with

$$T(p) = p_1^4 + p_2^4 + p_1^2 + p_2^2 \quad \text{and} \quad U(q) = q_1^4 + q_2^4 + q_1^3 + q_1^2 + 2q_2^2. \qquad (2.6)$$

This Hamiltonian has already been used as a test in Section 1.1.3; as already mentioned, it is reversible with respect to the transformation $p \leftrightarrow -p$ and it is separated into two systems with one degree of freedom each.

We observe in Figure 2.2 that the behaviour of the two methods over a long interval is the same. We were surprised that the error of the non-optimized method is slightly smaller than the one obtained with the optimized method: this is explained by looking at the error constants of the methods.

We recall that the error constant of a method of order r is given by

$$C = \frac{C_{r+1}}{\sigma(1)}$$

where C_{r+1} is defined by

$$\rho(e^h) - h\sigma(e^h) = C_{r+1}h^{r+1} + \mathcal{O}(h^{r+2}).$$

For the class of methods considered, the error constant is equal to

$$C = \lambda_4 = \frac{131 - 19s_1 + 11s_2}{720(1 + s_1 + s_2)}.$$

If we denote by C_p and C_q respectively the error constants for the method used for the integration of the momenta and the positions, then for plmm4c we have $C_p = C_q \approx 0.4858$, and for the non-optimized method we have $C_p = C_q \approx 0.2950$, which are smaller.
This explains why, even though the optimized method is more stable, it leads to a slightly larger error.

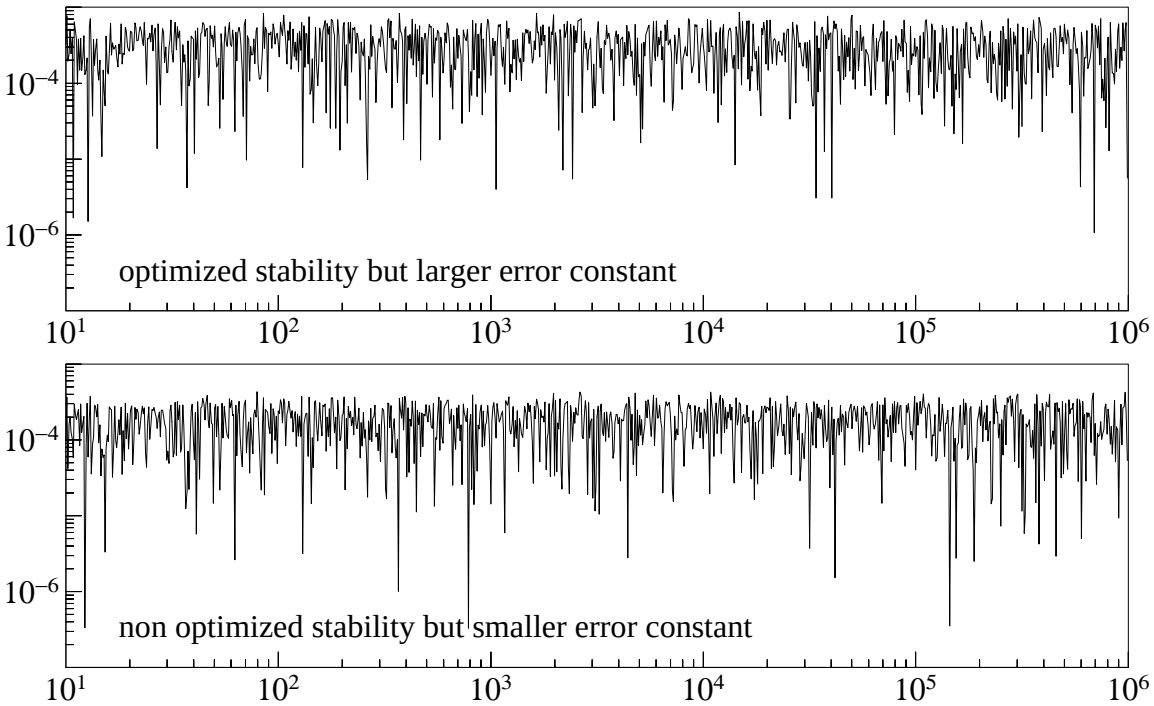


Figure 2.2 – Comparison of optimized and non-optimized formulations of the improved method of order 4: it is reported the error in the Hamiltonian as a function of time, $h = 0.005$. Initial approximations are computed using an implicit Runge Kutta method of order 8.

2.2.3 Stability optimization: partitioned LMM with 7 steps and order 6

The class of multistep methods of 7 steps and order 6 is given by the polynomials

$$\rho(\zeta) = (\zeta - 1)\left(\zeta^2 + 2a_1\zeta + 1\right)\left(\zeta^2 + 2a_2\zeta + 1\right)\left(\zeta^2 + 2a_3\zeta + 1\right)$$

and

$$\sigma(\zeta) = a\left(\zeta^6 + \zeta\right) + b\left(\zeta^5 + \zeta^2\right) + c\left(\zeta^4 + \zeta^3\right)$$

with

$$\begin{aligned} a &= \left(131s_1 - 19s_2 + 11s_3 + 461\right)/180 \\ b &= \left(89s_1 + 119s_2 - 31s_3 - 121\right)/60 \\ c &= \left(161s_1 + 191s_2 + 401s_3 + 311\right)/90 \end{aligned}$$

where $s_1 = a_1 + a_2 + a_3$, $s_2 = a_1a_2 + a_2a_3 + a_1a_3$, $s_3 = a_1a_2a_3$ and $|a_1|$, $|a_2|$, $|a_3| < 1$. The associated coefficients λ_6 and λ_8 of the expansion (2.1) are given by

$$\lambda_6 = -\frac{527s_1 - 271s_2 + 191s_3 - 4975}{60480(1 + s_1 + s_2 + s_3)}$$

and

$$\lambda_8 = \frac{17103 + 8006s1 - 314s_7 + 283s6 + 186s_4 + 1472s2 - 4416s3 - 1237s_5 + 352s_8 + 63s_9}{(1 + a_1)^2(1 + a_2)^2(1 + a_3)^2}$$

where $s_4 = a_1^2a_2 + a_1a_2^2 + a_1a_3^2 + a_1^2a_3 + a_2a_3^2 + a_2^2a_3$, $s_5 = a_1^2 + a_2^2 + a_3^2$, $s_6 = a_1^2a_2^2 + a_1^2a_3^2 + a_2^2a_3^2$, $s_7 = a_1^2a_2^2a_3 + a_1^2a_2a_3^2 + a_1a_2^2a_3^2$, $s_8 = a_1^2a_2a_3 + a_1a_2^2a_3 + a_1a_2a_3^2$, $s_9 = a_1^2a_2^2a_3^2$.

In this case we want to find a partitioned multistep method such that $\rho_p(\zeta)$ and $\rho_q(\zeta)$ have no roots in common except $\zeta = 1$, and

$$\lambda_6^p = \lambda_6^q \quad \text{and} \quad \lambda_8^p = \lambda_8^q. \tag{2.7}$$

As before, we solve the problem numerically: we fix four out of the six parameters and we find the remaining two by (2.7); all these parameters are used to find the configuration that maximizes the minimum of the distances between the roots of the polynomials $\rho_p(\zeta)$ and $\rho_q(\zeta)$.
The values for the angles (in radians) that we find in this way are

$$\begin{aligned} &\theta_1^p \approx 1.9871, \quad \theta_2^p \approx 0.1491, \quad \theta_3^p \approx 1.8546 \\ &\theta_1^q \approx 2.1176, \quad \theta_2^q \approx 1.2870, \quad \theta_3^q \approx 1.1152, \end{aligned} \tag{2.8}$$

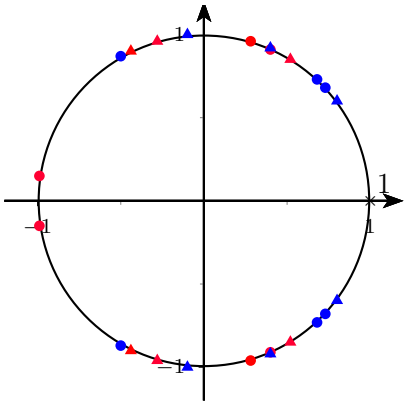


Figure 2.3 – Zeros of of the optimized order 6 method (red) and of the non optimized order 6 method (blue).

We want to compare the performance of the optimized method with another non-optimized method of the same class: we choose for example the method given by the angles

$$\begin{aligned} &\bar\theta_1^p \approx 2.3242, \quad \bar\theta_2^p \approx 2.3945, \quad \bar\theta_3^p \approx 1.0472 \\ &\bar\theta_1^q \approx 1.4706, \quad \bar\theta_2^q \approx 1.9823, \quad \bar\theta_3^q \approx 2.4981. \end{aligned} \tag{2.9}$$

Both the optimized and non-optimized configurations (2.8) and (2.9) are represented in red and in blue in Figure 2.3 respectively: the roots of $\rho_p(\zeta)$ and $\rho_q(\zeta)$ are represented by dots and triangles respectively.

Figure 2.4 shows the comparison of the errors in the Hamiltonian obtained with the methods defined by the angles (2.9) and (2.8), and we observe that with the optimized method (2.8) we obtain an error about 10 times smaller than the error obtained with (2.9).

Again, this can be explained with the error constant, that is

$$C = \lambda_6 = -\frac{527s_1 - 271s_2 + 191s_3 - 4975}{60480(1 + s_1 + s_2 + s_3)}.$$

In fact, for the optimized method (2.8) we have $C_p = C_q \approx 0.09109$, whereas for the non optimized method (2.9) we have $C_p = C_q \approx 0.7016$. This explains the results of the numerical experiments.

2.3 Some numerical examples of non-separable systems

In Chapter 1, we concentrated on the properties of near-preservation of energy and momenta of explicit partitioned multistep methods applied to separable Hamiltonian systems. In this section, we present several numerical experiments performed on different non-separable Hamiltonians.

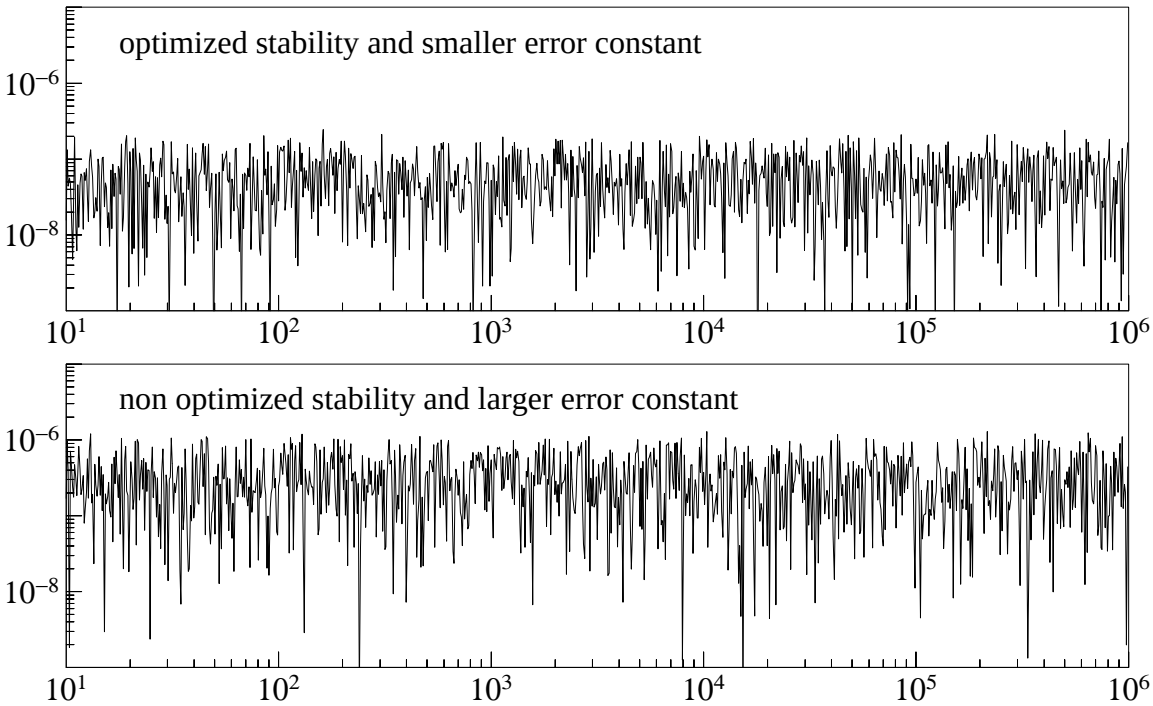


Figure 2.4 – Comparison of optimized and non-optimized formulations of the improved method of order 6: it is reported the error in the Hamiltonian as a function of time, $h = 0.005$. Initial approximations are computed using an implicit Runge Kutta method of order 8.

2.3.1 Numerical Examples: a polynomial non-separable Hamiltonian

The first Hamiltonian we want to investigate is given by

$$H(p,q) = \left(1 + \frac{p^2}{2}\right)^2 \left(1 + q^2\right) \tag{2.10}$$

whose associated Hamilton equations are

$$\begin{cases} \dot{p} &= -2q\left(1 + \frac{p^2}{2}\right)^2 \\ \dot{q} &= 2p(1+q^2)\left(1 + \frac{p^2}{2}\right) \end{cases} ; \tag{2.11}$$

it is described in [Chi09]. We consider as initial data the values $p_0 = 0.5$ and $q_0 = 0$.

We test on this system five different algorithms: plmm2 (described in Section 1.1.3), plmm4c (described in 1.2.3), the optimized order 4 method (2.5), and the non-optimized and the optimized order 6 methods (2.9) and (2.8); the initial values for all these methods are computed with an implicit Runge-Kutta method of order 8. We observe that the energy is preserved for very long time integration with all the proposed algorithms.

We compute as well the parasitic components: we recall that they are solutions of equations of the form

$$\begin{cases} \dot{u}_l = -\mu_l H_{pq}(p,q) u_l + \mathcal{O}(h) \\ v_l = \mathcal{O}(h) \end{cases} \tag{2.12}$$

and

$$\begin{cases} u_m = \mathcal{O}(h) \\ \dot{v}_m = \mu_m H_{qp}(p,q) v_m + \mathcal{O}(h) \end{cases} \tag{2.13}$$

respectively if they are associated to a root of $\rho_p(\zeta)$ or to a root of $\rho_q(\zeta)$. We solve (2.12) and (2.13) with $h = 0$ using an explicit Runge-Kutta method of order 4 and different initial data (depending on the size of the energy).

We observe that also the norm of the parasitic components stays small and bounded for a long time integration: its boundedness gives an explanation to the good behavior of the error on the energy on long time integrations.

2.3.2 Numerical examples: the spherical pendulum

A spherical pendulum is a mathematical pendulum in a tridimensional space: both the mass and the length of the pendulum are taken to be equal to one.
We show here some numerical experiments made on the Hamiltonian of this system, which in this case is described in spherical coordinates as in Figure 2.6. The Hamiltonian is

$$H(p,q) = \left(p_\theta^2 + \frac{p_\phi^2}{\sin^2\theta} \right) - \cos\theta \tag{2.14}$$

and the associated canonical equations are

$$\begin{cases} \dot{p}_\theta &= p_\phi^2 \frac{\cos\theta}{\sin^3\theta} - \sin\theta \\ \dot{p}_\phi &= 0 \\ \dot{\theta} &= p_\theta \\ \dot{\phi} &= \frac{p_\phi}{\sin^2\theta} \end{cases} . \tag{2.15}$$

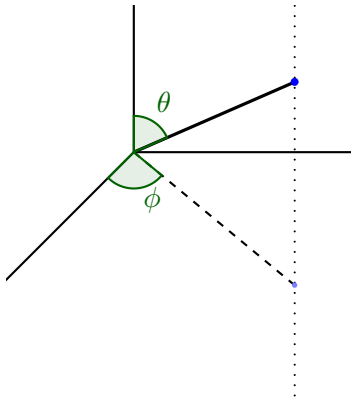


Figure 2.6 – Spherical Pendulum, notation in spherical coordinates.

As for the polynomial Hamiltonian, we use these equations to test the methods plmm2, plmm4c, the optimized order 4 method, as well as the non-optimized and optimized order 6 methods. We consider the initial data $p_0 = (-1.2, 1)$, $q_0 = (\pi/2, \pi/4)$. The error on the energy is shown in black in Figure 2.7 for $h = 0.005$, and we observe that we obtain near-preservation of the energy for long time.

As before, we solve the parasitic equations (2.12) and (2.13) with $h = 0$ with an explicit Runge-Kutta method of order 4; we remark that the parasitic component for the variable p_ϕ does not have to be computed, since that variable is integrated exactly.

In Figure 2.7 we show in red the norm of the remaining parasitic components, and we observe linear growth that is apparently in conflict with the good behaviour of the error in the energy. The explanation is that, since the variable ϕ does not appear in the Hamiltonian, the associated parasitic component does not influence the error in the energy: in fact, plotting the norm of all the parasitic components except for the parasitic component associated to ϕ, we observe that this norm stays small and bounded for long times, that explains the excellent behavior of the error in the Hamiltonian.

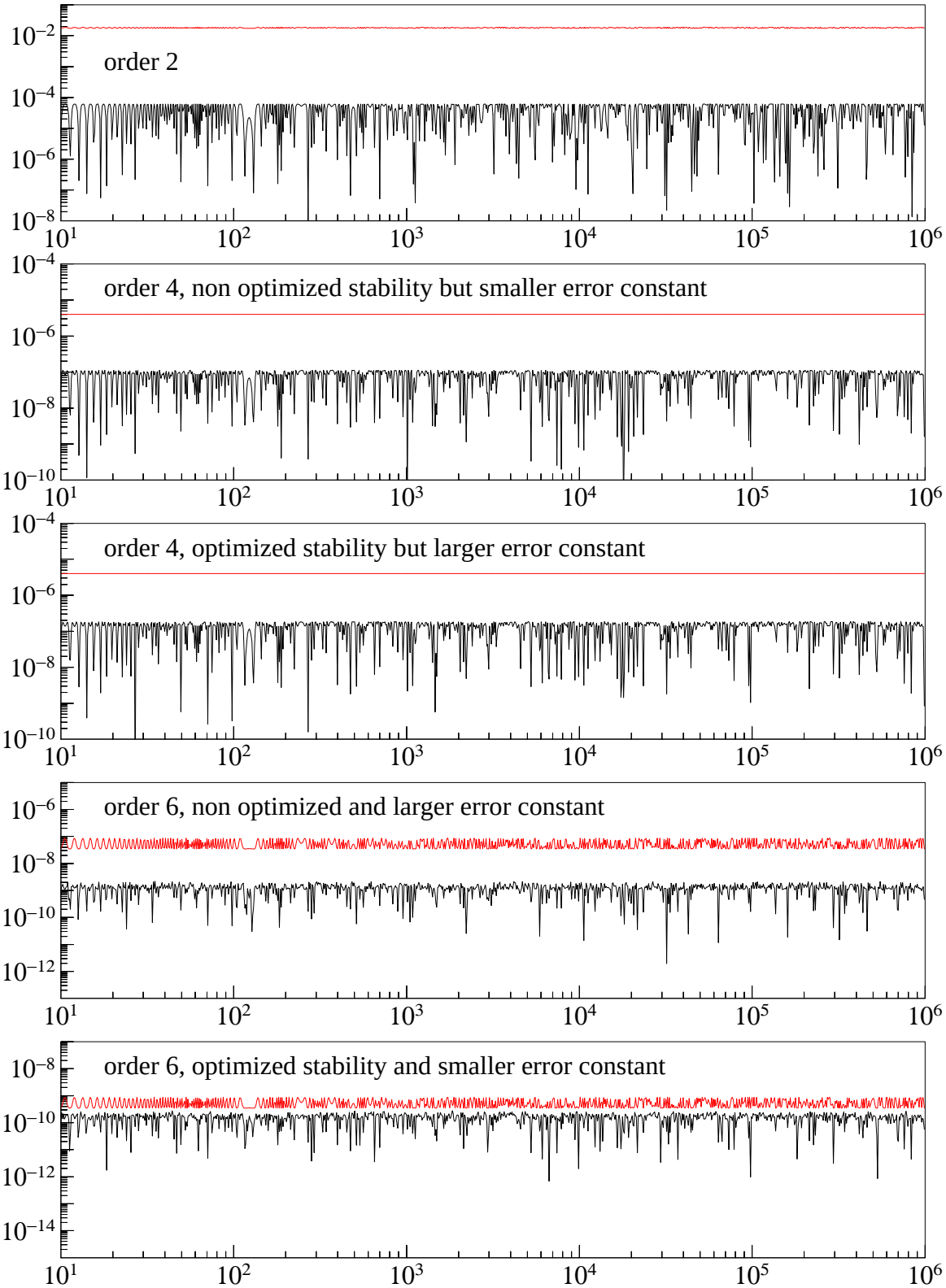


Figure 2.5 – (Polynomial non-separable Hamiltonian) Error in the Hamiltonian (black) and norm of the parasitic components (red) as functions of time, obtained using different methods, $h = 0.01$. Initial approximations are computed using an implicit Runge Kutta method of order 8.

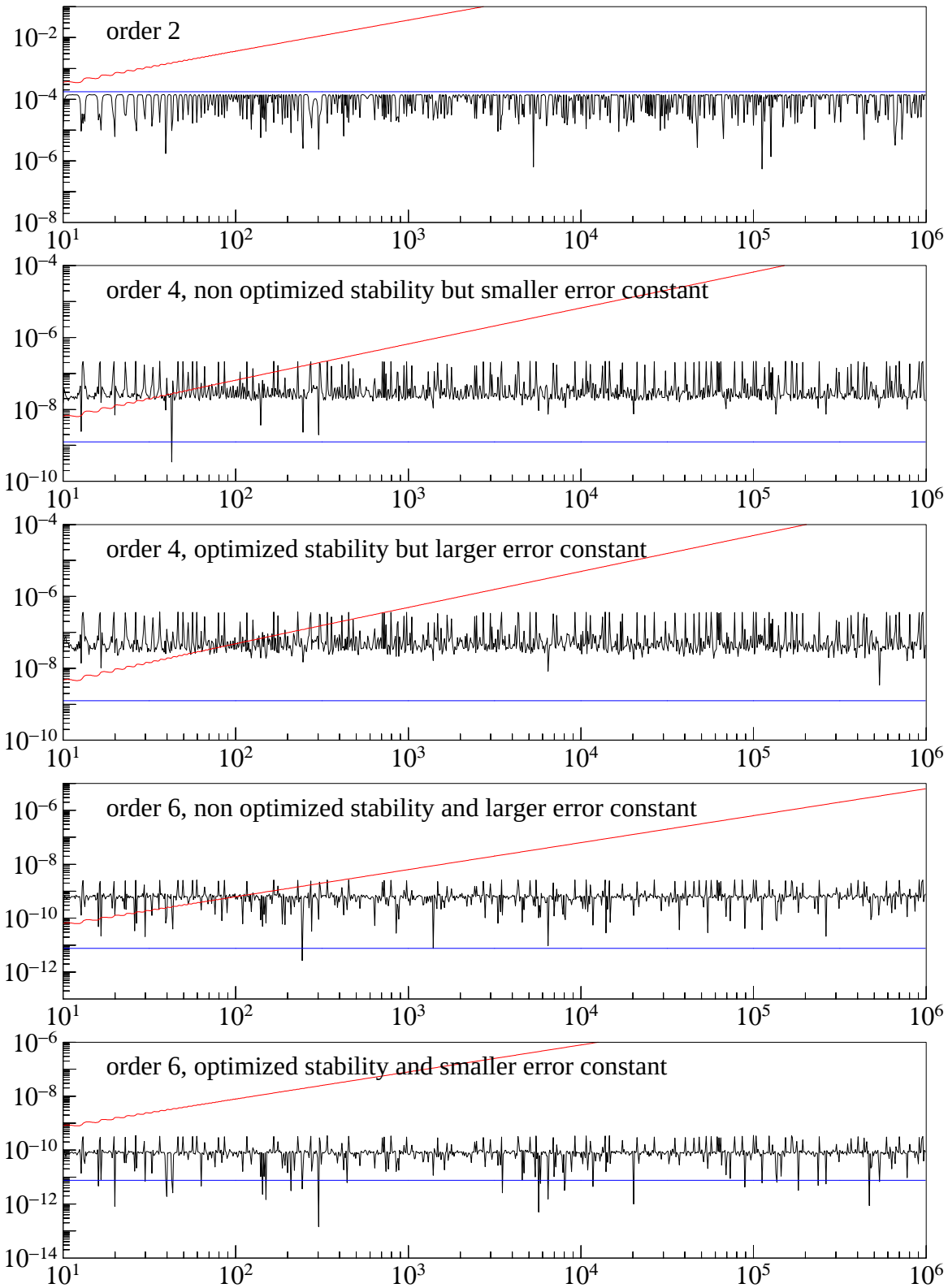


Figure 2.7 – (Spherical Pendulum) Error in the Hamiltonian (black), norm of the parasitic components (red) and norm of the parasitic components without the ϕ component as functions of time, obtained using different methods, $h = 0.005$. Initial approximations are computed using an implicit Runge Kutta method of order 8.

2.3.3 Numerical examples: the double pendulum

A double pendulum consists of a mathematical pendulum with another mathematical pendulum attached to its end. In this section we will consider both the pendula with unitary mass and length. We represent the double pendulum in polar coordinates, with the notation shown in Figure 2.8. With this notation, its Hamiltonian is

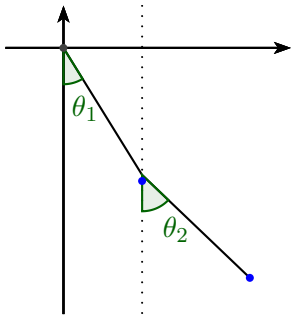


Figure 2.8 – Double pendulum, notation in polar coordinates.

$$H(p,q) = \frac{p_1^2 + 2p_2^2 - 2p_1 p_2 \cos(\theta_1 - \theta_2)}{2\left(1 + \sin^2(\theta_1 - \theta_2)\right)} - 2\cos(\theta_1) - \cos(\theta_2)\,. \tag{2.16}$$

As for the other Hamiltonians, we used the Hamiltonian (2.16) to test the methods plmm2, plmm4c, the optimized order 4 method and the non-optimized and optimized order 6 methods. We considered the initial data $p_0 = (-0.2, 0.3)$, $q_0 = (\pi/8, \pi/12)$ to obtain a non-chaotic motion, and we used $h = 0.005$.

Figure 2.9 shows in black the error in the Hamiltonian as a function of time, and in red the norm of the parastic components. As in the previous cases, we observe that they both stay small and bounded for long time integration.

We want now to show the performance of these methods when they are applied to a chaotic system: we choose as initial data $p_0 = (-0.2, 0.3)$, $q_0 = (\pi/2, \pi/2)$. Figure 2.10 shows that for this data the error in the Hamiltonian does not show the nice behavior observed for a regular system, but the error grows exponentially with time after relatively short time.

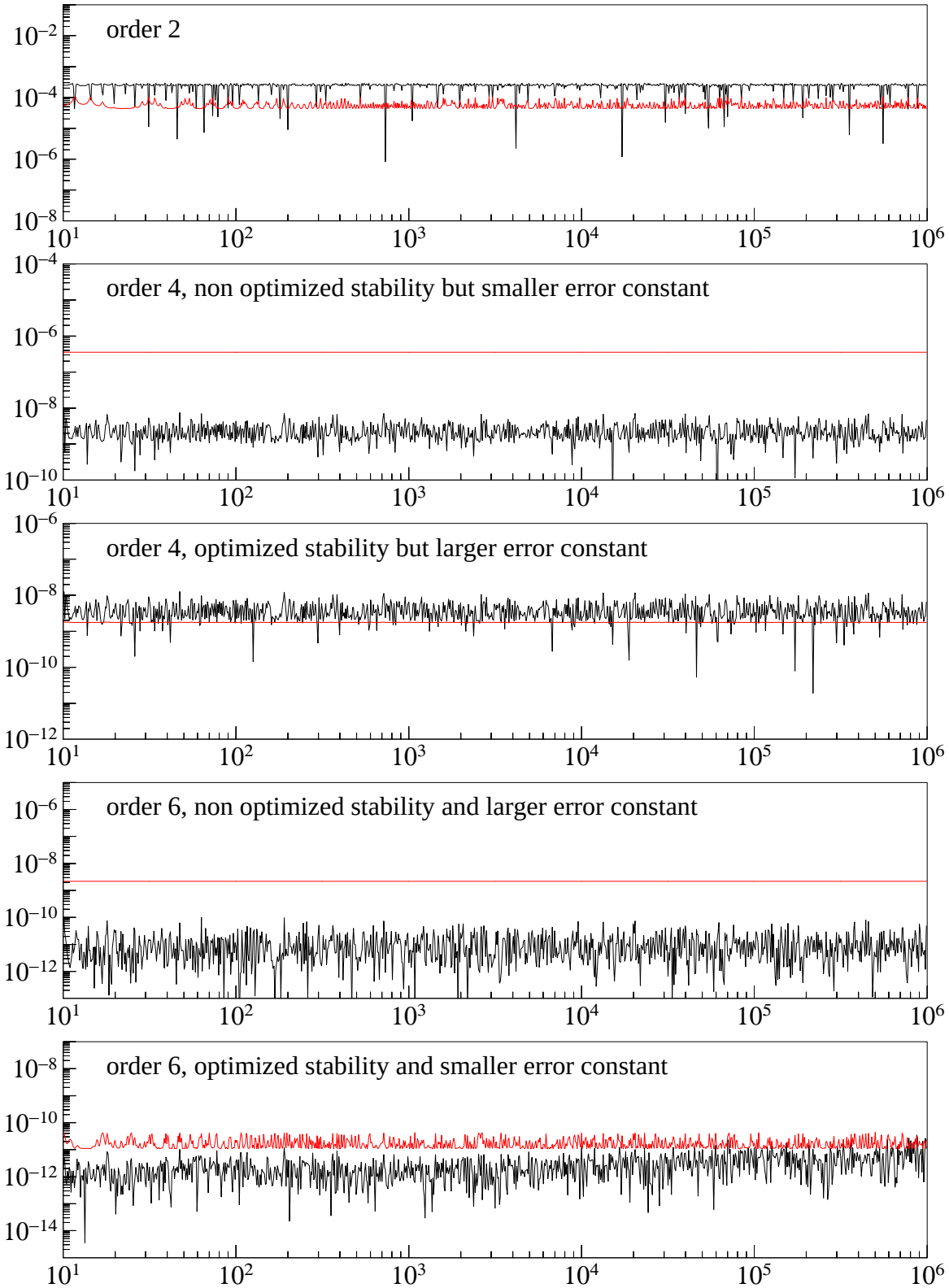


Figure 2.9 – (Double Pendulum, regular data) Error in the Hamiltonian (black) and norm of the parasitic components (red) as functions of time, obtained using different methods, $h = 0.005$. Initial approximations are computed using an implicit Runge Kutta method of order 8.

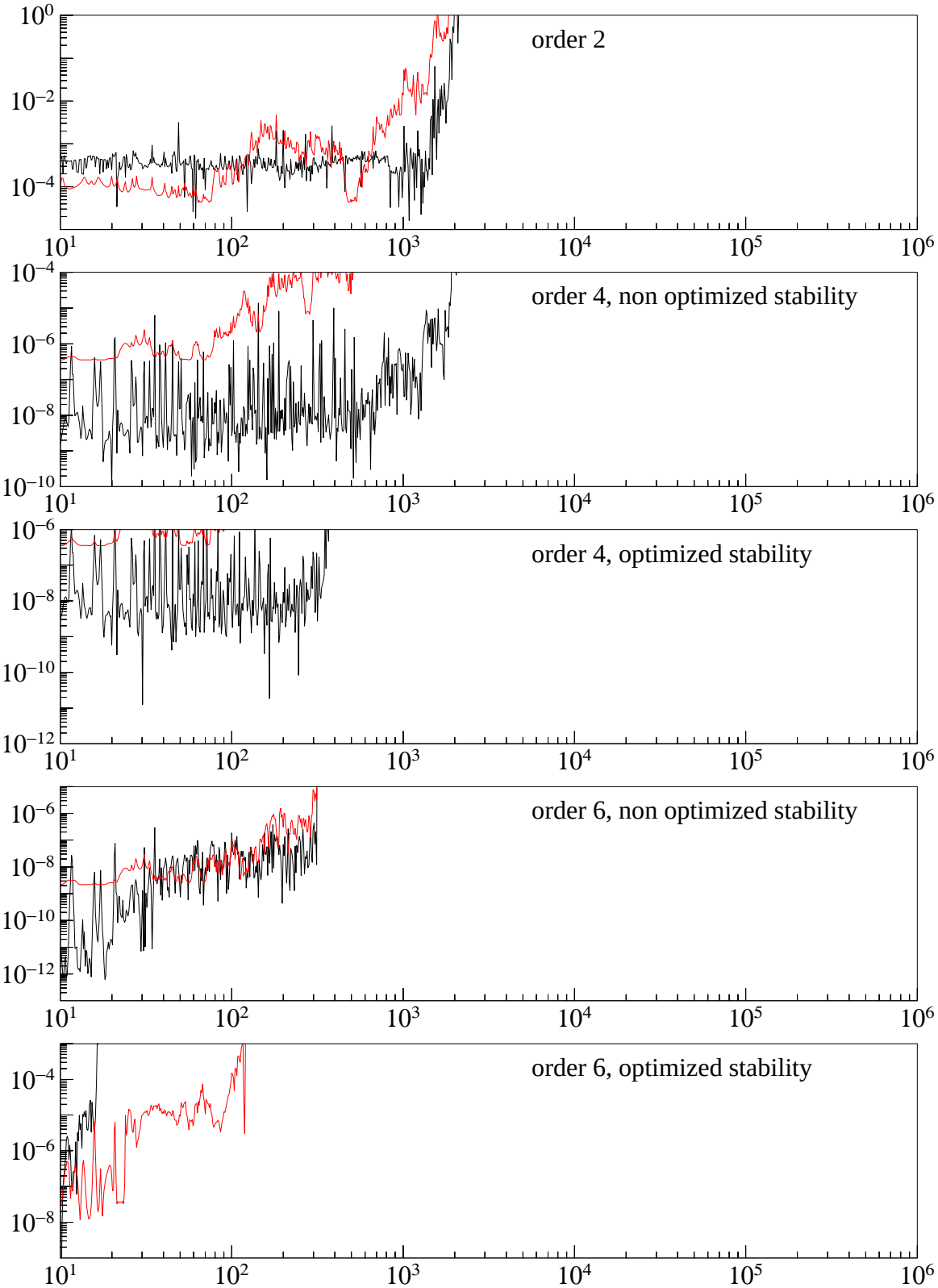


Figure 2.10 – (Double Pendulum, chaotic data) Error in the Hamiltonian (black) and norm of the parasitic components (red) as functions of time, obtained using different methods, $h = 0.005$. Initial approximations are computed using an implicit Runge Kutta method of order 8.

Chapter 3

Symmetric multistep methods for constrained Hamiltonian systems

Note: This chapter is identical to the paper [CHL13] in collaboration with E. Hairer and C. Lubich.

3.1 Introduction

The motion of mechanical systems is often constrained in the position coordinates (e.g., rigid body motion, frozen bonds in molecular dynamics). This typically leads to differential-algebraic equations of the form

$$\begin{aligned} M\ddot{q} &= -\nabla U(q) - G(q)^{\mathsf{T}}\lambda \\ 0 &= g(q), \end{aligned} \tag{3.1}$$

where $q \in \mathbb{R}^d$ is the vector of position coordinates, M is a positive definite mass matrix, $U(q)$ is a smooth real potential, $g(q) \in \mathbb{R}^m$ (with $m < d$) collects the constraints, and $G(q) = g'(q)$ is the matrix of partial derivatives. The term containing the Lagrange multiplier $\lambda \in \mathbb{R}^m$ forces the solution to satisfy the algebraic constraint. In addition to $g(q) = 0$ every solution of (3.1) also satisfies the differentiated relation $\frac{d}{dt}g(q) = G(q)\dot{q} = 0$. Initial values $q(0) = q_0$, $\dot{q}(0) = \dot{q}_0$ are said to be consistent if they satisfy both relations $g(q_0) = 0$ and $G(q_0)\dot{q}_0 = 0$. A second differentiation of the constraint leads to $\frac{\partial^2}{\partial q^2}g(q)(\dot{q},\dot{q}) + G(q)\ddot{q} = 0$ which, after insertion of (3.1), permits to express the Lagrange multiplier λ in terms of q and $\dot{q}$, provided that the matrix

$$G(q)M^{-1}G(q)^{\mathsf{T}} \qquad \text{is invertible} \tag{3.2}$$

along the solution. This will be assumed throughout this article. It implies that the differential-algebraic equation is of index 3.

Introducing the momentum $p = M\dot{q}$, the problem is seen to be Hamiltonian with total energy

$$H(q,p) = \tfrac{1}{2}\, p^{\mathsf{T}}M^{-1}p + U(q). \tag{3.3}$$

Elimination of the Lagrange multiplier λ from the system yields a differential equation on the manifold

$$\mathcal{M} = \{(q,p)\,;\, g(q) = 0,\, G(q)M^{-1}p = 0\}. \tag{3.4}$$

The flow is symplectic on $\mathcal{M}$, and the energy $H(q,p)$ is preserved along solutions of the system. In the spirit of geometric numerical integration one is interested in numerical simulations that share these properties as far as possible.

The most natural discretization of (3.1) is obtained when the second derivative is replaced by a central difference. This leads to the so-called SHAKE algorithm [RCB77]

$$\begin{aligned} q_{n+1} - 2q_n + q_{n-1} &= -h^2 M^{-1}\big(\nabla U(q_n) + G(q_n)^{\mathsf{T}}\lambda_n\big) \\ 0 &= g(q_{n+1}). \end{aligned} \tag{3.5}$$

The momentum approximation is given by $p_n = M(q_{n+1} - q_{n-1})/2h$ and does not enter the recursion (3.5). In general $G(q_n)M^{-1}p_n \neq 0$, so that the numerical solution (q_n, p_n) does not lie on the manifold $\mathcal{M}$.

An important modification, called RATTLE [And83], consists in writing the algorithm as a one-step method and to add a projection step, so that $(q_n, p_n) \in \mathcal{M}$. The algorithm is given by

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2}\Big(\nabla U(q_n) + G(q_n)^{\mathsf{T}}\theta_n\Big) \\ q_{n+1} &= q_n + hM^{-1}p_{n+1/2} \\ 0 &= g(q_{n+1}) \\ p_{n+1} &= p_{n+1/2} - \frac{h}{2}\Big(\nabla U(q_{n+1}) + G(q_{n+1})^{\mathsf{T}}\mu_{n+1}\Big) \\ 0 &= G(q_{n+1})M^{-1}p_{n+1}. \end{aligned} \tag{3.6}$$

It is symmetric, symplectic on the manifold $\mathcal{M}$, and convergent of order 2 (see [HLW06, Section VII.1] for details). Eliminating the momentum variables shows that the RATTLE approximation satisfies the two-term recursion (3.5) of SHAKE with $\lambda_n = (\theta_n + \mu_n)/2$.

The RATTLE algorithm is an excellent geometric integrator for low accuracy requirements (such as in molecular dynamics simulations). There are a few extensions of this algorithm to higher order. An easy way is by composition methods with the RATTLE scheme as basic integrator [Rei96]. Another extension is the partitioned Runge–Kutta method based on the Lobatto IIIA–IIIB pair. It is of order $2s-2$ and reduces to the RATTLE algorithm for $s = 2$ [Jay96]. The present article proposes a new extension, based on symmetric multistep methods.

The long-time behavior of symmetric linear multistep methods for unconstrained Hamiltonian systems $\ddot{q} = -\nabla U(q)$ has been studied in [HL04], see also [CH13a] for their applicability to more general Hamiltonian problems. Section 3.2 explains how these methods can be extended to constrained systems of the form (3.1). The main results on their long-time behaviour, in particular, the near-preservation of the total energy and the momentum over long time intervals, are reported in Section 3.3. The construction of stable symmetric methods is discussed in Section 3.4, and the coefficients of optimal-order methods are presented for orders 4, 6, and 8. The numerical experiments of Section 3.5 illustrate the excellent long-time behaviour of the methods in agreement with the theoretical results. Rigorous proofs are based on a backward error analysis. The long-time behaviour of "smooth" numerical solutions and their preservation of energy and momentum are discussed in Section 3.6. Bounds for parasitic solution components are the topic of Section 3.7. The results of Sections 3.6 and 3.7 are then combined to yield the main results.

3.2 Symmetric linear multistep methods

With the notation $f(q) = -\nabla U(q)$ for the force, linear multistep methods for differential-algebraic equations (3.1) are given by

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j q_{n+j} &= h^2 \sum_{j=0}^{k} \beta_j M^{-1} \Big(f(q_{n+j}) - G(q_{n+j})^{\mathsf{T}} \lambda_{n+j} \Big) \\ 0 &= g(q_{n+k}). \end{aligned} \tag{3.7}$$

For implicit methods ($\beta_k \neq 0$) this represents a nonlinear system for (q_{n+k}, λ_{n+k}). For explicit methods ($\beta_k = 0$) we insert q_{n+k} from the first relation into the second one to obtain a nonlinear equation for λ_{n+k-1}. As soon as λ_{n+k-1} is computed, the solution approximation q_{n+k} is given explicitly. The computational cost of an explicit multistep method is thus precisely the same as that for the SHAKE algorithm.

For the study of linear multistep methods it is convenient to introduce the generating polynomials

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j, \qquad \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j.$$

Throughout this article we assume that $\rho(\zeta)$ and $\sigma(\zeta)$ do not have common zeros (irreducibility). The method (3.7) is *stable* if all zeros of $\rho(\zeta)$ satisfy $|\zeta| \leq 1$, and if those of modulus one have a multiplicity not exceeding two. It is *consistent of order* r, if

$$\frac{\rho(\zeta)}{(\log \zeta)^2} - \sigma(\zeta) = \mathcal{O}\big((\zeta - 1)^r\big) \qquad \text{for} \quad \zeta \to 1. \tag{3.8}$$

In the present article we focus our interest on *symmetric* methods, which means that the coefficients satisfy

$$\alpha_j = \alpha_{k-j}, \qquad \beta_j = \beta_{k-j} \qquad \text{for all } j.$$

If a multistep method (3.7) is stable and symmetric, all zeros of $\rho(\zeta)$ are on the unit circle, and the order r is even. Furthermore, it follows from the irreducibility assumption that k is even (because symmetry implies for odd k that $\rho(-1) = \sigma(-1) = 0$), and that -1 cannot be a simple zero of $\rho(\zeta)$. The construction of explicit symmetric methods of optimal order will be discussed in Section 3.4 below.

An approximation of the momentum $p = M\dot{q}$ can be computed *a posteriori* by symmetric finite differences supplemented with a projection onto the manifold $\mathcal{M}$:

$$p_n = M \, \frac{1}{h} \sum_{j=-l}^{l} \delta_j q_{n+j} + h \, G(q_n)^{\mathsf{T}} \mu_n. \tag{3.9}$$

together with $G(q_n) M^{-1} p_n = 0$, which gives a linear system for μ_n. One typically chooses $l = k/2$, so that the approximations p_n are of the same order as q_n. This is not essential, because errors in p_n do not propagate.

Comments on the implementation. The formulation (3.7) is a straightforward extension of the SHAKE algorithm (3.5). To reduce the effect of round-off we consider momentum approximations $p_{n+1/2}$, as it was proposed in RATTLE. For explicit multistep methods this

yields

$$\begin{aligned} \sum_{j=0}^{k-1} \hat{\alpha}_j \, p_{n+j+1/2} &= h \sum_{j=1}^{k-1} \beta_j \Big(f(q_{n+j}) - G(q_{n+j})^{\mathsf{T}} \lambda_{n+j} \Big) \\ q_{n+k} &= q_{n+k-1} + h \, M^{-1} p_{n+k-1/2} \\ 0 &= g(q_{n+k}), \end{aligned} \tag{3.10}$$

where $\hat{\alpha}_j$ are the coefficients of $\rho(\zeta)/(\zeta - 1) = (\zeta - 1)\tilde{\rho}(\zeta)$. The approximation of the momenta becomes

$$\begin{aligned} p_n &= \sum_{j=-l}^{l-1} \hat{\delta}_j \, p_{n+j+1/2} + h \, G(q_n)^{\mathsf{T}} \mu_n \\ 0 &= G(q_n) M^{-1} p_n, \end{aligned} \tag{3.11}$$

where the coefficients $\hat{\delta}_j$ are given by $(\zeta - 1) \sum_{j=-l}^{l-1} \hat{\delta}_j \zeta^j = \sum_{j=-l}^{l} \delta_j \zeta^j$.

3.3 Main results

When linear multistep methods are applied to differential-algebraic equations of index 3, symmetric formulas are typically avoided because of their notorious weak instability and the standard choice is BDF schemes. There is some research on a partitioned treatment of the force term and the Lagrange multiplier (for example [AFS97]) such that also non-stiff integrators can be applied. However, little attention has been paid to long-time energy and momentum preservation with these integrators. This requires the use of symmetric methods. The present work shows that the suspected weak instability is not harmful for problems of the form (3.1) and for a special class of integrators.

For a favourable long-time behaviour we need the following properties of the generating polynomials:

$$\rho(\zeta) = 0 \quad \text{has only simple roots with the exception of the double root 1; all roots are on the unit circle.} \tag{3.12}$$

$$\sigma(\zeta) = 0 \quad \text{has only simple non-zero roots; all non-zero roots are on the unit circle.} \tag{3.13}$$

Symmetry of the method together with condition (3.12) is essential for good long-time behaviour in unconstrained problems (see [HL04]), and condition (3.13) is familiar from the convergence analysis of multistep methods for index-3 differential-algebraic equations.

For the starting values we assume

$$\begin{aligned} q_j - q(jh) &= \mathcal{O}(h^{r+2}) \quad \text{and} \quad g(q_j) = 0 \quad \text{for} \quad j = 0, \ldots, k-1 \\ \lambda_j - \lambda(jh) &= \mathcal{O}(h^r) \quad \text{for} \quad j = 1, \ldots, k-2 \end{aligned}$$

(the latter for the case of an explicit method with $\beta_{k-1} \neq 0$).

3.3.1 Energy conservation

It follows from differentiation of $H(q(t), p(t))$ that the total energy (3.3) is exactly preserved along solutions of the system (3.1). Recall that $p = M\dot{q}$.

Theorem 3.3.1. *Consider a symmetric linear multistep method (3.7) of order r with generating polynomials satisfying (3.12) and (3.13). Along the numerical solution of the constrained system (3.1) the total energy (3.3) is conserved up to $\mathcal{O}(h^r)$ over time $\mathcal{O}(h^{-r-1})$:*

$$H(q_n, p_n) = H(q_0, p_0) + \mathcal{O}(h^r) \quad \text{for} \quad nh \le h^{-r-1}.$$

The constant symbolized by $\mathcal{O}$ is independent of n and h subject to $nh \le h^{-r-1}$.

3.3.2 Momentum conservation

Constrained N-body systems preserve the total angular momentum if both the potential $U(q)$ and the constraint function $g(q)$ are invariant under rotations. More generally, the invariance properties

$$U(e^{\tau A} q) = U(q) \quad \text{and} \quad g(e^{\tau A} q) = g(q) \qquad \text{for all } \tau, q \tag{3.14}$$

with a matrix A such that MA is skew-symmetric, implies that the Lagrange function

$$\mathcal{L}(q, \dot q, \lambda) = \tfrac{1}{2}\, \dot q^{\mathsf T} M \dot q - U(q) - g(q)^{\mathsf T} \lambda$$

is invariant under the symmetry $q \mapsto e^{\tau A} q$. By Noether's theorem the momentum

$$L(q, p) = p^{\mathsf T} A q \tag{3.15}$$

is conserved along solutions of the constrained Hamiltonian system (3.1).

Theorem 3.3.2. *Consider a symmetric linear multistep method (3.7) of order r with generating polynomials satisfying (3.12) and (3.13). Along the numerical solution of the constrained system (3.1) satisfying (3.14) the momentum (3.15) is conserved up to $\mathcal{O}(h^r)$ over time $\mathcal{O}(h^{-r-1})$:*

$$L(q_n, p_n) = L(q_0, p_0) + \mathcal{O}(h^r) \quad \text{for} \quad nh \le h^{-r-1}.$$

The constant symbolized by $\mathcal{O}$ is independent of n and h subject to $nh \le h^{-r-1}$.

Remark 3.3.3. Symplectic one-step methods (like the Rattle algorithm) conserve the momentum exactly. This is not the case for linear multistep methods, because their underlying one-step method cannot be symplectic (see [HLW06, Section XV.4.1]).

3.4 Examples of higher order methods

Symmetric linear k-step multistep methods (3.7) with even k can be constructed as follows. We define the ρ-polynomial by

$$\rho(\zeta) = (\zeta - 1)^2 \prod_{j=1}^{k/2-1} (\zeta^2 + 2a_j \zeta + 1),$$

where a_j are distinct real numbers satisfying $-1 < a_j < 1$. This implies the assumption (3.12). The order condition (3.8) then uniquely determines the σ-polynomial of degree $k-1$ such that the method is explicit and of order $r = k$. The resulting method is symmetric.

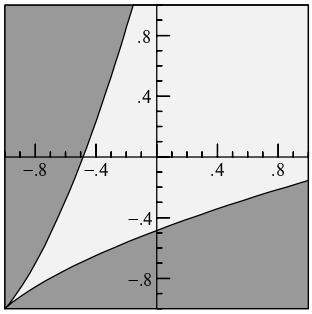


Figure 3.1 – The dark grey region shows the (a_1, a_2) values for which the corresponding σ-polynomial (case $k = 6$) has all non-zero roots on the unit circle.

3.4.1 Coefficients of methods up to order 8

For methods up to order 8 we investigate for which values of a_j the corresponding σ-polynomial satisfies assumption (3.13).

Order $r = k = 4$: The σ-polynomial is given by

$$\sigma(\zeta) = (7 + a_1)(\zeta^3 + \zeta)/6 + (-1 + 5a_1)\zeta^2/3.$$

We see that condition (3.13) is satisfied for all choices of $-1 < a_1 < 1$.

Order $r = k = 6$*:* The σ-polynomial is given by

$$\sigma(\zeta) = \alpha(\zeta^5 + \zeta) + \beta(\zeta^4 + \zeta^2) + \gamma\zeta^3$$

with

$$\begin{aligned} \alpha &= (79 + 9\,(a_1 + a_2) - a_1a_2)/60 \\ \beta &= (-14 + 26\,(a_1 + a_2) + 6\,a_1a_2)/15 \\ \gamma &= (97 + 7\,(a_1 + a_2) + 97\,a_1a_2)/30. \end{aligned}$$

It has double zeros on the unit circle if $\beta^2 = 4\alpha(\gamma - 2\alpha)$. This curve separates the region where all non-zero roots of $\sigma(\zeta) = 0$ are of modulus 1, from that where at least one root is outside the unit disc, see Figure 3.1.

Order $r = k = 8$*:* The σ-polynomial is given by

$$\sigma(\zeta) = \alpha(\zeta^7 + \zeta) + \beta(\zeta^6 + \zeta^2) + \gamma(\zeta^5 + \zeta^3) + \delta\zeta^4$$

with

$$\begin{aligned} \alpha &= (10993 + 1039\,s_1 - 95\,s_2 + 31\,s_3)/7560 \\ \beta &= (-2215 + 2279\,s_1 + 473\,s_2 - 73\,s_3)/1260 \\ \gamma &= (16661 + 491\,s_1 + 8261\,s_2 + 2171\,s_3)/2520 \\ \delta &= (-8723 + 7027\,s_1 + 1357\,s_2 + 12067\,s_3)/1890, \end{aligned}$$

where $s_1 = a_1 + a_2 + a_3$, $s_2 = a_1a_2 + a_1a_3 + a_2a_3$, and $s_3 = a_1a_2a_3$. We remark that none of the methods presented in Table 7.1 of [HLW06, Sect. XV.7] (including a method proposed in [QT90]) satisfies the condition (3.13). However, if two among the parameters a_j are not too far from -1 and the third one is not far from 1, then condition (3.13) is satisfied. In particular, the choice

$$a_1 = -0.8, \qquad a_2 = -0.4, \qquad a_3 = 0.7$$

gives a method that satisfies both conditions (3.12) and (3.13).

Coefficients $\hat{\delta}_j$ of (3.11): Symmetric multistep methods of order $r = k$ are complemented by a difference formula (3.11) for the computation of the momenta. We use the coefficients $\hat{\delta}_j$, $j = -k/2, \dots, k/2 - 1$ given by:

$$\begin{aligned} k = 2: &\quad \tfrac{1}{2}\big(1, 1\big), \\ k = 4: &\quad \tfrac{1}{12}\big(-1, 7, 7, -1\big), \\ k = 6: &\quad \tfrac{1}{60}\big(1, -8, 37, 37, -8, 1\big), \\ k = 8: &\quad \tfrac{1}{840}\big(-3, 29, -139, 533, 533, -139, 29, -3\big). \end{aligned}$$

3.4.2 Linear stability - interval of periodicity

When applied to the harmonic oscillator $\ddot{q} = -\omega^2 q$, the numerical solution of a symmetric linear multistep method is determined by the roots of the equation

$$\rho(\zeta) + (h\omega)^2 \sigma(\zeta) = 0. \tag{3.16}$$

According to [LW76] we say that the method has interval of periodicity $(0, \Omega)$ if, for all $h\omega \in (0, \Omega)$, these roots are bounded by 1. For the method (3.5) of order 2 the interval of periodicity is $(0, 2)$, which implies that the method is stable only for $0 \le h\omega < 2$.

The assumption (3.12) and the symmetry of the method imply that the roots of (3.16) stay on the unit circle for small $h\omega > 0$. Consequently, the interval of periodicity is always non-empty.

Order $r = k = 4$: Studying the roots of (3.16) as a function of $h\omega$, one observes that a root can leave the unit circle only when two roots collapse at the point -1. This implies that

$$\Omega = \sqrt{-\frac{\rho(-1)}{\sigma(-1)}} = \sqrt{\frac{6(1 - a_1)}{2 - a_1}}.$$

For orders $r = k \ge 6$, the value Ω of the interval of periodicity can be computed numerically as function of the parameters a_j. For example, for values of (a_1, a_2) in the dark grey region of Figure 3.1, we have $0 < \Omega < 0.8$, and the largest values of Ω are attained away from the border of the square.

3.5 Numerical experiments

We have implemented symmetric linear multistep methods as proposed in Section 3.2. The following numerical experiments illustrate an excellent long-time behaviour for constrained Hamiltonian systems confirming our theoretical results.

Example 3.5.1 (Triple pendulum)**.** We consider three connected mathematical pendulums moving in the plane and suspended at the origin. Denoting by (q_1, q_2), (q_3, q_4), (q_5, q_6) their endpoints, the constraints $g_i(q) = 0$ are given by

$$q_1^2 + q_2^2 = 1, \quad (q_3 - q_1)^2 + (q_4 - q_2)^2 = 1, \quad (q_5 - q_3)^2 + (q_6 - q_4)^2 = 1.$$

The potential due to gravity is $U(q) = q_2 + q_4 + q_6$. We consider initial positions

$$q(0) = \Big(\tfrac{1}{2}, -\tfrac{\sqrt{3}}{2}, \tfrac{1}{2} + \tfrac{\sqrt{2}}{2}, -\tfrac{\sqrt{3}}{2} - \tfrac{\sqrt{2}}{2}, \tfrac{1}{2} + \tfrac{\sqrt{2}}{2} + 1, -\tfrac{\sqrt{3}}{2} - \tfrac{\sqrt{2}}{2}\Big),$$

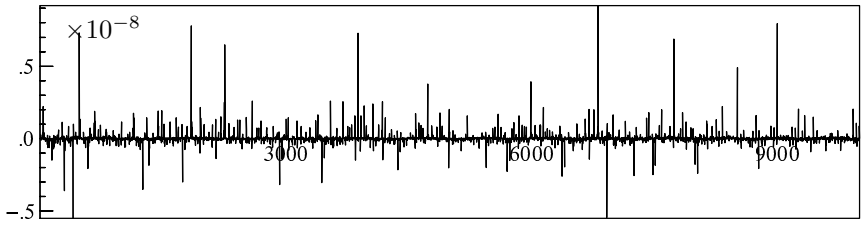


Figure 3.2 – Triple pendulum: error in the Hamiltonian for the symmetric multistep method (A) of order $r = k = 6$, applied with step size $h = 0.01$.

which correspond to angles of 30°, 45°, and 90°, and momenta $p(0) = (0, \dots, 0)$. This choice of initial values produces a chaotic behaviour of the solution.

To illustrate the necessity of the condition (3.13) we apply two symmetric multistep schemes of order $r = k = 6$, which are constructed as explained in Section 3.4:

(A) $a_1 = -0.7$, $a_1 = 0.4$, the σ-polynomial satisfies (3.13);

(B) $a_1 = -0.1$, $a_1 = 0.4$, the σ-polynomial does not satisfy (3.13).

The numerical Hamiltonian is shown in Figure 3.2 for method (A). The error remains bounded without any drift, and an application with reduced step size shows that it is of size $\mathcal{O}(h^6)$. For the step size $h = 0.01$ this behaviour can be observed on much longer intervals than shown in Figure 3.2 (numerically verified on $[0, 200\,000]$). For method (B), the error explodes after about 130 steps (independent of the step size). This is due to the fact that the σ-polynomial has a zero of modulus larger than 1.

Let us remark that the above description of the problem is extremely simple compared to the equations using minimal coordinates (angles). The long-time behaviour of method (A) in Figure 3.2 should be compared with that of partitioned multistep methods applied to the equation in minimal coordinates (see [CH13a, Section I.3]), where no energy preservation could be achieved in the chaotic regime.

Example 3.5.2 (Two-body problem on the sphere)**.** We consider two particles moving on the unit sphere which are attracted by each other. As potential we take

$$U(q) = -\frac{\cos\vartheta}{\sin\vartheta}, \qquad \cos\vartheta = \langle Q_1, Q_2 \rangle, \tag{3.17}$$

where $Q_1 = (q_1, q_2, q_3)^\mathsf{T}$, $Q_2 = (q_4, q_5, q_6)^\mathsf{T}$ are the positions of the two particles, and ϑ is their distance along a geodesics. The constraints are

$$g_1(q) = Q_1^\mathsf{T} Q_1 - 1, \qquad g_2(q) = Q_2^\mathsf{T} Q_2 - 1.$$

The equations of motion have the total angular momentum

$$L(p, q) = Q_1 \times P_1 + Q_2 \times P_2$$

as conserved quantity. Here, we use the notation $P_1 = (p_1, p_2, p_3)^\mathsf{T}$, $P_2 = (p_4, p_5, p_6)^\mathsf{T}$.

In view of a comparison with the experiments of [HH03] we consider initial values given in spherical coordinates by

$$Q_i = \big(\cos\phi_i \sin\theta_i, \sin\phi_i \sin\theta_i, \cos\theta_i\big)^\mathsf{T}$$

with $(\phi_1, \theta_1) = (0.8, 0.6)$ and $(\phi_2, \theta_2) = (0.5, 1.5)$ for the positions, and with $(\dot\phi_1, \dot\theta_1) = (1.1, -0.2)$ and $(\dot\phi_2, \dot\theta_2) = (-0.8, 0.0)$ for the velocities. In our numerical experiment we

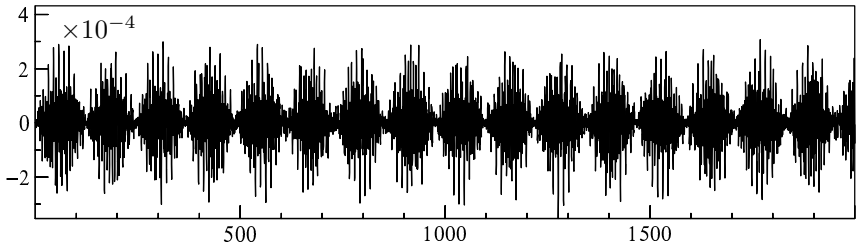


Figure 3.3 – Two-body problem on the sphere: error in the first component of the angular momentum for a symmetric multistep method of order $r = k = 8$ applied with step size $h = 0.02$.

consider the multistep method of order $r = k = 8$ with parameters $a_1 = -0.8$, $a_2 = -0.4$, and $a_3 = 0.7$ (see Section 3.4). Figure 3.3 shows the error in the first component of the angular momentum. In perfect agreement with Theorem 3.3.2 we have an error of size $\mathcal{O}(h^8)$, and no drift can be observed over long time intervals (this is numerically checked on intervals as long as $T = 10^6$). A similar behavior is true for the other two components of the angular momentum and for the total energy.

Since the same problem was treated numerically in [HH03, Section 5.3] with a composition method of order 8 and Rattle as basic integrator, this is the moment to say a few words on a comparison between symmetric linear multistep methods (as considered in the present work) and high order composition methods. Both are explicit and can have high order of accuracy. Which one is more efficient? From the experiment of [HH03] we see that an error in the energy of size $8 \cdot 10^{-6}$ is obtained with step size $h = 0.15$. For the composition method of order 8 (with 17 Rattle applications per step) this corresponds to 226 666 force evaluations for an integration over an interval of length 2 000. With the multistep method we need a step size $h = 0.0125$ to achieve the same accuracy. This corresponds to 160 000 force evaluations, which is an improvement of about 30%. Needless to say that such comparisons are problem dependent. We believe that it is important to consider both approaches.

Example 3.5.3 (Rigid body - heavy top)**.** The configuration space of a rigid body with one point fixed is the rotation group $SO(3)$. The motion is described by an orthogonal matrix $Q(t)$ that satisfies

$$\begin{aligned} \ddot{Q}D &= -\nabla_Q U(Q) - Q\Lambda \\ 0 &= Q^{\mathsf{T}}Q - I, \end{aligned} \tag{3.18}$$

where the diagonal matrix $D = \operatorname{diag}(d_1, d_2, d_3)$ is related to the moments of inertia I_1, I_2, I_3 via

$$I_1 = d_2 + d_3, \quad I_2 = d_3 + d_1, \quad I_3 = d_1 + d_2,$$

and Λ is a symmetric matrix consisting of Lagrange multipliers. The potential, due to gravity, is given by $U(Q) = q_{33}$. For a more detailed description see [HLW06, Section VII.5]. With $P = \dot{Q}D$, we are thus concerned with the Hamiltonian

$$H(P, Q) = \tfrac{1}{2} \operatorname{trace}(PD^{-1}P^{\mathsf{T}}) + U(Q).$$

The equation (3.18) is of the form (3.1) and satisfies the regularity condition (3.2).

With the abbreviation

$$\begin{aligned}\hat{\alpha}_{k-1}\widetilde{P}_{n+k-1/2} &= -\sum_{j=0}^{k-2}\hat{\alpha}_j P_{n+j+1/2} - h\beta_{k-1}\nabla_Q U(Q_{n+k-1}) \\ &\quad - h\sum_{j=1}^{k-2}\beta_j\Big(\nabla_Q U(Q_{n+j}) + Q_{n+j}\Lambda_{n+j}\Big)\end{aligned} \tag{3.19}$$

and $\gamma_{k-1} = \beta_{k-1}/\hat{\alpha}_{k-1}$ the multistep formula (3.10) becomes

$$\begin{aligned}P_{n+k-1/2} &= \widetilde{P}_{n+k-1/2} - h\gamma_{k-1}Q_{n+k-1}\Lambda_{n+k-1} \\ Q_{n+k} &= Q_{n+k-1} + hP_{n+k-1/2}D^{-1}.\end{aligned}$$

These formulas are similar to those for the Rattle algorithm. We work with the auxiliary matrix

$$\Omega_{n+k-1} = Q_{n+k-1}^{\mathsf{T}}P_{n+k-1/2}D^{-1},$$

so that, for given $(Q_{n+j}, P_{n+j-1/2}, \Lambda_{n+j-1}), j \le k-1$, the approximations Q_{n+k}, $P_{n+k-1/2}$, Λ_{n+k-1} are obtained as follows:

- compute $\widetilde{P}_{n+k-1/2}$ from (3.19);
- find an orthogonal matrix $I + h\Omega_{n+k-1}$ such that

  $$\Omega_{n+k-1}D = Q_{n+k-1}^{\mathsf{T}}\widetilde{P}_{n+k-1/2} - h\gamma_{k-1}\Lambda_{n+k-1}$$

  holds with a symmetric matrix Λ_{n+k-1};
- compute $Q_{n+k} = Q_{n+k-1}(I + h\Omega_{n+k-1})$;
- compute $P_{n+k-1/2} = Q_{n+k-1}\Omega_{n+k-1}D$.

Steps 1, 3, and 4 are straightforward computations. Step 2 requires the iterative solution of a nonlinear (quadratic) equation for Λ_{n+k-1}.

If an approximation P_n is required for output, it can be obtained from $P_n = Q_n\Omega_n$, where Ω_n and the symmetric matrix K_n are given by

$$\begin{aligned}\Omega_n &= Q_n^{\mathsf{T}}\sum_{j=-l}^{l-1}\hat{\delta}_j P_{n+j+1/2} + hK_n \\ 0 &= \Omega_n D^{-1} + D^{-1}\Omega_n^{\mathsf{T}}.\end{aligned}$$

These two equations constitute a linear system for Ω_n and K_n. The computations can be done efficiently by representing orthogonal matrices in terms of quaternions (see [HLW06, Section VII.5.3]).

3.6 Backward error analysis for smooth numerical solutions

For the proof of the main theoretical results we adapt the presentation of [HL04] to the case of constrained Hamiltonian systems. For the problem (3.1) we use the notation $f(q) = -\nabla U(q)$ and, without loss of generality, we assume the mass matrix M to be the identity, i.e., $M = I$.

3.6.1 Modified differential-algebraic system

Proposition 3.6.1 (Existence). *Let a consistent linear multistep method (3.7) be applied to the problem (3.1). Then, there exist unique h-independent functions $f_j(q,v)$ such that for every truncation index N, every solution $(y(t),\mu(t))$ of the modified differential-algebraic system*

$$\begin{aligned} \ddot{y} &= f(y) + hf_1(y,\dot{y}) + \cdots + h^{N-1}f_{N-1}(y,\dot{y}) - G(y)^{\mathsf{T}}\mu \\ 0 &= g(y) \end{aligned} \tag{3.20}$$

satisfies the multistep relation

$$\sum_{j=0}^{k} \alpha_j y(t+jh) = h^2 \sum_{j=0}^{k} \beta_j \Big(f\big(y(t+jh)\big) - G\big(y(t+jh)\big)^{\mathsf{T}} \mu(t+jh) \Big) + \mathcal{O}(h^{N+2}). \tag{3.21}$$

If the method is of order r, then $f_j(q,v) = 0$ for $j < r$. If it is symmetric, then $f_j(q,v) = 0$ for all odd j, and $f_j(q,-v) = f_j(q,v)$ for all even j.

Proof. We write the Taylor series of a function as $z(t+h) = e^{hD}z(t)$, where D denotes differentiation with respect to time. The identity (3.21) is then of the form

$$\rho(e^{hD})y = h^2\sigma(e^{hD})\big(f(y) - G(y)^{\mathsf{T}}\mu\big) + \mathcal{O}(h^{N+2}). \tag{3.22}$$

With $x^2\sigma(e^x)/\rho(e^x) = 1 + \vartheta_1 x + \vartheta_2 x^2 + \ldots$ this relation becomes

$$\ddot{y} = \big(1 + \vartheta_1 hD + \vartheta_2 h^2 D^2 + \ldots\big)\big(f(y) - G(y)^{\mathsf{T}}\mu\big) + \mathcal{O}(h^N). \tag{3.23}$$

With the exception of the h-independent term we replace $\mu(t)$ by $\mu(y(t),\dot{y}(t))$, where $\mu(q,v)$ is the expression obtained by differentiating twice the algebraic relation in (3.20). The coefficient functions $f_j(q,v)$ can then be obtained exactly as in the non-constrained case of [HL04]. □

In the modified differential-algebraic system (3.20) we have achieved uniqueness of the coefficient functions by imposing the term with the Lagrange multiplier to be independent of h.

3.6.2 Modified energy

We still assume that $M = I$ so that the momenta equal the velocities, $p = \dot{q}$. In this situation the total energy is given by

$$H(q,p) = \tfrac{1}{2}p^{\mathsf{T}}p + U(q).$$

It is preserved along the flow of the differential-algebraic system (3.1).

Proposition 3.6.2 (Energy preservation). *Consider a symmetric multistep method of order r applied to (3.1). Then, there exist unique h-independent functions $H_j(q,p)$ such that for every truncation index N the modified energy*

$$H_h(q,p) = H(q,p) + h^r H_r(q,p) + h^{r+2}H_{r+2}(q,p) + \ldots,$$

truncated at the $\mathcal{O}(h^N)$ term, satisfies

$$\frac{d}{dt}\,H_h\big(y(t),\dot{y}(t)\big) = \mathcal{O}(h^N)$$

along solutions of the modified differential-algebraic system (3.20).

Proof. Instead of dividing (3.22) by $\rho(e^{hD})$, we divide by $\sigma(e^{hD})$. This yields

$$\big(1+\gamma_1 hD+\gamma_2 h^2D^2+\ldots\big)\,\ddot y = -\nabla U(y) - G(y)^{\mathsf T}\mu + \mathcal{O}(h^N) \tag{3.24}$$

with coefficients γ_j given by $\rho(e^x)/(x^2\sigma(e^x)) = 1+\gamma_1 x+\gamma_2 x^2+\ldots$. We take the scalar product with $\dot y$ and note that $G(y)\dot y = 0$, which follows from $g(y)=0$ by differentiation with respect to time. The rest of the proof is the same as that of Proposition 1 in [HL04].□

3.6.3 Modified momentum

We assume that $M = I$ and that A is a skew-symmetric matrix for which the invariance (3.14) holds.

Proposition 3.6.3 (Momentum preservation)**.** *Consider a symmetric multistep method of order r applied to (3.1). Then, there exist unique h-independent functions $L_j(q,p)$ such that for every truncation index N the modified momentum*

$$L_h(q,p) = L(q,p) + h^r L_r(q,p) + h^{r+2}L_{r+2}(q,p) + \ldots,$$

truncated at the $\mathcal{O}(h^N)$ term, satisfies

$$\frac{d}{dt}\,L_h\big(y(t),\dot y(t)\big) = \mathcal{O}(h^N)$$

along solutions of the modified differential-algebraic system (3.20).

Proof. We take the scalar product of (3.24) with Ay and note that the invariance (3.14) implies

$$f(y)^{\mathsf T}Ay = 0 \quad\text{and}\quad G(y)Ay = 0 \qquad \text{for all } y.$$

The rest of the proof is the same as that of Proposition 2 in [HL04]. □

3.7 Long-term analysis of parasitic solution components

We consider irreducible, stable, symmetric linear multistep methods (3.7), we denote the double root of $\rho(\zeta)=0$ by $\zeta_0=1$, and we assume that the remaining roots $\zeta_i, \zeta_{-i}=\overline{\zeta}_i$ for $1\le i<k/2$ are simple. As a consequence of stability and symmetry we have $|\zeta_i|=1$. Furthermore, we denote by $\zeta_i,\zeta_{-i}=\overline{\zeta}_i$ for $k/2\le i<k$ complex pairs of roots of $\sigma(\zeta)=0$ (not including 0 for explicit methods).

We consider the index set $\mathcal{I}_\rho = \{i\in\mathbb{Z}\,;\,1\le|i|<k/2\}$ corresponding to the roots of $\rho(\zeta)=0$ different from 1, and the index set $\mathcal{I}_\sigma=\{i\in\mathbb{Z}\,;\,k/2\le|i|<k-l\}$ (with $l=0$ for implicit methods, and $l>0$ else) corresponding to the non-zero roots of $\sigma(\zeta)=0$. We denote $\mathcal{I}=\mathcal{I}_\rho\cup\mathcal{I}_\sigma$.

3.7.1 Linear problems with constant coefficients

To motivate the analysis of this section we consider the linear problem

$$\begin{aligned} \ddot q &= -A\,q - G^{\mathsf T}\lambda \\ 0 &= G\,q, \end{aligned} \tag{3.25}$$

where $q \in \mathbb{R}^d$, $\lambda \in \mathbb{R}^m$, the matrix A is symmetric, and G is of full rank. For this problem the multistep formula (3.7) reads

$$\sum_{j=0}^{k} \alpha_j q_{n+j} = -h^2 \sum_{j=0}^{k} \beta_j (A\, q_{n+j} + G^{\mathsf{T}} \lambda_{n+j}), \qquad G\, q_{n+k} = 0. \tag{3.26}$$

If the initial values are consistent, i.e., $G\, q_j = 0$ for $j = 0, \ldots, k-1$, then $G\, q_n = 0$ for all $n \geq 0$, and a multiplication by G of the multistep relation yields

$$\sum_{j=0}^{k} \beta_j\, G(A\, q_{n+j} + G^{\mathsf{T}} \lambda_{n+j}) = 0, \tag{3.27}$$

which permits to eliminate the Lagrange multipliers from the multistep formula. We thus obtain

$$\sum_{j=0}^{k} \alpha_j q_{n+j} = -h^2 \sum_{j=0}^{k} \beta_j \Big(I - G^{\mathsf{T}} (G G^{\mathsf{T}})^{-1} G \Big) A\, q_{n+j}.$$

This formula shows that the numerical solution $\{q_n\}$ depends only on the starting values $q_0, \ldots, q_{k-1}$, and is not affected by $\lambda_0, \ldots, \lambda_{k-1}$. Since we are concerned with a linear homogeneous difference equation with characteristic polynomial $\rho(\zeta)$ for $h = 0$, its solution is of the form

$$q_n = y(nh) + \sum_{i \in \mathcal{I}_\rho} \zeta_i^n\, z_i(nh), \tag{3.28}$$

where $y(t)$ and $z_i(t)$ are smooth functions in the sense that all their derivatives are bounded independently of h. The Lagrange multiplier is obtained from the difference relation (3.27) and satisfies

$$\lambda_n = -(G G^{\mathsf{T}})^{-1} G A\, q_n + \sum_{i \in \mathcal{I}_\sigma} \zeta_i^n \nu_i$$

with constant vectors ν_i that are determined by the initial approximations $\lambda_0, \ldots, \lambda_{k-1}$ for implicit methods, and by $\lambda_1, \ldots, \lambda_{k-2}$ for methods satisfying $\beta_k = 0$ and $\beta_{k-1} \neq 0$. Whereas only the zeros of the ρ-polynomial are important for the approximations $\{q_n\}$, also those of the σ-polynomial come into the game for the Lagrange multipliers $\{\lambda_n\}$.

3.7.2 Differential-algebraic system for parasitic solution components

Motivated by the analysis for the linear problem we aim at writing the numerical solution in the form (3.28) also for nonlinear problems. Due to the dependence of G on q we have to take the sum over $\mathcal{I}_\rho$ and $\mathcal{I}_\sigma$. It is easy to guess that $y(t)$ will be a solution of (3.20). It remains to study the smooth functions $z_i(t)$.

Proposition 3.7.1 (Differential-algebraic system). *Consider a symmetric linear multistep (3.7) of order r and assume that, with exception of the double root $\zeta_0 = 1$, all roots of $\rho(\zeta)$ are simple. For $i \in \mathcal{I}_\rho$ we let $\theta_i = \sigma(\zeta_i)/(\zeta_i \rho'(\zeta_i))$. We further assume that all non-zero roots of $\sigma(\zeta)$ are simple and of modulus 1.*

Then, there exist h-independent matrix-valued functions $A_{i,l}(y, v)$, $B_{i,l}(y, v)$, and $C_{i,l}(y, v)$, such that for every truncation index M and for every solution of the combined system (3.20) and

$$\begin{aligned} \dot z_i &= \big(h A_{i,1}(y, \dot y) + \cdots + h^{M-1} A_{i,M-1}(y, \dot y) \big) z_i - \theta_i h\, G(y)^{\mathsf{T}} \nu_i \\ 0 &= G(y)\, z_i \end{aligned} \tag{3.29}$$

for $i \in \mathcal{I}_\rho$, and

$$\begin{aligned} \dot\nu_i &= \big(B_{i,0}(y,\dot y) + \cdots + h^{M-3} B_{i,M-3}(y,\dot y)\big)\nu_i \\ z_i &= \big(h^3 C_{i,3}(y,\dot y) + \cdots + h^M C_{i,M}(y,\dot y)\big)\nu_i \end{aligned} \tag{3.30}$$

for $i \in \mathcal{I}_\sigma$, with initial values satisfying $z_{-i}(0) = \overline{z}_i(0)$ and $\nu_{-i}(0) = \overline{\nu}_i(0)$ the following holds: as long as $\|z_i(t)\| \le \delta$ for all $i \in \mathcal{I}_\rho$ and $h^2 \|G(y(t))^{\mathsf T} \nu_i(t)\| \le \delta$ for $i \in \mathcal{I}_\sigma$ (with sufficiently small δ), the functions[1]

$$\widehat y(t) = y(t) + \sum_{i\in\mathcal{I}} \zeta_i^{t/h}\, z_i(t), \qquad \widehat\mu(t) = \mu(t) + \sum_{i\in\mathcal{I}} \zeta_i^{t/h}\, \nu_i(t) \tag{3.31}$$

satisfy $g\big(\widehat y(t)\big) = \mathcal{O}(\delta^2)$ and

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j \widehat y(t+jh) &= h^2 \sum_{j=0}^{k} \beta_j \Big(f\big(\widehat y(t+jh)\big) - G\big(\widehat y(t+jh)\big)^{\mathsf T} \widehat\mu(t+jh)\Big) \\ &\qquad + \mathcal{O}(h^{N+2} + h^{M+1}\delta + \delta^2). \end{aligned} \tag{3.32}$$

Proof. Taylor expansion yields

$$f\big(\widehat y(t)\big) = f\big(y(t)\big) + \sum_{i\in\mathcal{I}} \zeta_i^{t/h} f'\big(y(t)\big)\, z_i(t) + \mathcal{O}(\delta^2),$$

and similarly

$$\begin{aligned} G\big(\widehat y(t)\big)^{\mathsf T} \widehat\mu(t) &= G\big(y(t)\big)^{\mathsf T} \mu(t) \\ &\quad + \sum_{i\in\mathcal{I}} \zeta_i^{t/h} \Big(G\big(y(t)\big)^{\mathsf T} \nu_i(t) + \big(G'\big(y(t)\big)\, z_i(t)\big)^{\mathsf T} \mu(t)\Big) + \mathcal{O}(h^{-2}\delta^2), \end{aligned}$$

because we have $h^2\nu_i(t) = \mathcal{O}(\delta)$ on the considered interval. These relations show that (3.32) is satisfied if the functions $y(t)$ and $\mu(t)$ are solutions of (3.22) and the functions $z_i(t)$ and $\nu_i(t)$ satisfy the relation

$$\rho(\zeta_i e^{hD}) z_i = h^2 \sigma(\zeta_i e^{hD}) \Big(f'(y) z_i - G(y)^{\mathsf T} \nu_i - (G'(y) z_i)^{\mathsf T} \mu\Big) + \mathcal{O}(h^{M+1}\delta).$$

Similar to the proof of Proposition 3.6.1 we divide by $\rho(\zeta_i e^{hD})$ and use the expansion

$$\frac{\sigma(\zeta_i e^x)}{\rho(\zeta_i e^x)} = \theta_{i,-1} x^{-1} + \theta_{i,0} + \theta_{i,1}\, x + \theta_{i,2}\, x^2 + \dots \ .$$

For $i \in \mathcal{I}_\rho$, where $\theta_{i,-1} \ne 0$, the above equation for z_i becomes

$$\begin{aligned} \dot z_i = h\Big(\theta_{i,-1} + \theta_{i,0} hD + \dots\Big) \Big(f'(y) z_i - G(y)^{\mathsf T} \nu_i - (G'(y) z_i)^{\mathsf T} \mu\Big) \\ + \mathcal{O}(h^M \delta). \end{aligned} \tag{3.33}$$

As in the proof of Proposition 3.6.1, the elimination of higher derivatives gives a differential equation of the form (3.29). The Lagrange multipliers ν_i are determined by the condition $G(y) z_i = 0$, which is needed for having $g(\widehat y) = \mathcal{O}(\delta^2)$.

[1] Note that the analogous expression in [Hai99] and [HL04] has a sum over an index set that includes also finite products of ζ_i. This is not necessary for the investigations of the present work.

For $i \in \mathcal{I}_\sigma$, where $\theta_{i,-1} = \theta_{i,0} = 0$ and $\theta_{i,1} \neq 0$, the equation for z_i becomes

$$z_i = h^2\Big(\theta_{i,1}hD + \theta_{i,2}(hD)^2 + \ldots\Big)\Big(f'(y)z_i - G(y)^\mathsf{T}\nu_i - (G'(y)z_i)^\mathsf{T}\mu\Big) + \mathcal{O}(h^{M+1}\delta). \tag{3.34}$$

We insert the equations (3.30) into (3.34) and express the higher derivatives of z_i and ν_i recursively in terms of ν_i. Equating powers of h yields for the h^3 term $C_{i,3} = -\theta_{i,1}((G'(y)\dot{y})^\mathsf{T} + G(y)^\mathsf{T}B_{i,0})$. The condition $G(y)z_i = 0$ yields $GC_{i,3} = 0$, so that multiplication of the above equation with $G(y)$ determines $B_{i,0}$, which in turn gives $C_{i,3}$. The same construction is used to determine the matrices for higher powers of h. This construction ensures that the relations (3.33) and (3.34) are satisfied, which completes the proof. □

Having found differential-algebraic equations for the smooth and parasitic solution components, we still need initial values for the combined system (3.20), (3.29), (3.30). We note that for given $y(0) = y_0$ and $\dot{y}(0) = \dot{y}_0$ satisfying $G(y_0)\dot{y}_0 = 0$, the function $\mu(t)$ is determined for all $t \geq 0$. For $i \in \mathcal{I}_\rho$, if in addition to $y_0, \dot{y}_0$ also $z_i(0) = z_{i,0}$ satisfying $G(y_0)z_{i,0} = 0$ is given, the functions $z_i(t)$ and $\nu_i(t)$ are determined for all $t \geq 0$ by (3.29). For $i \in \mathcal{I}_\sigma$ we need the initial value $\nu_i(0) = \nu_{i,0}$, which then determines $\nu_i(t)$ and $z_i(t)$ for all t by (3.30).

The next lemma shows how initial values $y_0, \dot{y}_0, z_{i,0}$ $(i \in \mathcal{I}_\rho), \nu_{i,0}$ $(i \in \mathcal{I}_\sigma)$ can be obtained form starting approximations $q_0, q_1, \ldots, q_{k-1}$ and $\lambda_1, \ldots, \lambda_{k-2}$ for explicit methods satisfying $\beta_{k-1} \neq 0$. In general, there are $k-2l$ starting values $\lambda_l, \ldots, \lambda_{k-l-1}$, where l is the multiplicity of the root 0 in $\sigma(\zeta)$. In the following, we only consider the most interesting case $l = 1$ for simplicity.

Proposition 3.7.2 (Initial values)**.** *Under the assumptions of Proposition 3.7.1 consider the starting values $q_0, q_1, \ldots, q_{k-1}$ and $\lambda_1, \ldots, \lambda_{k-2}$. We assume that $g(q_j) = 0$, $q_j - q(jh) = \mathcal{O}(h^s)$ for $j = 0, 1, \ldots, k-1$, $\lambda_j - \lambda(jh) = \mathcal{O}(h^{s-2})$ for $j = 1, \ldots, k-2$, where $(q(t), \lambda(t))$ is a solution of (3.1) and $1 \leq s \leq r+2$. Then there exist (locally) unique consistent initial values $y_0, \dot{y}_0, z_{i,0}$ $(i \in \mathcal{I}_\rho), \nu_{i,0}$ $(i \in \mathcal{I}_\sigma)$ of the combined system (3.20), (3.29), (3.30) such that its solution satisfies*

$$q_j = y(jh) + \sum_{i\in\mathcal{I}} \zeta_i^j\, z_i(jh) + G\big(y(jh)\big)^\mathsf{T}\kappa_j, \quad j = 0, \ldots, k-1 \tag{3.35}$$

$$\lambda_j = \mu(jh) + \sum_{i\in\mathcal{I}} \zeta_i^j\, \nu_i(jh), \qquad j = 1, \ldots, k-2, \tag{3.36}$$

where, with $\delta = h^s$, we have $\kappa_j = \mathcal{O}(\delta^2)$. The initial values satisfy $z_{-i,0} = \overline{z}_{i,0}$ for $i \in \mathcal{I}_\rho$ and $\nu_{-i,0} = \overline{\nu}_{i,0}$ for $i \in \mathcal{I}_\sigma$, and

$$\begin{aligned} &y_0 - q(0) = \mathcal{O}(\delta), && h\dot{y}_0 - h\dot{q}(0) = \mathcal{O}(\delta), \\ &z_{i,0} = \mathcal{O}(\delta),\ i \in \mathcal{I}_\rho, && h^2\nu_{i,0} = \mathcal{O}(\delta),\ i \in \mathcal{I}_\sigma. \end{aligned} \tag{3.37}$$

Proof. The equations (3.35)-(3.36) together with $g(y_0) = 0$, $G(y_0)\dot{y}_0 = 0$, and $G(y_0)z_{i,0} = 0$ constitute a nonlinear system $F(\mathbf{x}) = 0$ for the vector $\mathbf{x} = \big(y_0, h\dot{y}_0, (z_{i,0}; i \in \mathcal{I}_\rho), (h^2\nu_{i,0}; i \in \mathcal{I}_\sigma), (\kappa_j; j = 0, \ldots, k-1)\big)$. An approximation of its solution is $\mathbf{x}_0 = \big(q(0), h\dot{q}(0), 0, \ldots, 0\big)$. Using assumption (3.2), the inverse of the Jacobian matrix $F'(\mathbf{x}_0)$ can be shown to be bounded, and we have $F(\mathbf{x}_0) = \mathcal{O}(\delta)$. A convergence theorem for Newton's method thus

proves the estimates (3.37). A sharper estimate for the variables κ_j follows from the fact that

$$\begin{aligned} 0 = g(q_j) - g(y(jh)) &= G(y(jh))(q_j - y(jh)) + \mathcal{O}\big(\|q_j - y(jh)\|^2\big) \\ &= G(y(jh))G(y(jh))^{\mathsf{T}}\kappa_j + \mathcal{O}(\delta^2), \end{aligned}$$

because $G(y)G(y)^{\mathsf{T}}$ has a bounded inverse. We have used that $q_j - y(jh) = q_j - q(jh) + q(jh) - y(jh)$ is bounded by $\mathcal{O}(\delta + h^{r+2})$. □

For given $q_0, \ldots, q_{k-1}$ and $\lambda_1, \ldots, \lambda_{k-2}$ (in the case of explicit methods) the numerical approximations q_k and λ_{k-1} are simultaneously obtained from (3.7).

Proposition 3.7.3 (Local error). *Under the assumptions of Propositions 3.7.1 and 3.7.2 consider the solution of the combined system (3.20), (3.29), (3.30) that corresponds to the starting approximations $q_0, \ldots, q_{k-1}$ and $\lambda_1, \ldots, \lambda_{k-2}$. Then the numerical approximation after one step satisfies*

$$\begin{aligned} q_k &= y(kh) + \sum_{i\in\mathcal{I}} \zeta_i^k\, z_i(kh) + \mathcal{O}(h^{N+2} + h^{M+1}\delta + \delta^2), \\ \lambda_{k-1} &= \mu((k-1)h) + \sum_{i\in\mathcal{I}} \zeta_i^k\, \nu_i((k-1)h) + \mathcal{O}(h^N + h^{M-1}\delta + h^{-2}\delta^2). \end{aligned}$$

Proof. Using the notation (3.31) and subtracting (3.32) from the multistep formula (3.7), it follows from Proposition 3.7.2 that

$$\begin{aligned} \alpha_k(q_k - \widehat{y}(kh)) + \mathcal{O}(\delta^2) &= h^2\beta_{k-1}G(q_{k-1})^{\mathsf{T}}(\lambda_{k-1} - \widehat{\mu}((k-1)h)) \\ &\quad + \mathcal{O}(h^{N+2} + h^{M+1}\delta + \delta^2). \end{aligned}$$

Inserting q_k from this formula into $g(q_k) = 0$ and using $g(\widehat{y}(kh)) = \mathcal{O}(\delta^2)$ yields the estimate for λ_{k-1}, and consequently also for q_k. □

3.7.3 Bounds on parasitic solution components

We next prove that the parasitic solution components $z_i(t)$ remain bounded and small on long time intervals.

Proposition 3.7.4 (Near-invariants). *Under the assumptions of Proposition 3.7.1 there exist h-independent matrix-valued functions $E_{i,l}(y,v)$ such that for every truncation index M and for every solution of the combined system (3.20), (3.29), (3.30) the functions*

$$K_i(y,v,z_i) = \|z_i\|^2 + \overline{z}_i^{\mathsf{T}}\Big(h^2E_{i,2}(y,v) + \ldots + h^{M-1}E_{i,M-1}(y,v)\Big)z_i$$

for $i \in \mathcal{I}_\rho$ and

$$\begin{aligned} K_i(y,v,\nu_i) &= \|h^2G(y)^{\mathsf{T}}\nu_i\|^2 \\ &\quad + h^4\,\overline{\nu}_i^{\mathsf{T}}\Big(hE_{i,1}(y,v) + \ldots + h^{M-1}E_{i,M-1}(y,v)\Big)\nu_i \end{aligned}$$

for $i \in \mathcal{I}_\sigma$ are near-invariants of the system; more precisely, we have

$$\begin{aligned} K_i\big(y(t),\dot{y}(t),z_i(t)\big) &= K_i\big(y(0),\dot{y}(0),z_i(0)\big) + \mathcal{O}(th^M\delta^2), \quad i \in \mathcal{I}_\rho \\ K_i\big(y(t),\dot{y}(t),\nu_i(t)\big) &= K_i\big(y(0),\dot{y}(0),\nu_i(0)\big) + \mathcal{O}(th^M\delta^2), \quad i \in \mathcal{I}_\sigma \end{aligned}$$

as long as $(y(t),\dot{y}(t))$ stays in a compact set and $\|z_i(t)\| \le \delta$ for $i \in \mathcal{I}_\rho$ and $h^2\|G(y(t))^{\mathsf{T}}\nu_i(t)\| \le \delta$ for $i \in \mathcal{I}_\sigma$.

Proof. We start as in the proof of Proposition 3.7.1. However, instead of dividing by $\rho(\zeta_i e^{hD})$ we divide this time by $\sigma(\zeta_i e^{hD})$. This yields

$$\Big(\frac{\rho}{\sigma}\Big)(\zeta_i e^{hD})\, z_i = h^2\Big(f'(y)\, z_i - G(y)^{\mathsf{T}}\nu_i - \big(G'(y)z_i\big)^{\mathsf{T}}\mu\Big) + \mathcal{O}(h^{M+1}\delta).$$

We multiply this relation with the transposed of $\overline{z}_i = z_{-i}$. The second term on the right-hand side vanishes, because of $G(y)z_{-i} = 0$. The first term on the right-hand side is real, because $f(y) = -\nabla U(y)$ so that $f'(y)$ is a symmetric matrix. This is also the case for the third term.

For the study of the left-hand side we consider the expansion (see [HL04, formula (4.16)])

$$\Big(\frac{\rho}{\sigma}\Big)(\zeta_i e^{ix}) = \sum_{l\geq -1} c_{i,l}\, x^l \qquad \text{with real coefficients} \quad c_{-i,l} = (-1)^l c_{i,l},$$

where $c_{i,-1} = c_{i,0} = 0$ and $c_{i,1} \neq 0$ for $i \in \mathcal{I}_\rho$, and $c_{i,-1} \neq 0$ for $i \in \mathcal{I}_\sigma$. We are thus concerned with the expression

$$\sum_{l\geq -1} c_{i,l}(-ih)^l\, \overline{z}_i^{\mathsf{T}} z_i^{(l)}, \tag{3.38}$$

where for $l = -1$ we define in view of (3.34)

$$z_i^{(-1)} = h^3\Big(\theta_{i,1} + \theta_{i,2}(hD) + \dots\Big)\Big(f'(y)z_i - G(y)^{\mathsf{T}}\nu_i - (G'(y)z_i)^{\mathsf{T}}\mu\Big) + \mathcal{O}(h^{M+1}\delta) \tag{3.39}$$

such that $\dot{z}_i^{(-1)} = z_i$.

For $i \in \mathcal{I}_\rho$, we note that $2\mathrm{Re}\,(\overline{z}_i^{\mathsf{T}}\dot{z}_i) = z_{-i}^{\mathsf{T}}\dot{z}_i + \dot{z}_{-i}^{\mathsf{T}} z_i = \frac{d}{dt}\|z_i\|^2$. For the higher order expressions we have the telescoping sums

$$\begin{aligned}
\mathrm{Re}\,\Big(\overline{z}_i z_i^{(2m+1)}\Big) &= \frac{1}{2}\frac{d}{dt}\Big(\sum_{j=0}^{2m}(-1)^j(\overline{z}_i^{(j)})^{\mathsf{T}} z_i^{(2m-j)}\Big)\\
\mathrm{Im}\,\Big(\overline{z}_i z_i^{(2m)}\Big) &= \frac{1}{2i}\frac{d}{dt}\Big(\sum_{j=0}^{2m-1}(-1)^j(\overline{z}_i^{(j)})^{\mathsf{T}} z_i^{(2m-j-1)}\Big)
\end{aligned}$$

so that the imaginary part of (3.38) is a total derivative of a quadratic function in z_i and its derivatives. Using the system (3.29), first and higher order derivatives of z_i can be expressed as a linear function of z_i with coefficients depending on y and $\dot{y}$. Dividing the first formula of the present proof by $c_{i,1}(-ih)/2$, and then taking the real part gives

$$\frac{d}{dt}K_i\big(y(t), \dot{y}(t), z_i(t)\big) = \mathcal{O}(h^M\delta^2)$$

with a quadratic function in z_i of the desired form.

For $i \in \mathcal{I}_\sigma$, we note that

$$2\,\mathrm{Re}\,(\overline{z}_i^{\mathsf{T}} z_i^{(-1)}) = 2\,\mathrm{Re}\,(\overline{\dot{z}_i^{(-1)}}^{\mathsf{T}} z_i^{(-1)}) = \frac{d}{dt}\|z_i^{(-1)}\|^2.$$

The same argument as above yields a near-invariant that is quadratic in $h^{-1}z_i^{(-1)}$. By formula (3.39) the leading term in $h^{-1}z_i^{(-1)}$ is given by $-h^2\theta_{i,1}G(y)^{\mathsf{T}}\nu_i$ and the higher-order terms can be expressed as linear functions in ν_i. This proves the statement of the proposition. □

Let us collect the assumptions that are required for proving the boundedness of the parasitic solution components.

(A1) The multistep method (3.7) is symmetric and of order r. All roots of $\rho(\zeta)$, with the exception of the double root $\zeta_0 = 1$, are simple. All non-zero roots of $\sigma(\zeta)$ are simple and of modulus one.

(A2) The potential $U(q)$ and the constraint function $g(q)$ of (3.1) are defined and smooth in an open neighbourhood of a compact set K.

(A3) The starting approximations $q_0, \dots, q_{k-1}$ and $\lambda_1, \dots, \lambda_{k-2}$ are such that the initial values for the differential-algebraic system (3.20), (3.29), (3.30) obtained from Proposition 3.7.2 satisfy

$$y(0) \in K, \qquad \|\dot y(0)\| \le M,$$
$$\|z_i(0)\| \le \delta/2, \ i \in \mathcal{I}_\rho \quad \text{and} \quad \|h^2 G(y(0))^{\mathsf T} \nu_i(0)\| \le \delta/2, \ i \in \mathcal{I}_\sigma.$$

(A4) The numerical solution $\{q_n\}$, for $0 \le nh \le T$, stays in a compact set K_0 that has a positive distance to the boundary of K.

Theorem 3.7.5 (Long-time bounds for the parasitic components)**.** *Assume (A1)–(A4). For sufficiently small h and δ and for fixed truncation indices N and M that are large enough such that $h^N = \mathcal{O}(\delta^2)$ and $h^M = \mathcal{O}(\delta)$, there exist functions $y(t), \mu(t)$ and $z_i(t), \nu_i(t)$ for $i \in \mathcal{I}$ on an interval of length*

$$T = \mathcal{O}(h\delta^{-1})$$

such that

- $q_n = y(nh) + \sum_{i \in \mathcal{I}} \zeta_i^n z_i(nh)$ *for* $0 \le nh \le T$;
- $\lambda_n = \mu(nh) + \sum_{i \in \mathcal{I}} \zeta_i^n \nu_i(nh)$ *for* $0 \le nh \le T$;
- *on every subinterval $[nh, (n+1)h)$ the functions $y(t), \mu(t)$ and $z_i(t), \nu_i(t)$ for $i \in \mathcal{I}$ are a solution of the system (3.20), (3.29), (3.30);*
- *the functions $y(t), h^2\mu(t)$ and $z_i(t), h^2\nu_i(t)$ for $i \in \mathcal{I}$ have jump discontinuities of size $\mathcal{O}(\delta^2)$ at the grid points nh;*
- *for $0 \le t \le T$, the parasitic components are bounded by*

$$\|z_i(t)\| \le \delta, \ i \in \mathcal{I}_\rho \quad \text{and} \quad \|h^2 G(y(t))^{\mathsf T} \nu_i(t)\| \le \delta, \ i \in \mathcal{I}_\sigma.$$

Proof. To define the functions $y(t), \mu(t), z_i(t), \nu_i(t)$ on the interval $[nh, (n+1)h)$ we consider the consecutive numerical solution values $q_n, q_{n+1}, \dots, q_{n+k-1}$ and $\lambda_{n+1}, \dots, \lambda_{n+k-2}$. We compute initial values for the system (3.20), (3.29), (3.30) according to Proposition 3.7.2, and we let $y(t), \mu(t), z_i(t), \nu_i(t)$ be its solution on $[nh, (n+1)h)$. By Proposition 3.7.3 this construction yields jump discontinuities of size $\mathcal{O}(\delta^2)$ at the grid points.

It follows from Proposition 3.7.4 that $K_i(y(t), \dot y(t), z_i(t))$ for $i \in \mathcal{I}_\rho$ and $K_i(y(t), \dot y(t), \nu_i(t))$ for $i \in \mathcal{I}_\sigma$ remain constant up to an error of size $\mathcal{O}(h^{M+1}\delta^2)$ on the interval $[nh, (n+1)h)$. Taking into account the jump discontinuities of size $\mathcal{O}(\delta^2)$, we find that

$$\begin{aligned} K_i(y(t), \dot y(t), z_i(t)) &\le K_i(y(0), \dot y(0), z_i(0)) + C_1 t h^{-1}\delta^3 + C_2 t h^M \delta^2 \\ K_i(y(t), \dot y(t), \nu_i(t)) &\le K_i(y(0), \dot y(0), \nu_i(0)) + C_1 t h^{-1}\delta^3 + C_2 t h^M \delta^2 \end{aligned}$$

as long as $\|z_i(t)\| \le \delta$ for $i \in \mathcal{I}_\rho$ and $\|h^2 G(y(t))^\mathsf{T} \nu_i(t)\| \le \delta$ for $i \in \mathcal{I}_\sigma$. By Proposition 3.7.4 this then implies with $C_3 = C_1 + hC_2$, for $i \in \mathcal{I}_\rho$,

$$\|z_i(t)\|^2 \le \|z_i(0)\|^2 + C_3 t h^{-1} \delta^3 + C_4 h^2 \delta^2.$$

For $i \in \mathcal{I}_\sigma$ we obtain

$$\|h^2 G(y(t))^\mathsf{T} \nu_i(t)\|^2 \le \|h^2 G(y(0))^\mathsf{T} \nu_i(0)\|^2 + C_3 t h^{-1} \delta^3 + C_4 h \delta^2.$$

The assumptions $\|z_i(t)\| \le \delta$ and $\|h^2 G(y(t))^\mathsf{T} \nu_i(t)\| \le \delta$ are certainly satisfied as long as $C_3 t\delta \le h/4$ and $C_4 h \le 1/4$, so that the right-hand side of the above estimates is bounded by δ^2. This proves not only the estimate for $\|z_i(t)\|$ and $\|h^2 G(y(t))^\mathsf{T} \nu_i(t)\|$, but at the same time it guarantees recursively that the above construction of the functions $y(t), \mu(t), z_i(t), \nu_i(t)$ is feasible. □

3.7.4 Proof of the main results

The proof of Theorem 3.3.1 combines Theorem 3.7.5 and Proposition 3.6.2. For the piecewise smooth function $y(t)$ of Theorem 3.7.5 we have

$$H_h(y(t), \dot y(t)) = H_h(y(0), \dot y(0)) + \mathcal{O}(t h^N) + \mathcal{O}(t h^{-1} \delta^2),$$

where the first error term comes from the truncation of the modified energy and the second error term comes from the discontinuity at the grid points. By the bounds for the parasitic components z_i we have

$$q_n = y(nh) + \mathcal{O}(\delta) \quad \text{and} \quad p_n = \dot y(nh) + \mathcal{O}(h^{-1}\delta + h^r)$$

because the differentiation formula is of order r. We therefore obtain

$$H_h(q_n, p_n) = H_h(q_0, p_0) + \mathcal{O}(t h^N) + \mathcal{O}(t h^{-1} \delta^2) + \mathcal{O}(h^{-1}\delta + h^r).$$

With $\delta = h^{r+2}$, Theorem 3.3.1 now follows by using the estimate between the modified energy H_h and the original energy H as given by Proposition 3.6.2.

Theorem 3.3.2 is obtained in the same way using Proposition 3.6.3.

Chapter 4

Complements on symmetric LMM for constrained Hamiltonian systems

4.1 Introduction

This chapter presents some complements to the study of symmetric linear multistep methods applied to constrained Hamiltonian systems.
In Section 4.2 the interval of periodicity of the classes of methods of Section 3.4 is studied; this study is supplemented by some figures representing the stability regions of the classes of methods of order 4, 6 and 8.
In Section 4.3 a similar study is done for the error constant, which has been computed for the same classes of methods and represented graphically.

4.2 Stability issues

In the previous chapter we studied the properties of near preservation of energy and momenta of symmetric multistep methods applied to constrained Hamiltonian systems, and some stability properties of these methods. In this paragraph we will analyze in detail these properties for the classes of methods described in Section 3.4.

4.2.1 Overview on stability for linear multistep methods

We consider a linear multistep method for second order equations $\ddot{q} = f(q)$,

$$\sum_{j=0}^{k} \alpha_j q_{n+j} = h^2 \sum_{j=0}^{k} \beta_j f(q_{n+j}) \tag{4.1}$$

with generating polynomials $\rho(\zeta)$ and $\sigma(\zeta)$; as remarked in Section 3.4.2, we use the equation of the harmonic oscillator $\ddot{q} = -\omega^2 q$ to study the stability of the method.
Applying (4.1) to the test equation, we obtain the following difference equation

$$\sum_{j=0}^{k} \alpha_j q_{n+j} = -h^2\omega^2 \sum_{j=0}^{k} \beta_j q_{n+j}.$$

whose corresponding characteristic polynomial is

$$\rho(\zeta) - z^2\sigma(\zeta)$$

where we denote $z = \mathrm{i}h\omega$.

Definition 4.2.1 (Interval of Periodicity)**.** We define *interval of periodicity* of the multistep method with generating polynomials $\rho(\zeta)$, $\sigma(\zeta)$ the interval

$$[0,\Omega] = \left\{ H \geq 0;\ \text{all the roots of}\ \ \rho(\zeta) + H^2\sigma(\zeta) = 0\ \ \text{have modulus one} \right\} \tag{4.2}$$

where $H = \mathrm{i}z$. As remarked in Section 3.4.2, if the polynomial $\rho(\zeta)$ of a zero-stable symmetric method has all simple roots, then a continuity argument shows that $[0,\Omega] \neq \emptyset$.

4.2.2 Study of stability

In this paragraph we want to study the stability of the classes of symmetric multistep methods of order $4, 6$ and 8; thus we consider a polynomial $\rho(\zeta)$ of the form

$$\rho(\zeta) = (\zeta - 1)^2 \prod_{j=1}^{k/2-1} \left(\zeta^2 + 2a_j\zeta + 1\right) :$$

and after having checked if the corresponding polynomial $\sigma(\zeta)$ satisfies the root condition, we study the interval of periodicity (4.2) as a function of the parameters a_j.
In this paragraph we will assume $\omega = 1$ and, as in Section 3.4.2, we consider separately the different cases $k = 4, 6, 8$.

- $k = 4$: We have only one parameter a_1 and we know from Section 3.4 that every value of $|a_1| < 1$ gives rise to a polynomial $\sigma(\zeta)$ that satisfies the root condition.
  As in Section 3.4.2, it is possible to study when two roots collapse at the point -1, obtaining the analytic expression for the interval of periodicity

  $$[0,\Omega] = \left[0, \sqrt{\frac{6(1-a_1)}{2-a_1}}\right]. \tag{4.3}$$

  Another way to approach the problem is the numerical study of the roots of $\rho(\zeta) + H^2\sigma(\zeta)$ as a function of the parameter $|a_1| < 1$; we implemented the problem in Matlab, using for the parameter a_1 a grid of equidistant points.
  The curve shown in Figure 4.1 represents the length of the interval of periodicity as a function of a_1, and it corresponds to the analytic expression reported in (4.3).

- $k = 6$: In this case, we have two parameters a_1, a_2, and we know from Section 3.4 that some values of these parameters do not make $\sigma(\zeta)$ satisfy the root condition.
  As before we implemented the problem in Matlab, using a equidistant grid for the parameters a_1 and a_2. We excluded part of the border of the square because of round-off; although we expect that the length of the interval of periodicity tends to zero for values of the parameters closer and closer to the boundary of the square.
  Figure 4.2 shows the square representing the two parameters a_1 and a_2: the different colors represent the different sizes of the region of stability.
  As reported in Section 3.4.2 the maximal size of the interval of periodicity is obtained with a_1 and a_2 respectively close to 0.66 and 0.26; for these values, we have $[0,\Omega] \approx [0, 1.05]$.

- $k = 8$: In this case we have three parameters a_1, a_2, a_3: from Section 3.4 we know that the polynomial $\sigma(\zeta)$ satisfies the roots conditions only for some values of these parameters.

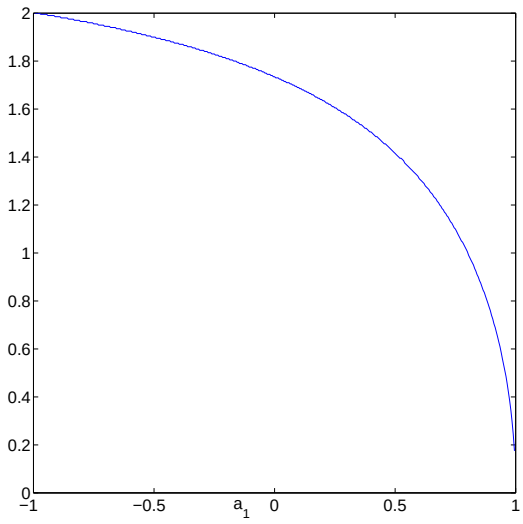


Figure 4.1 – Length of the interval of periodicity for the class of methods of order 4.

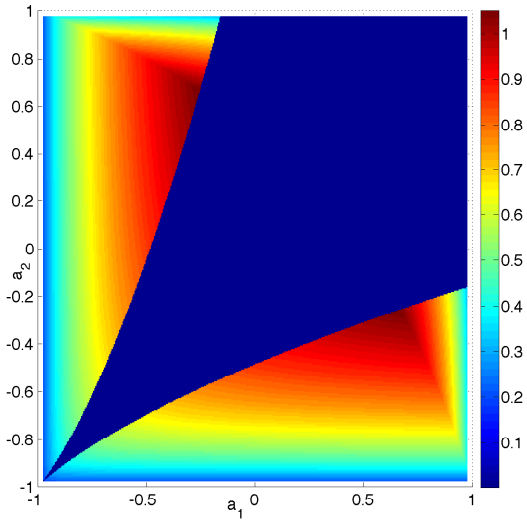


Figure 4.2 – Length of the interval of periodicity for the class of methods of order 6. The value zero corresponds to the parameters a_1, a_2 for which $\sigma(\zeta)$ doesn't satisfy the root condition.

To study the stability we fixed 12 equidistants values of a_3, and on these "slices" we fixed a discretization for a_1, a_2 and $H = \omega h$: the results of these computations are shown in Figure 4.3.

Using a refined grid in a smaller region, we found that the values of the parameters that maximize the interval of periodicity are close to $a_1 = -0.305$, $a_2 = 0.585$ and $a_3 = -0.8975$: the corresponding interval of periodicity is $[0, \Omega] \approx [0, 1.0075]$.

4.3 Study of the error constant

In this paragraph we want to study the error constant for the classes of methods of order 4, 6 and 8 described in Section 3.4. We know that the error constant for a linear multistep method of order p is defined by

$$C = \frac{C_{p+2}}{\sigma(1)}, \tag{4.4}$$

where C_{p+2} is given by

$$\rho(e^h) - h^2\sigma(e^h) = C_{p+2}h^{p+2} + \mathcal{O}(h^{p+3}).$$

We studied this quantity numerically as a function of the parameters a_i. We distinguish the cases $k = 4, 6, 8$.

- *k=4*: as reported in the previous paragraph, we have only one parameter $|a_1| < 1$, and the error constant is given by

$$C = -\frac{1}{240}\frac{-9 + a_1}{1 + a_1}.$$

Figure 4.4 shows C as a function of a_1: we notice that, because of the term $1 + a_1$ in the denominator, is desirable to choose a_1 far from -1 in order to have a small error constant.

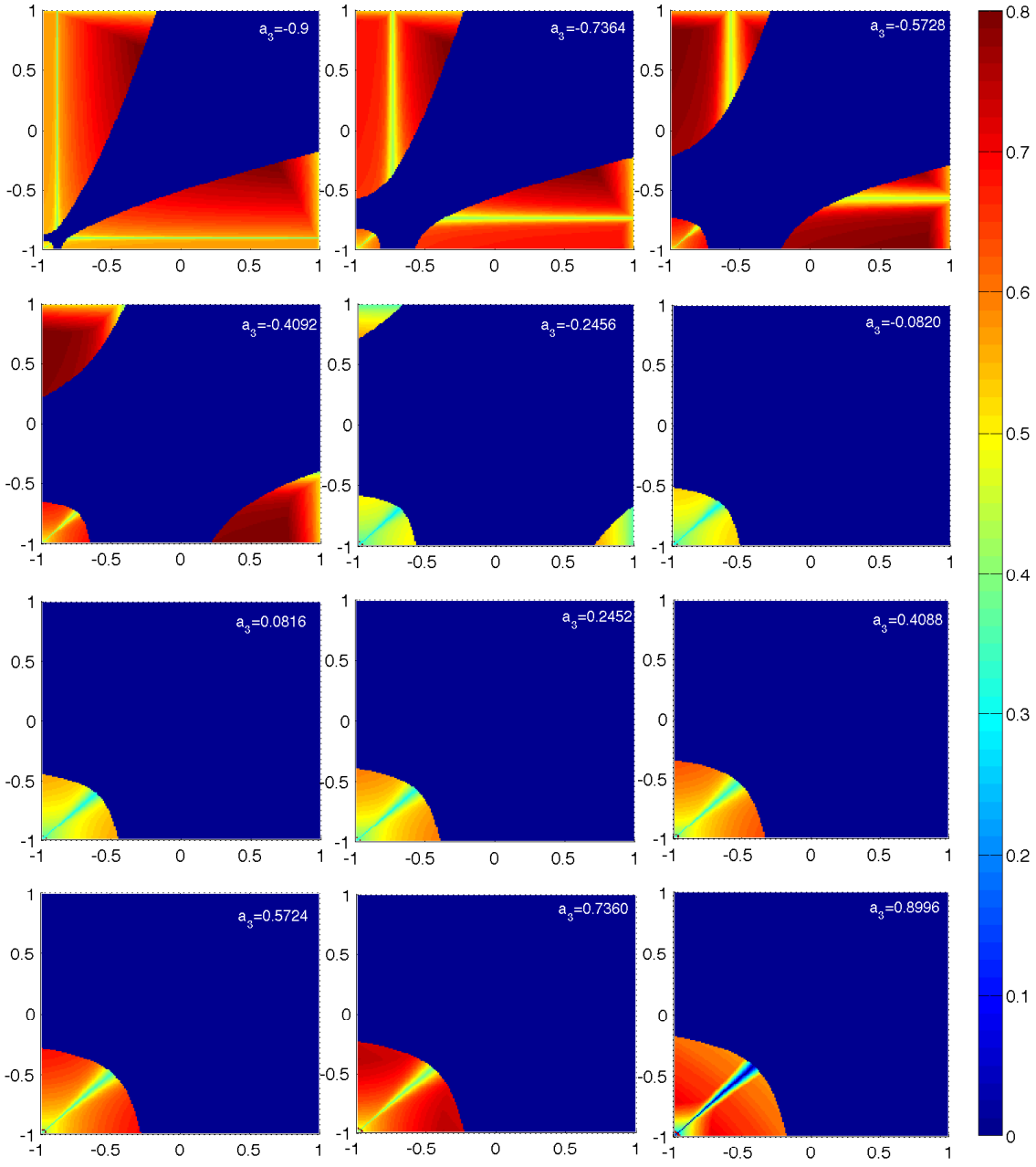


Figure 4.3 – Length of the interval of periodicity for the class of methods of order 8. The value zero corresponds to the parameters a_1, a_2 for which $\sigma(\zeta)$ doesn't satisfy the root condition: the horizontal and vertivcal axis represent respectively a_1 and a_2.

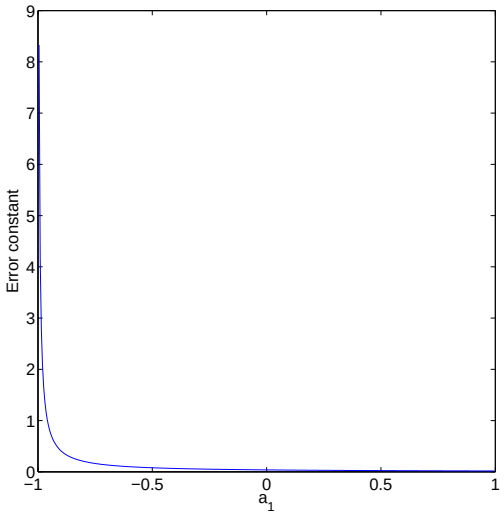


Figure 4.4 – Error constant for the class of methods of order 4 as a function of a_1.

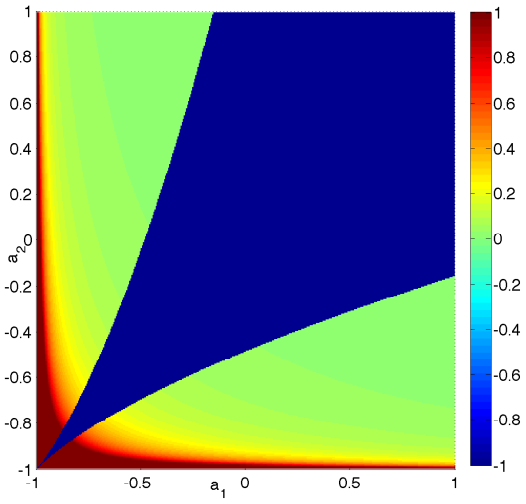


Figure 4.5 – Error constant for the class of methods of order 6 as a function of a_1 and a_2. The value $C = -1$ corresponds to the parameters a_1, a_2 that give rise to $\sigma(\zeta)$ that doesn't satisfy the root condition.

- *k=6*: in this case we have two parameters a_1 and a_2 which have some constraints given by the root condition on the polynomial $\sigma(\zeta)$.
  The error constant is given by

$$C = \frac{1}{60480}\frac{-95(a_1+a_2)+31a_1a_2+1039}{1+a_2+a_1+a_1a_2}.$$

  that is shown in Figure 4.5 as function of a_1 and a_2: we reported all the $C > 1$ as equal to 1.
  We notice that it is preferable to choose both the parameters far from -1 to avoid having large values of C.

- *k=8*: in this case we have three parameters a_1, a_2 and a_3, which have some constraints due to the roots condition on $\sigma(\zeta)$.
  The error constant is given by

$$C = -\frac{1}{3628800}\frac{2209(a_1+a_2+a_3)+289a_1a_2a_3-641(a_1a_2+a_2a_3+a_1a_3)-28961}{1+a_1+a_2+a_3+a_1a_2a_3+a_1a_2+a_2a_3+a_1a_3}$$

  Figure 4.6 shows C as a function of a_1, a_2 and a_3 and again we report all the $C > 1$ as equal to 1.
  As in the previous cases we see that if one of the three parameters is too close to -1 the error constant increases.

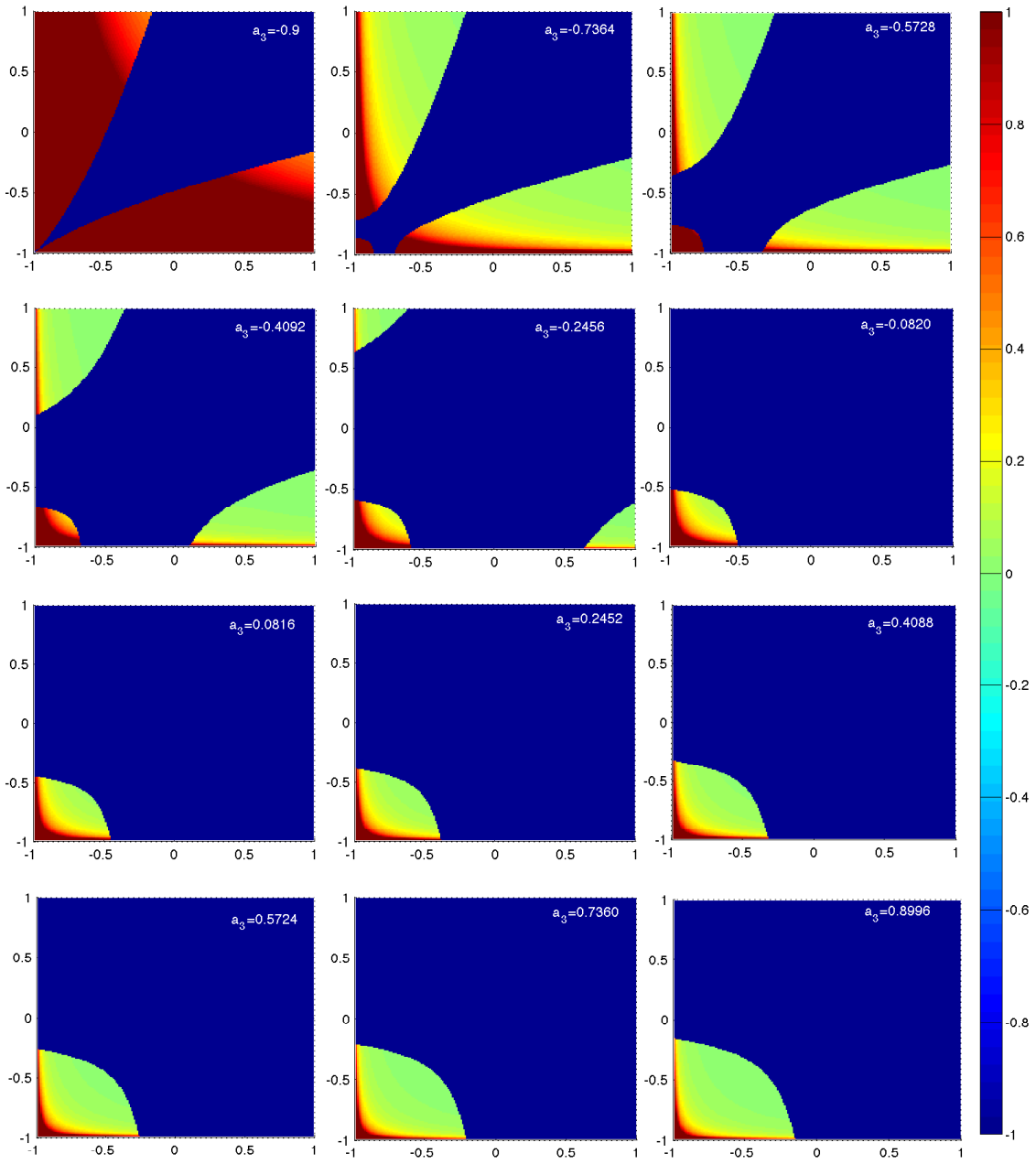


Figure 4.6 – Error constant for the class of methods of order 8. The value $C = -1$ corresponds to the parameters a_1, a_2 that give rise to $\sigma(\zeta)$ that doesn't satisfy the root condition.

Chapter 5

Implementation and round-off error optimization

Note: Part of this Chapter is a different presentation of the results of the article [CH13b] in collaboration with E. Hairer.

The results of this Chapter are obtained by the Fortran 90 code of Appendix A, whereas those of [CH13b] are obtained by a simplified implementation in Fortran 77.

5.1 Introduction

This chapter is focused on the implementation in Fortran 90 of the class of methods shown in Chapter 3, and on the issues related to the optimization of round-off error.
In the previous chapters we described how to build a class of methods with high order for the solution of constrained Hamiltonian systems: since with very accurate methods it is easy to achieve errors of the size of the machine precision it also becomes necessary to optimize round-off errors.

In Section 5.2 it is shown how a simple implementation can lead to a linear growth of round-off error when the errors reach the size of 10^{-16}; then we describe all the techniques which are useful to solve this problem and to achieve an optimized round-off error that behaves like a random walk. All the analyses are supplemented with numerical experiments showing the improvements produced by each described technique.

In Section 5.3 further numerical experiments are reported, and we show the behaviour of the algorithm when it is applied to chaotic systems (the final version of this routine can be found in Appendix A).

Finally, in Section 5.4, the probabilistic interpretation of the results of this chapter is given.

5.2 Round-off error: comparing standard and optimized implementations

In this section we want to compare different implementations of the algorithm presented in Chapter 3, remarking how they can lead to different behaviours of the error when the discretization error reaches the size of the machine precision.
We show here the techniques used to eliminate all the deterministic errors that can lead

to a linear growth of the error: this is something we wish to avoid since we want to have a very accurate algorithm.

The comparison in this section are made using as a test the two-body problem on the sphere with initial data described in 3.5.2: we chose a very regular system so that the typical long-time behaviour can be better observed. Other numerical experiments will be provided in the next section of this chapter. All the stepsizes of the computations will be chosen in order to obtain a discretization error small enough to make the round-off error visible.

5.2.1 "SHAKE-like" vs "RATTLE-like" implementations

As described in Section 3.2, the standard or "SHAKE-like" (from the algorithm described in [RCB77]) formulation of the algorithm we described is

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j q_{n+j} &= h^2 \sum_{j=0}^{k} \beta_j \left(f(q_{n+j}) - G(q_{n+j})^{\mathsf{T}} \lambda_{n+j} \right) \\ 0 &= g(q_{n+k}) \end{aligned}$$

where $\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j$ and $\sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j$ are the characteristic polynomials associated to a symmetric multistep method of order r: in this section, as in Chapter 3, we will focus only on explicit methods (i. e. $\sigma(\zeta)$ will have degree $k-1$). The approximations of the momenta are computed a-posteriori using a central finite differences formula (in general of the same order as the multistep method to obtain an algorithm of order r), followed by a projection on the constraint $G(q_n)M^{-1}p_n = 0$, that is

$$\begin{aligned} p_n &= M \frac{1}{h} \sum_{j=-k/2}^{k/2} \delta_j q_{n+j} + h G(q_n)^{\mathsf{T}} \mu_n \\ 0 &= G(q_n) M^{-1} p_n. \end{aligned}$$

As remarked in Section 3.2, this is not the best formulation for this algorithm, because for $h \to 0$ the double root in $\zeta = 1$ of the polynomial $\rho(\zeta)$ leads to unbounded solution of the difference equation associated to the method.

To avoid this we use a *stabilized* (or "RATTLE-like", from [And83]) formulation: we split the double root in $\zeta = 1$ considering the new variable $p_{n+j+1/2}$ and, denoting by $(\hat{\alpha}_j)_{j=0}^{k-1}$ the coefficients of $\hat{\rho}(\zeta) = \rho(\zeta)/(\zeta - 1)$, the algorithm becomes

$$\begin{aligned} \sum_{j=0}^{k-1} \hat{\alpha}_j p_{n+j+1/2} &= h \sum_{j=1}^{k-1} \beta_j \left(f(q_{n+j}) - G(q_{n+j})^{\mathsf{T}} \lambda_{n+j} \right) \\ q_{n+k} &= q_{n+k-1} + h M^{-1} p_{n+k-1/2} \\ 0 &= g(q_{n+k}). \end{aligned} \tag{5.1}$$

This algorithm consists of three principal parts: the solution of the nonlinear system to obtain the Lagrange multipliers, the computation of the momenta on the intermediate grid and the computation of the positions. We observe that this formulation is less affected by round-off because both of the difference equations associated to (5.1) have bounded solutions for $h \to 0$. We can also use the approximations of the momenta on the intermediate

grid to reformulate the computation of p_n, which can be rewritten as

$$\begin{aligned} p_n &= \sum_{j=-k/2}^{k/2-1} \hat{\delta}_j p_{n+j+1/2} + hG(q_n)^{\mathsf{T}}\mu_n \\ 0 &= G(q_n)M^{-1}p_n, \end{aligned}$$

where the coefficients $\hat{\delta}_j$ are given by $(\zeta-1)\sum_{j=-l}^{l-1}\hat{\delta}_j\zeta^j = \sum_{j=-l}^{l}\delta_j\zeta^j$: anyway this is less important because the momenta p_n don't enter in the recurrency.

In Figure 5.1 we show the comparison of the error in the Hamiltonian obtained using both formulations described in this section of an algorithm of order 8, applied on the two body problem on the sphere described in Section 3.5.2: the error increases like $\sqrt{t}$ and there is no linear growth (it will be explained in Section 5.4). The figure confirms that the "RATTLE-like" formulation (5.1) leads to an error that is smaller than the one obtained with the "SHAKE-like" formulation.

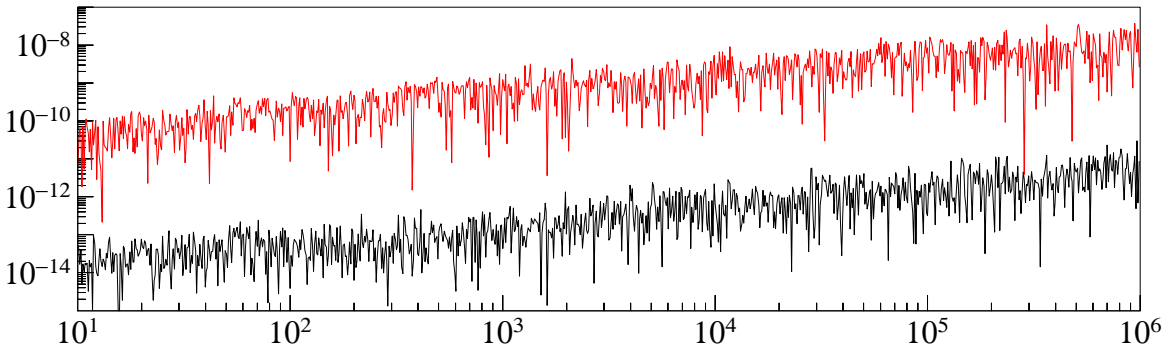


Figure 5.1 – (Two-Body problem on the sphere). Comparison of "SHAKE-like" (red) and "RATTLE-like" (black) formulations: there is reported the error in the Hamiltonian of an 8th order algorithm as function of time, $h = 0.001$.

To explain the difference between the errors found with the two algorithms, Figure 5.2 shows the errors obtained with the "SHAKE-like" and "RATTLE-like" routines with different stepsizes. We observe that with big stepsizes we obtain the same errors with the two algorithms (this is expected since they are two different formulations of the same algorithm) but we notice that, decreasing the stepsize, the error of the "SHAKE-like" decreases only until the size of about 10^{-12}, after which it starts increasing because of round-off. On the other hand, the error obtained with the stabilized formulations decreases until the size of 10^{-14} and afterwards starts increasing because of round-off. This happens because the round-off error due to the "SHAKE-like" formulation is bigger than the round-off error obtained with the "RATTLE-like" formulation: this confirms the better performance of the formulation (5.1).

In the following subsections we will work on (5.1) eliminating all the possible deterministic errors.

5.2.2 Influence of the initial approximations

To use the method (5.1), first of all we need to compute the initial approximations $p_{1/2}, \ldots, p_{k-1/2}$, $q_1, \ldots, q_{k-1}$ from the initial data p_0, q_0: we can do this by applying a one-step method of the form

$$y_j = y_{j-1} + h\delta_j, \quad y_j = (p_j, q_j)$$

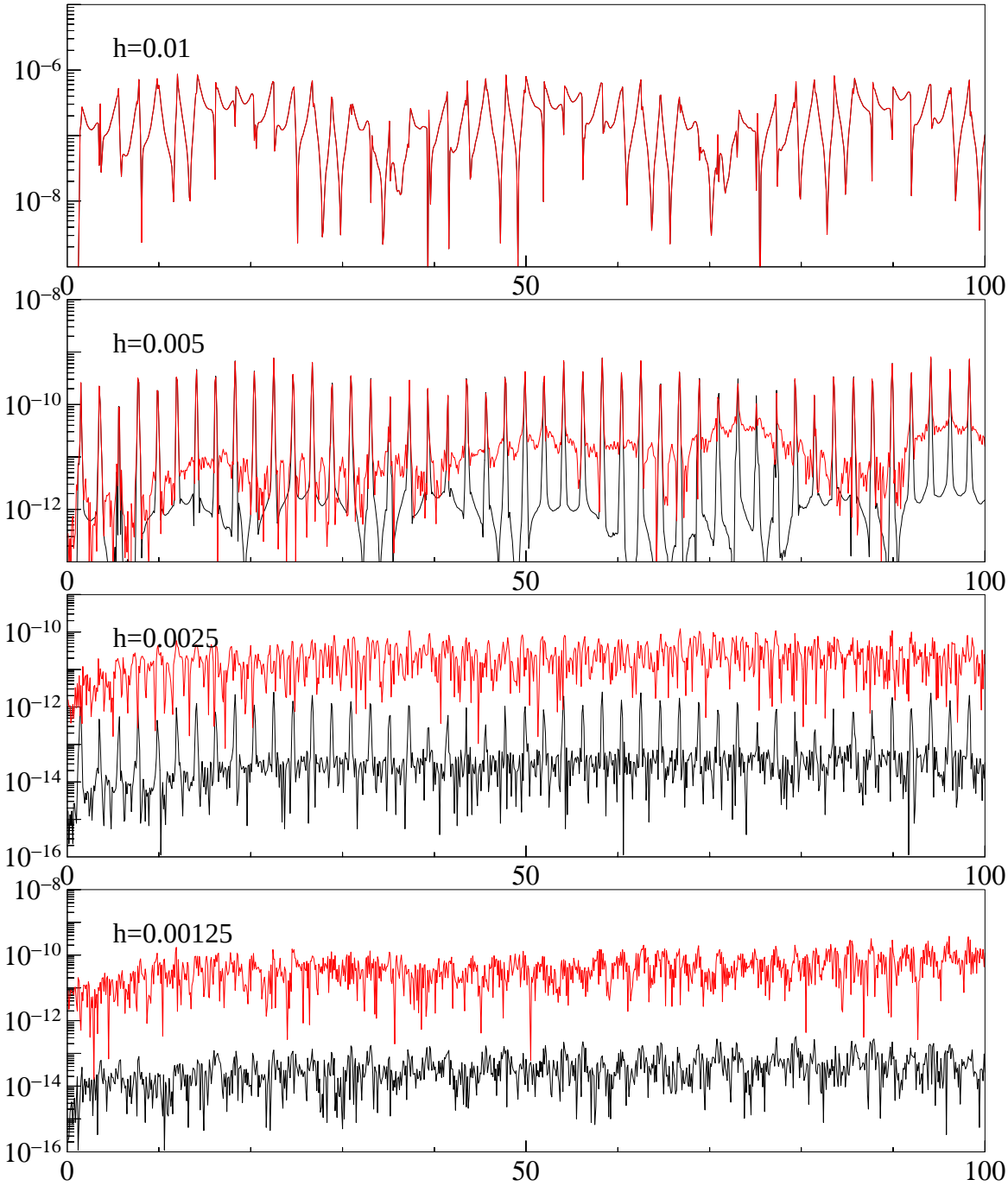


Figure 5.2 – (Two-body problem on the sphere). Comparison of "SHAKE-like" (red) and "RATTLE-like" (black) formulations: the figure shows the error in the Hamiltonian of an 8th order algorithm as a function of time using different stepsizes.

to the original first-order Hamiltonian system

$$\begin{aligned} \dot{p} &= -\nabla U(q) - G(q)^{\mathsf{T}}\lambda \\ \dot{q} &= M^{-1}p. \end{aligned}$$

To avoid cancellation errors we do not compute the initial approximations of the momenta on the intermediate grid as $p_{j-1/2} = (q_j - q_{j-1})/h$, but with the relation $p_{j-1/2} = M\delta_j$. We used different one-step methods to compute the initial approximations to check their influence in the accuracy of the algorithm.

In Figure 5.3 we compare the error obtained from the algorithm (5.1) using initial approximations obtained from an explicit Runge Kutta method of order 4 (ERK4) and an implicit Runge Kutta method of order 8 (IRK8): for the computation of the initial approximations with ERK4 we divided the stepsize by an appropriate integer number in order to obtain the same discretization error obtained with IRK8 and thus to correctly compare the numerical solutions. We observe that the choice of method for the computation of initial approximations doesn't influence the numerical solution, which depends only on the discretization error.

We decided to use the implicit Gauss Runge Kutta method of order 8 for the computation of the initial approximation in the Fortran implementation reported in Appendix A, and all the numerical experiments reported in this chapter have been also made using this method.

The initial Lagrange multipliers $\lambda_0, \ldots, \lambda_{k-2}$ are computed using *index reduction*: we solve with respect to λ the linear system given by the second time derivative of the constraint, i. e.

$$0 = \frac{\partial}{\partial q}\left(G(q)H_p(p,q)\right)H_p(p,q) - G(q)H_{pp}(p,q)\left(H_q(p,q) + G(q)^{\mathsf{T}}\lambda\right). \tag{5.2}$$

We observe that all the terms in (5.2) can be easily calculated by the functions we need as input for the algorithm, except for $\frac{\partial}{\partial q}(G(q)H_p(p,q))$; so in the final implementation of the routine the function $\frac{\partial}{\partial q}(G(q)H_p(p,q))H_p(p,q)$ is an input function that has to be provided by the user.

5.2.3 Treatment of the coefficients of a multistep method

In this section we compare the errors obtained after manipulating the coefficients of the multistep formula in (5.1); in particular we want to compare the error found using integer coefficients to that obtained with the use of rational coefficients.

We know from Section 3.4 that

$$\hat{\rho}(\zeta) = \sum_{j=0}^{k-1}\hat{\alpha}_j\zeta^j = (\zeta-1)\prod_{j=1}^{k/2-1}\left(\zeta^2 + 2a_j\zeta + 1\right) \qquad \text{with } |a_j| < 1,$$

and so to make the $\hat{\alpha}_j$ integer it is sufficient to consider the least common denominator of $(a_j)_{j=1}^{k/2-1}$, which we will call d_α.
The coefficients of $\sigma(\zeta)$ are in general rational, and so calling d_β their least common denominator and defining *den* the least common denominator of $\{d_\alpha, d_\beta\}$, we can rewrite

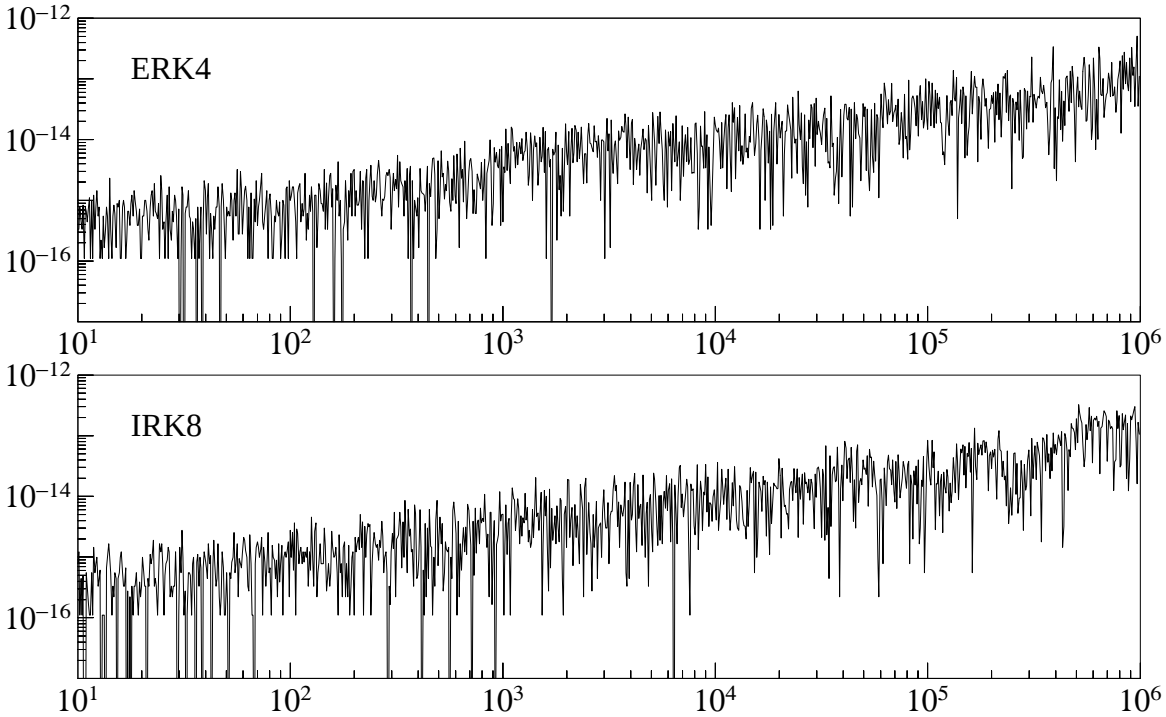


Figure 5.3 – (Two-body problem on the sphere). Comparison of the error in the Hamiltonian of an 8th order algorithm as a function of time, $h = 0.001$, using different initial approximations computed with an explicit Runge Kutta method of order 4 and a Gauss Implicit Runge Kutta method of order 8.

the algorithm as

$$\begin{aligned} p_{n+k-1/2} &= p_{n+1/2} + \frac{1}{den}\left(\sum_{j=1}^{k-2}\hat{\alpha}'_j p_{n+j+1/2} - h\sum_{j=1}^{k-1}\beta'_j\left(f(q_{n+j}) - G(q_{n+j})^{\mathsf{T}}\lambda_{n+j}\right)\right) \\ q_{n+k} &= q_{n+k-1} + hM^{-1}p_{n+k-1/2} \\ 0 &= g(q_{n+k}), \end{aligned} \tag{5.3}$$

where $\hat{\alpha}'_j = \hat{\alpha}_j \cdot den$ and $\beta'_j = \beta_j \cdot den$; these coefficients are all integers by construction.

Another possibility is keeping a_j in its decimal representation but multiplying $\sigma(\zeta)$ by common denominator d_β of the coefficients $(\beta_j)_{j=1}^{k-1}$: in this way we obtain an algorithm of the form

$$\begin{aligned} p_{n+k-1/2} &= \sum_{j=0}^{k-2}\hat{\alpha}_j p_{n+j+1/2} - \frac{h}{d_{\beta_j}}\sum_{j=1}^{k-1}\beta''_j\left(f(q_{n+j}) - G(q_{n+j})^{\mathsf{T}}\lambda_{n+j}\right) \\ q_{n+k} &= q_{n+k-1} + hM^{-1}p_{n+k-1/2} \\ 0 &= g(q_{n+k}), \end{aligned} \tag{5.4}$$

where $\beta''_j = \beta_j \cdot d_\beta$.

In Figure 5.4 we compare the error obtained from the basic algorithm (5.1) (with a_j decimal in quadruple precision, and $\rho(\zeta)$ and $\sigma(\zeta)$ with rational coefficients) with (5.3) and (5.4) with a_j in decimal representation in both quadruple and double precision. We observe that the manipulation of the coefficients does not give rise to remarkable differences among the errors in the Hamiltonian.

We therefore use formulation (5.3), which appears to be the most stable of the three.

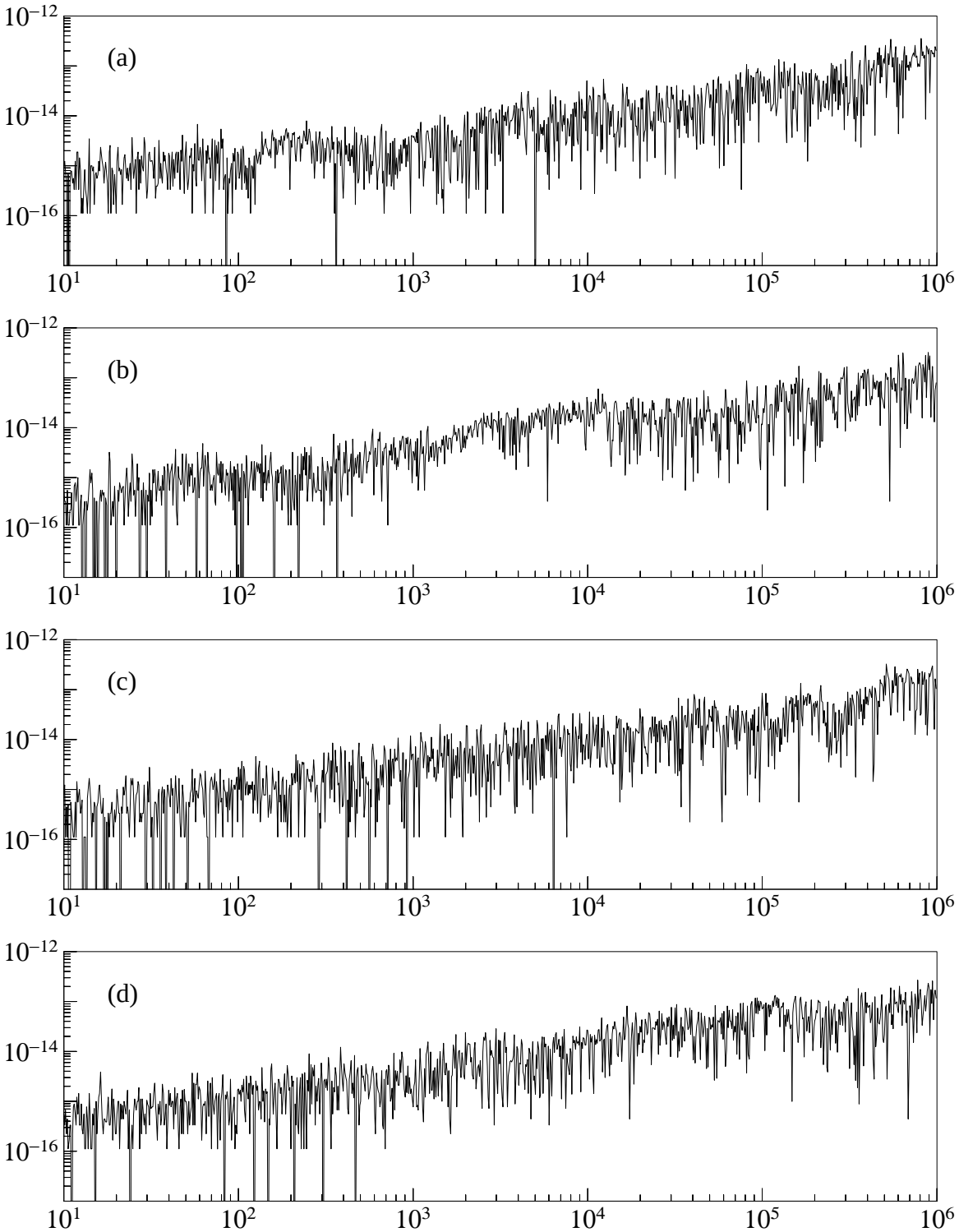


Figure 5.4 – (Two-body problem on the sphere). Comparison of the error in the Hamiltonian of an 8th order algorithm as a function of time, $h = 0.001$, using different implementations of the coefficients. In (a) there is reported the error obtained with the method (5.4) with a_j in double precision; in (b) the error of the method (5.4) with a_j in quadruple precision; in (c) the error of the method (5.3) with all integer coefficients and in (d) the error of the method (5.1) with α_j,β_j and a_j rational. All of these routines have been implemented using all the techniques described in this chapter.

5.2.4 Compensated summations

It is well known that the computations of the form

$$y_{n+1} = y_n + \delta_n, \tag{5.5}$$

where δ_n has smaller size than y_n, can be treated using *compensated summations*, a technique due to Gill (1951), Kahan (1965) and Möller (1965) that simulates calculations in quadruple precision using computations in double precision: in this way we can get more precision with cpu time comparable to an implementation in double precision.

The algorithm for the recursion (5.5) is the following:

$e = 0$
for $n = 0, 1, 2, \ldots$ do
  $a = y_n$
  $e = e + \delta_n$
  $y_{n+1} = a + e$
  $e = e + (a - y_{n+1})$
end do

We observe that the computation of the position is exactly of the form (5.5), and so compensated summations can be applied in order to reduce round-off error.

Similarly, the momenta are computed with a sum of quantities of different sizes, and so a similar technique is needed to reduce round-off error due to this sum, even if (5.5) cannot be applied straightforwardly.

In fact, the general term of the recursion for p on the intermediate grid is given by (here for $k = 4$)

$$p_{n+7/2} = p_{n+1/2} + \frac{1}{\delta}\underline{\left(\hat{\alpha_1}\left(p_{n+3/2} - p_{n+5/2}\right) + h(\beta_1\left(f_{n+1} + f_{n+3/2}\right) + \beta_2 f_{n+2})\right)}.$$

We observe that the underlined term has a smaller size than $p_{n+1/2}$, since it consists of a sum of an $\mathcal{O}(h)$ and $\hat{\alpha_1}\left(p_{n+3/2} - p_{n+5/2}\right)$, which is small because it is the difference of two close values of p. Furthermore we observe that the 4-step method can be rewritten as the combination of three recurrences such as

$$\begin{aligned} u_{k+1} &= u_k + \delta_k \\ v_{k+1} &= v_k + \gamma_k \\ z_{k+1} &= z_k + \eta_k \end{aligned}$$

where

$$\begin{aligned} u_k &= p_{n+3k+1/2} \\ v_k &= p_{n+3k+3/2} \\ z_k &= p_{n+3k+5/2} \end{aligned}$$

and we can then apply compensated summations to each of these: we need thus $k - 1$ vectors to collect small errors for a k-step multistep method.

As described in [CH13b] we create k vectors $e_{j+1/2}$, $j = 0, \ldots, k-1$ to save small errors of the momenta $p_{j+1/2}$: we can now apply the recurrency formula on them and use the result to update the output of the current iteration. Furthermore we create a vector e to save the small error in the computation of the positions and we apply straightforwardly the standard technique.

The compensated summations for both positions and momenta of (5.3) can be written as follows

$$
\begin{aligned}
&e = 0, e_j = 0,\ j = 0, \dots, k-1\\
&\text{for } n = 0, 1, 2, \dots \text{ do}\\
&\quad s_1 = h\left(\textstyle\sum_{j=1}^{k/2-1} \beta_j (f_{n+k-j} + f_{n+j}) + \beta_{k/2} f_{n+k/2}\right)\\
&\quad s_2 = -\textstyle\sum_{j=1}^{k/2-1} \hat{\alpha}_j \left(p_{n+k-j-1/2} - p_{n+j+1/2}\right)\\
&\quad d = -\textstyle\sum_{j=1}^{k/2-1} \hat{\alpha}_j \left(e_{k-j-1/2} - e_{j+1/2}\right)\\
&\quad a = p_{n+1/2}\\
&\quad e_{k-1/2} = (s_1 + s_2)/den + (d/den + e_{1/2})\\
&\quad p_{n+k-1/2} = a + e_{k-1/2}\\
&\quad e_{k-1/2} = \left(a - p_{n+k-1/2}\right) + e_{k-1/2}\\
&\quad \text{for } j = 1, \dots, k-1 \text{ do}\\
&\qquad e_{j-1/2} = e_{j+1/2}\\
&\quad \text{end do}\\
&\quad b = q_{n+k-1}\\
&\quad e = h p_{n+k-1/2} + e\\
&\quad q_{n+k} = b + e\\
&\quad e = (b - q_{n+k}) + e\\
&\text{end do}
\end{aligned}
$$

We remark that we have here k vectors instead of $k-1$, but the value $e_{k-1/2}$ can be saved in the position reserved to $e_{1/2}$.

In Figure 5.5 we compare the error in the Hamiltonian obtained with a symmetric multistep method of order 8 using compensated summations, with the error obtained using the same method but implementing (5.3) without. We observe that, unlike what happens for unconstrained systems (as checked in [CH13b]), here compensated summations are not enough to improve the accuracy of the algorithm.

However, as we will see in the next paragraph, they are nevertheless essential for building a tool that will succeed in improving accuracy.

5.2.5 Solution of nonlinear system: accurate constraint

In this section we show how compensated summations can also be used to fix a problem with the computation of the Lagrange multipliers.

The simplified Newton iteration that is used for the computation of the Lagrange multipliers is of the form

$$
\begin{aligned}
\delta^{(i+1)} &= -\left(\frac{\partial}{\partial\lambda} g(q_{k+1}(\lambda^{(0)}))\right)^{-1} g(q_{k+1}(\lambda^{(i)}))\\
\lambda^{(i+1)} &= \lambda^{(i)} + \delta^{(i+1)}
\end{aligned}
\tag{5.6}
$$

where $\lambda^{(0)}$ is the value of the Lagrange multiplier computed at the previous step.

During the computation of the increment $\delta^{(i+1)}$ at each iteration q will make $g(q)$ closer to 0 and, for example, in the case of a quadratic constraint (like the two-body problem on the sphere or for a pendulum) of the form $g(q) = q_1^2 + q_2^2 - 1$, the subtraction becomes more and more affected by cancellation error increasing the round-off error. Therefore we need to compute $g(q)$ with very high accuracy in order to avoid the loss of significant digits due to this problem.

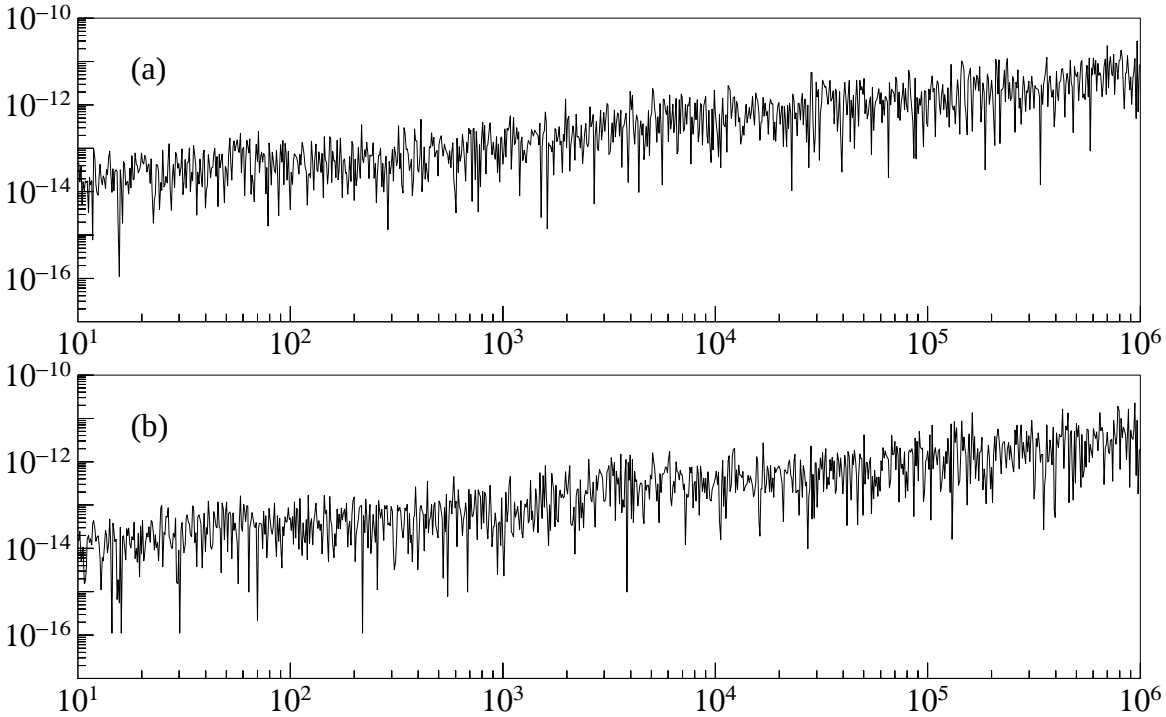


Figure 5.5 – (Two-body problem on the sphere). Comparison of the errors in the Hamiltonian of an 8th order symmetric multistep method as a function of time, obtained with compensated summations (b) and without (a). $h = 0.001$.

A trivial way to solve this problem is to evaluate the constraint in quadruple precision, convert the result to double precision and use this value for the computation of $\delta^{(i+1)}$, but on long time integrations this can increase the cpu time needed for the computation.

Another way to attack the problem is, as reported in [CH13b], to exploit the small error for the position accumulated in e with compensated summations and to use it to obtain a more accurate form of the constraint with a cpu time comparable to the less accurate formulation. For example, for a quadratic constraint, in order to avoid the dangerous subtraction, we can approximate $q_i \approx k_i/k$ where k_i and k are integers and k is very large, compute $d_i = (kq_i - k_i) + ke_i$ (where $e = (e_1, e_2)$ is the vector that collects the small errors obtained with compensated summations), and evaluate the constraint as

$$g(q+e) = \frac{1}{k^2}\left((k_1^2 + k_2^2 - k^2) + 2(k_1 d_1 + k_2 d_2) + d_1^2 + d_2^2\right) : \tag{5.7}$$

in this way the dangerous subtraction between two close real numbers is transformed in a safer subtraction between integers.

In Figure 5.6 we compare the errors in the Hamiltonian obtained with a routine of order 8 using both the standard and accurate formulation of the constraint. We remark that using the accurate formulation we obtain an algorithm that is about 10 time more precise.

We now want to check if the remarkable improvement shown in Figure 5.6 depends only on the elimination of the dangerous subtraction in the constraint or if it also depends on compensated summations.

In Figure 5.7 it is shown the comparison of the error obtained using the formulation of the constraint (5.7) with $e = 0$ (i. e. without compensated summations) and with compensated summations: we observe that the improvement seen in Figure 5.6 depends on the use of both the techniques, and so compensated summations remain a very important

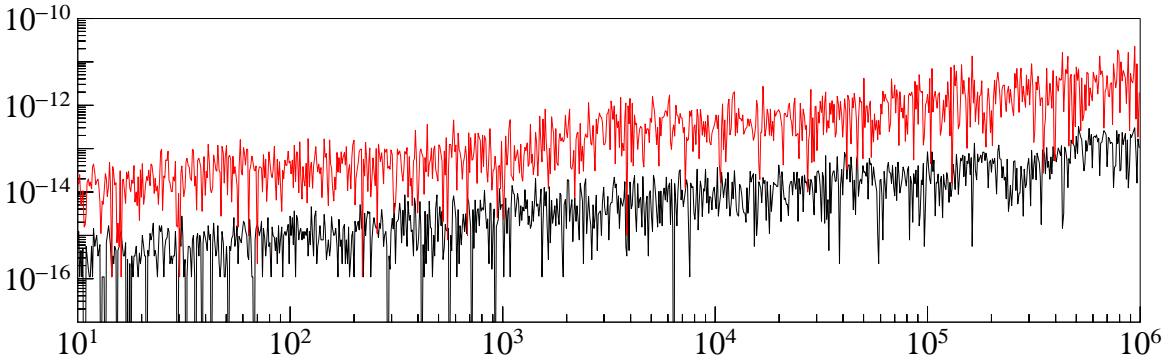


Figure 5.6 – (Two-body problem on the sphere). Comparison of the errors in the Hamiltonian of an 8th order symmetric multistep method as function of time, obtained with the standard constraint (red) and with the accurate formulation (black). Both algorithms have been implemented with compensated summations. $h = 0.001$.

tool in the application of multistep methods even for the constrained case.

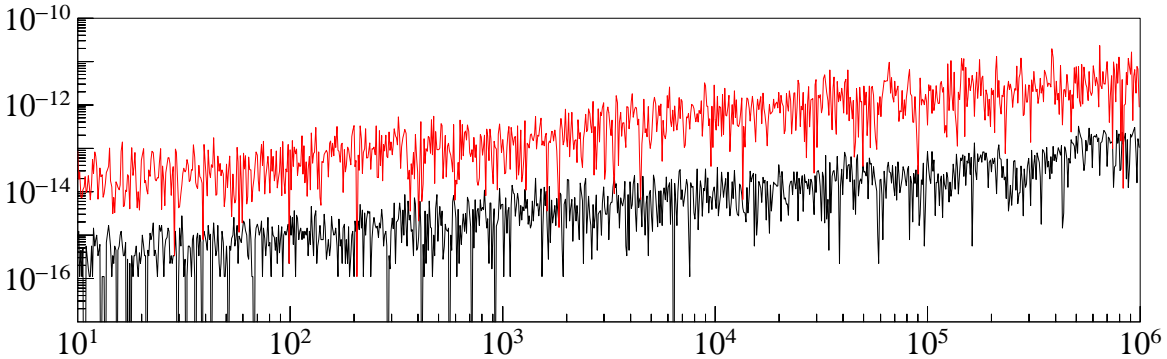


Figure 5.7 – (Two-body problem on the sphere). Comparison of the errors in the Hamiltonian of an 8th order symmetric multistep method as function of time, obtained with compensated summations (black) and without (red): both algorithms have been implemented with the accurate constraint (5.7). $h = 0.001$.

5.2.6 Solution of nonlinear aystem: iteration until convergence

Another issue with the computation of the Lagrange multipliers is the choice of the stop criterion for the Newton iteration (5.6): a very basic one is

$$\left\| \delta^{(i+1)} \right\| \leq tol \tag{5.8}$$

where *tol* is a tolerance fixed by the user.

This standard method can be improved by using a machine independent stop criterion called *iteration until convergence*, described by the following expression

$$\delta^{(i+1)} = 0 \quad \text{or} \quad \left\| \delta^{(i+1)} \right\| \geq \left\| \delta^{(i)} \right\|, \tag{5.9}$$

i.e. the Newton iteration stops if the increment $\delta^{(i)}$ is zero or if it starts to oscillate because of round-off: when the stopping criterion is satisfied the method returns $\lambda^{(i)}$.
In this way we make the numerical solution independent of the tolerance, eliminating a possible cause of deterministic error. We compare these stopping criteria in Figure 5.8: using (5.8) we notice a linear growth of the error after about 10^7 steps, but we see that using (5.9) the error behaves like a random walk for definitely longer times.

For this reason we keep the stopping criterion (5.9) in the final implementation reported in Appendix A.

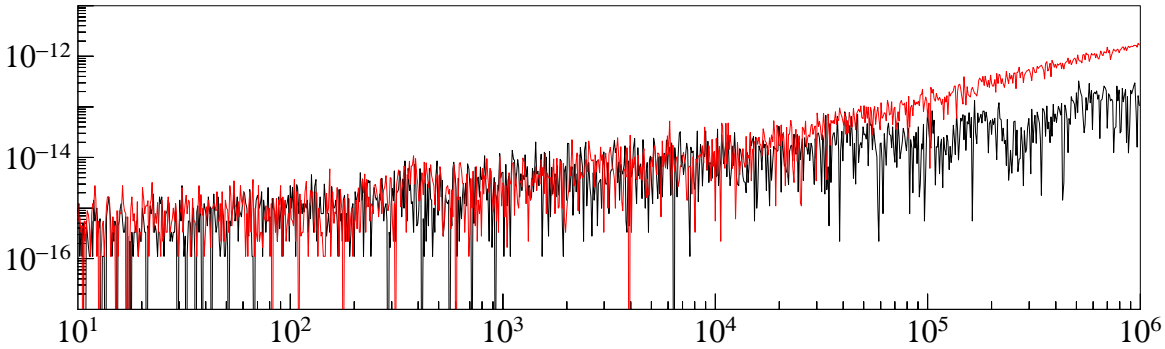


Figure 5.8 – (Two-body problem on the sphere). Comparison of the errors in the Hamiltonian of an 8th order symmetric multistep method as function of time, obtained using iteration until convergence (black) and the standard stopping criterion with $tol = 10^{-17}$ (red). Both algorithms have been implemented using the accurate formulation of the constraint. $h = 0.001$.

5.2.7 Programming choices

For a good implementation of a multistep method it is important to keep its symmetry by implementing it as follows (here for k=4)

$$p_{n+7/2} = p_{n+1/2} + \frac{1}{den}\left(\hat{\alpha}_1\left(p_{n+3/2} - p_{n+5/2}\right) + h\left(f_{n+1} + f_{n+3/2}\right) + \beta_2 f_{n+2}\right),$$

where $f_n = f(q_n) - G(q_n)^{\mathsf{T}}$.

Another way to avoid deterministic errors is to decrease the number of operations to be made. We note in the computation of q_{n+k} in (5.3), the momenta only appear multiplied by h: so we can compute with the multistep method the variable $hp_{n+k-1/2}$ instead of $p_{n+k-1/2}$.

This choice also has advantages in the solution of the nonlinear system: the variable of the Newton iteration is $\lambda = h^2\lambda_{n+k-1}$, which can be directly used for the computation of $hp_{n+k-1/2}$ and then for q_{n+k-1}.

We pass to the subroutine that computes the Lagrange multipliers the values a determined by the expression

$$hp_{n+k-1/2} = a - h^2\beta_{k-1}G(q_{n+k-1})^{\mathsf{T}}\lambda_{n+k-1}/\hat{\alpha}_0,$$

q_{n+k-1} and their corresponding errors obtained with compensated summations; they are used to update the Lagrange multiplier with the simplified Newton iteration in (5.6) using the constraint described in (5.7), where e is a temporary vector created using the small errors corresponding to a and q_{n+k-1}. After the computation of $h^2\lambda_{n+k-1}$, we can update $hp_{n+k-1/2}$ and q_{n+k-1} and compute the corresponding errors $e_{k-1/2}$ and e that will be passed to the next iteration.

5.3 Further numerical experiments

5.3.1 Constrained masses-springs system

We want to test the routine in Appendix A on a system consisting a set of six masses connected by four springs of constant κ (here considered all equal to 1) and five rigid bars as in Figure 5.9 (this system has been described in [LS94]).

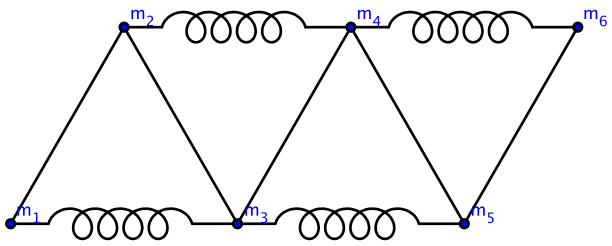


Figure 5.9 – Constrained masses-springs system

Using cartesian coordinates to solve the problems we have $p, q \in \mathbb{R}^{12}$, $M = I_{12}$, and $U = \frac{5}{2} q^{\mathsf{T}} K q$, where

$$K = \kappa \begin{pmatrix} I_2 & 0 & -I_2 & 0 & 0 & 0 \\ 0 & I_2 & 0 & -I_2 & 0 & 0 \\ -I_2 & 0 & 2I_2 & 0 & -I_2 & 0 \\ 0 & -I_2 & 0 & 2I_2 & 0 & -I_2 \\ 0 & 0 & -I_2 & 0 & I_2 & 0 \\ 0 & 0 & 0 & -I_2 & 0 & I_2 \end{pmatrix}.$$

The constraints are given by the fixed lengths (here equal to 1) of the bars, that is

$$g(q) = \begin{pmatrix} (q_1 - q_3)^2 + (q_2 - q_4)^2 - 1 = 0 \\ (q_3 - q_5)^2 + (q_4 - q_6)^2 - 1 = 0 \\ (q_5 - q_7)^2 + (q_6 - q_8)^2 - 1 = 0 \\ (q_7 - q_9)^2 + (q_8 - q_{10})^2 - 1 = 0 \\ (q_9 - q_{11})^2 + (q_{10} - q_{12})^2 - 1 = 0 \end{pmatrix}.$$

We use two different sets of initial data for the numerical experiments: in Figure 5.10 we show the error obtained using $q_0 = (0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1)$ (all the springs have initial extension equal to zero) and $p_0 \approx (0.5, 0.35, 0.5, 0.35, 0.4, 0.35, 0.3, 0.35, 0.53, 0.35, 0.72, 0.17)$, and in Figure 5.11 the error obtained using $q_0 = (0, 0, 0, 1, 0, 0, 0, 1, 0, 0, \sqrt{2}/2, \sqrt{2}/2)$ (all the springs have initial extension equal to zero except the last one, whose extension is about 0.7654) and initial velocities all equal to zero; we make the computations using the routines of order 4, 6 and 8.

We see that in both cases the error in the energy stays bounded for very long times integrations, confirming the theoretical results explained in Chapter 3; the results shown are obtained for a relatively large stepsize, anyway the order of the method has been verified halving the stepsize.

5.3.2 Triple pendulum

In this section we want to report the results obtained using the routine in Appendix A on a system less regular than the two-body problem seen in the previous section.

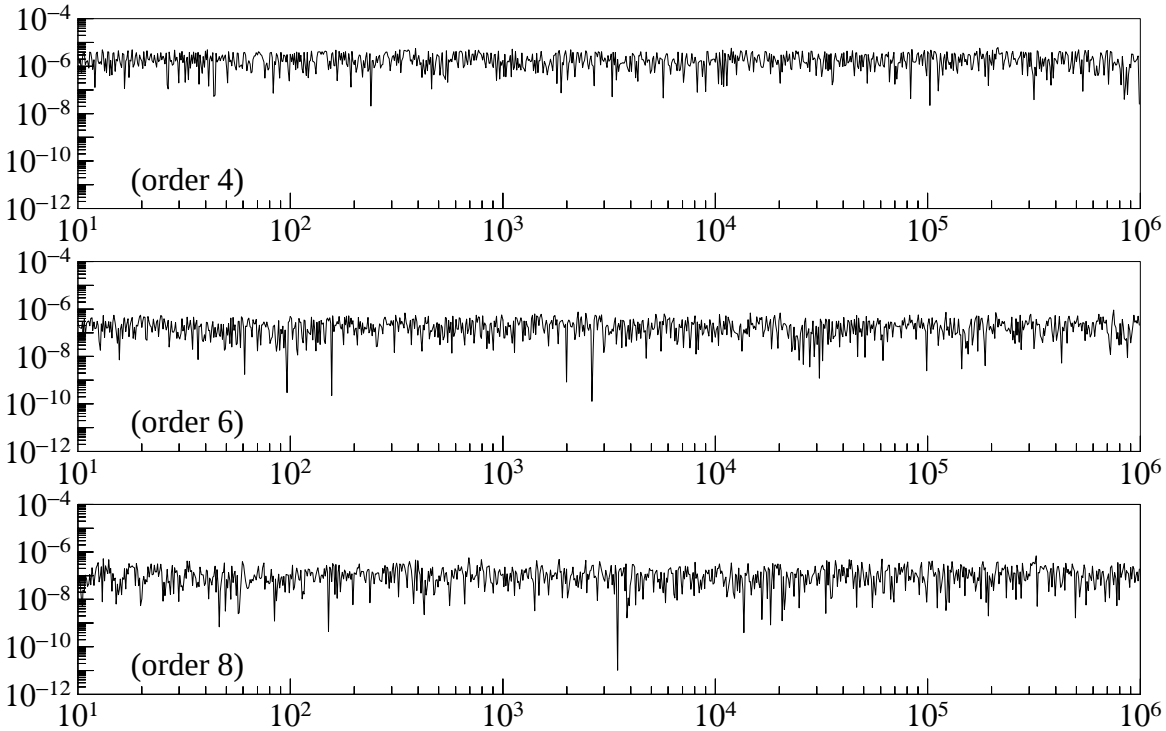


Figure 5.10 – (Constrained masses-springs system). Errors in the Hamiltonian of a 4th, 6th and 8th order algorithm as functions of time with $q_0 = (0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1)$ and $p_0 \approx (0.5, 0.35, 0.5, 0.35, 0.4, 0.35, 0.3, 0.35, 0.53, 0.35, 0.72, 0.17)$, $h = 0.01$.

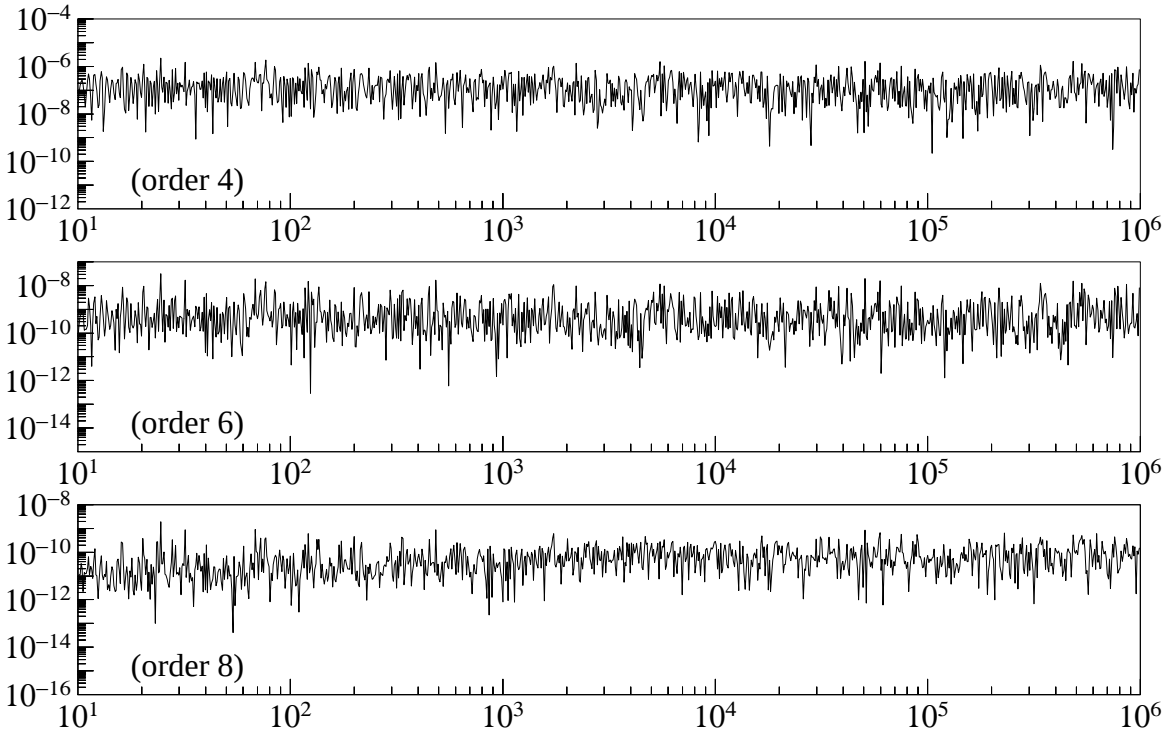


Figure 5.11 – (Constrained masses-springs system). Errors in the Hamiltonian of a 4th, 6th and 8th order algorithm as functions of time with $q_0 = q_0 = (0, 0, 0, 1, 0, 0, 0, 1, 0, 0, \sqrt{2}/2, \sqrt{2}/2)$ and $p_0 = 0$, $h = 0.01$.

We test the final routine on the triple pendulum using different sets of initial data, and with stepsizes that make the discretization error about the size of the machine precision.
The system has been described in Section 3.5: it is a mathematical pendulum suspended at the origin, and it is written in Cartesian coordinates. We make simulations with the four sets of data described by Figure 5.12; in all the cases the initial momenta are $p(0) = (0, \ldots, 0)$.

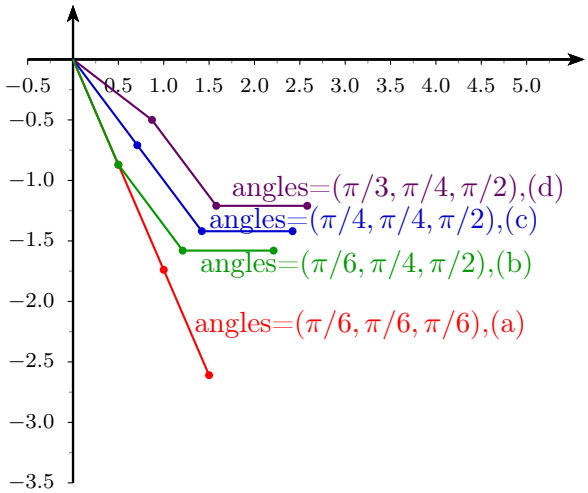


Figure 5.12 – Triple pendulum: initial configurations

We apply to these systems the routine of order 4 and the routine of order 6: we remark that the configurations (b), (c) and (d) give rise to chaotic motion.
In Figure 5.13 we report the errors in the Hamiltonian of the order 4 method used on the four systems described above: we easily observe that using the same stepsize the error is smaller if the system is less chaotic.

In Figure 5.14 it is reported the error in the Hamiltonian of the order 6 method applied to the equations of the triple pendulum. We can observe sudden growths of the error in the graphics related to chaotic systems ((b), (c) and (d)); these increases are due to violent change of motions caused by the chaoticity of the system.

This idea is confirmed by Figure 5.15, which reports the error in the Hamiltonian obtained using the same routine on the equations of the triple pendulum with the initial sets of data (b), (c) and (d), but with halved stepsize: we observe the increase of the error later and this is superposed with the erratic behaviour due to round-off error.

5.4 Probabilistic explanation of the error growth

We now give a probabilistic interpretation of the random-walk behaviour of the round-off errors, often called Brouwer's Law, described in [Hen62]. This interpretation has already been explained in [Vil08] .
The error in the Hamiltonian after one step is

$$H(p_{n+1}, q_{n+1}) - H(p_n, q_n) = \epsilon_n$$

where ϵ_n is a random variable with variance proportional to eps^2 (eps is the machine precision).
Eliminating all the sources of deterministic errors means setting $E(\epsilon_n) = 0$, and so removing the linear growth of the round-off error: the sum of ϵ_n after N steps has mean zero and variance proporitional to $Neps^2$, and so the error in the Hamiltonian after N steps grows like

$$Var(err_N)^{1/2} = c\, eps\sqrt{N},$$

i.e. like a random walk.

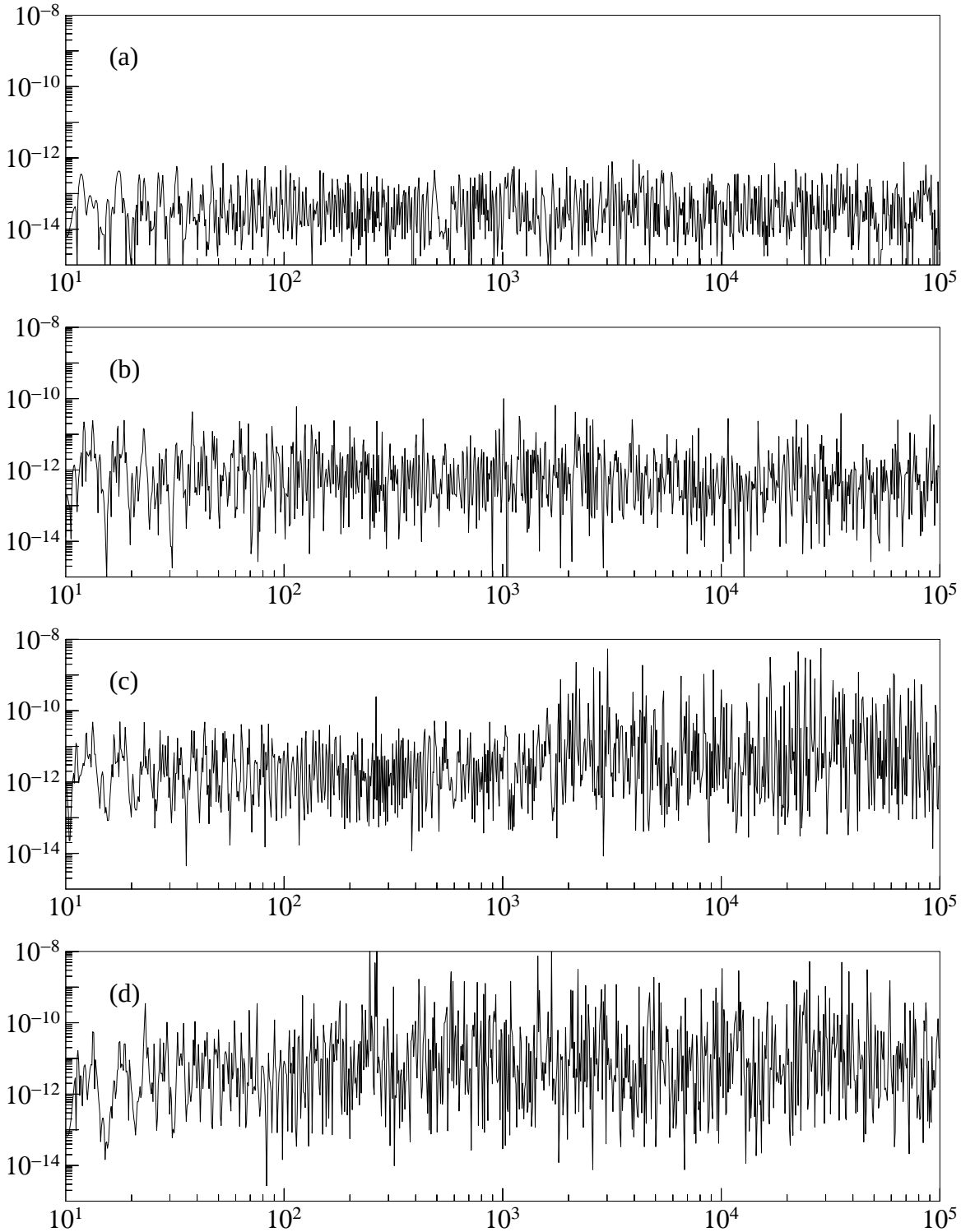


Figure 5.13 – (Triple pendulum). Error in the Hamiltonian of a 4th order algorithm as a function of time, $h = 0.001$. (a), (b), (c) and (d) respectively show the errors obtained using the initial data corresponding to the angles $(\pi/6, \pi/6, \pi/6)$, $(\pi/6, \pi/4, \pi/2)$, $(\pi/4, \pi/4, \pi/2)$ and $(\pi/6, \pi/4, \pi/2)$. All these routines have been implemented using the techniques described in this chapter.

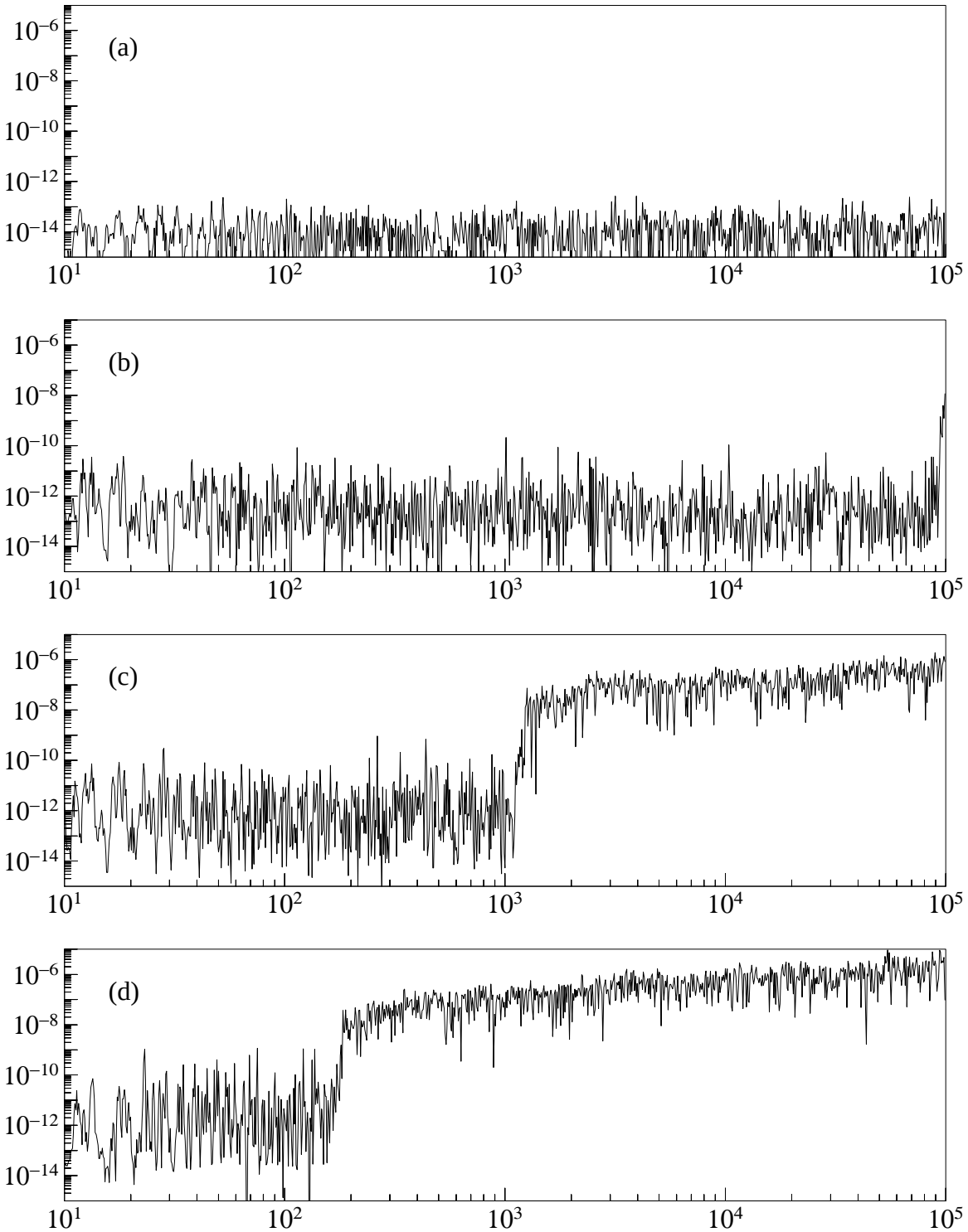


Figure 5.14 – (Triple pendulum). Error in the Hamiltonian of a 6th order algorithm as a function of time, $h = 0.005$. (a), (b), (c) and (d) respectively show the errors obtained using the initial data corresponding to the angles $(\pi/6, \pi/6, \pi/6)$, $(\pi/6, \pi/4, \pi/2)$, $(\pi/4, \pi/4, \pi/2)$ and $(\pi/6, \pi/4, \pi/2)$. All these routines have been implemented using the techniques described in this chapter.

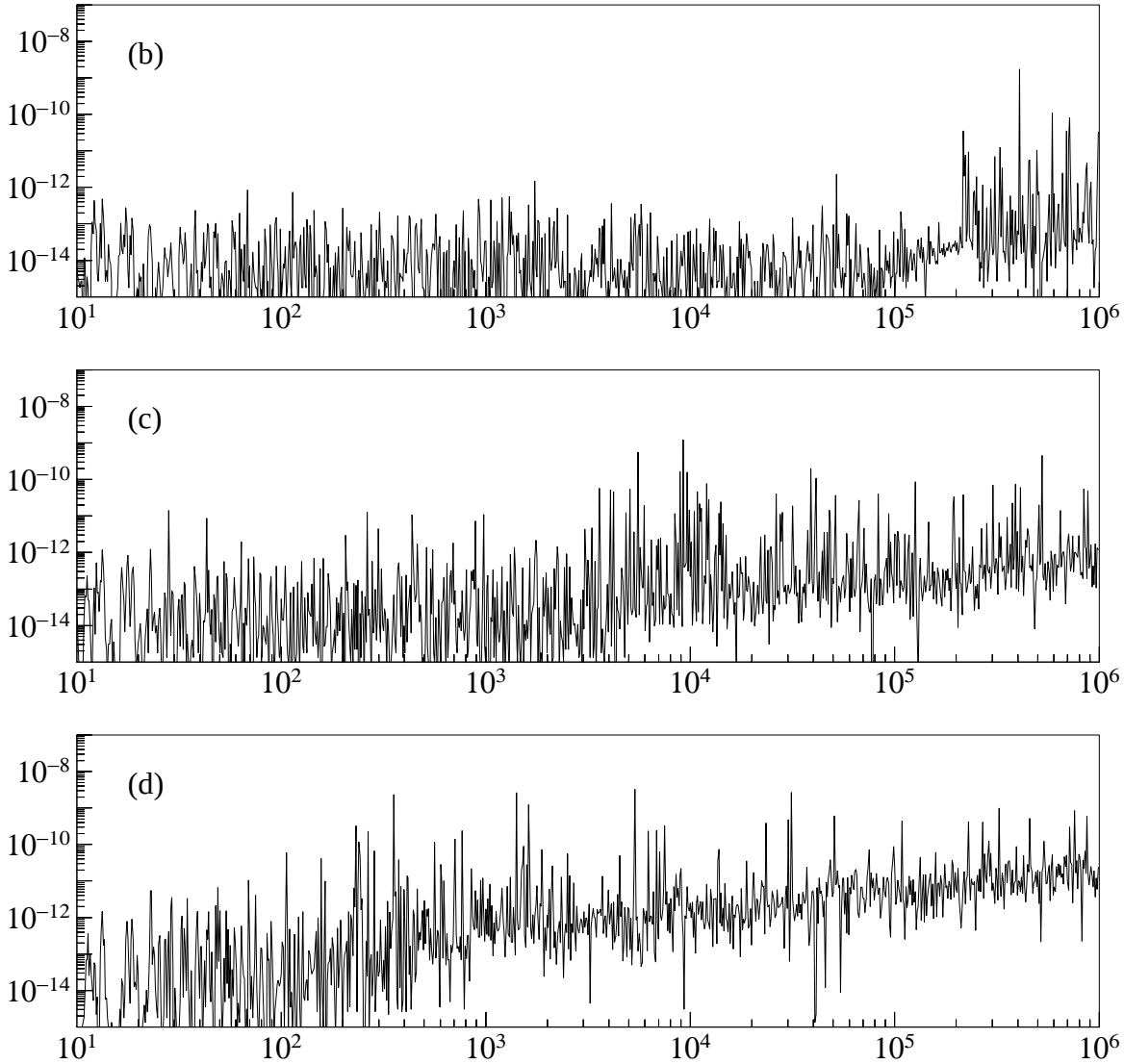


Figure 5.15 – (Triple pendulum). Error in the Hamiltonian of a 6th order algorithm as a function of time, $h = 0.0025$. (b), (c) and (d) respectively show the errors obtained using the initial data corresponding to the angles $(\pi/6, \pi/4, \pi/2)$, $(\pi/4, \pi/4, \pi/2)$ and $(\pi/6, \pi/4, \pi/2)$. All these routines have been implemented using the techniques described in this chapter.

Appendix A

Symmetric LMM for Constrained Hamiltonian Systems: Fortran 90 routine

In this Section the Fortran 90 routine implementing the classes of methods described in Chapter 3 is provided.

We remark that the use of LAPACK package (*http://www.netlib.org/lapack/*) is needed.

```
!!$-------------------------------------------------------------
!!$
!!$                  Linear Symmetric Multistep Methods
!!$                  for Constrained Hamiltonian Systems
!!$
!!$                   Version of May, 25th, 2013
!!$
!!$  e-mail contact address : paola.console@unige.ch
!!$
!!$-------------------------------------------------------------
!!$  This routine solves constrained second order equations like
!!$
!!$                        q'' = f(q),    g(q)=0
!!$
!!$  based on symmetric linear multistep methods
!!$              described in Chapter 3
!!$
!!$  and in the publication
!!$
!!$      P. Console, E. Hairer, C. Lubich -
!!$          Symmetric Multistep Methods
!!$      for Constrained Hamiltonian Systems
!!$
!!$
!!$  subroutine
!!$     lmmconstrained(m,n,k,q0,p0,t0,tf,MAT,h,f,Jg,g,indred,
!!$                    iout,solout)
!!$
!!$  INPUT..
!!$
!!$     m           Dimension of the constraint g(q)
!!$
```

```
!!$     n           Dimension of the positions q and of the
!!                  momenta p
!!$
!!$     k           Order of the Method
!!$                 Available choices: 4, 6, 8
!!$
!!$     q0,p0       Initial positions and momenta
!!$
!!$     t0,tf       Initial and final time of integration
!!$
!!$     MAT         Symmetric positive definite mass matrix
!!$
!!$     h           timestep
!!$
!!$     f           name (external) of subroutine computing f(q)
!!$                 subroutine f(q,res)
!!$                     real(kind=dp), dimension(n), intent(in)
!!$                                    :: q
!!$                     real(kind=dp), dimension(n), intent(out)
!!$                                    :: res
!!$
!!$     g           name (external) of subroutine computing the
!!$                 constraint g(q). errq is the error obtained
!!$                 with compensated summations.
!!$                   subroutine g(q,errq,res)
!!$                     real(kind=dp), dimension(n), intent(in)
!!$                                    :: q,errq
!!$                     real(kind=dp), dimension(m), intent(out)
!!$                                    :: err
!!$
!!$     Jg          name (external) of subroutine computing the
!!$                 Jacobian of the constraint g(q)
!!$                   subroutine jacconstraint(q,jac)
!!$                    real(kind=dp), dimension(n), intent(in)
!!$                                   :: q
!!$                    real(kind=dp), dimension(m,n), intent(out)
!!$                                   :: jac
!!$
!!$     indred      name (external) of subroutine computing the
!!$                 quantity (J(q)M^(-1)p)_qM^(-1)p where _q
!!$                 represents the Jacobian with respect to q:
!!$                 MAT is the LU factorization of the mass
!!$                 matrix, ipvn the integer pivot vector
!!$                 obtained during the factorization
!!$
!!$                   subroutine indred(q,p,MAT,ipivn,res)
!!$                      real(kind=dp), dimension(n), intent(in)
!!$                                     :: p,q
!!$                      real(kind=dp), dimension(n,n), intent(in)
!!$                                     :: MAT
!!$                      integer, dimension(n) :: ipivn
!!$                      real(kind=dp), dimension(m) :: res
!!$
!!$     iout        switch for calling the subroutine solout
!!$                      iout=0: subroutine is never called
!!$                      iout=1: subroutine is available for input
```

```
!!$
!!$     solout      name (external) of subroutine providing
!!$                 the numerical solution during integration.
!!$                 if iout=1, it is called after every step.
!!$                 j is an integer reporting the current timestep
!!$                 supply a dummy subroutine if iout=0.
!!$                    subroutine solout(j,newp,newq,n)
!!$                       integer(kind=dp), intent(in) :: j,n
!!$                       real(kind=dp), dimension(n) :: newp,newq
!!$
!!$
!!$  OUTPUT..
!!$     p0,q0       numerical solution at tf


module precision

  integer, parameter :: dp=kind(1.0d0)
  integer, parameter :: qp=selected_real_kind(22)

end module precision



module initial

  use precision
  implicit none

contains


  subroutine projectionq(n,m,q,h,MAT,f,indred,Jg,g,projq)

    external Jg,g,f,indred,dgetri,dgetrs
    integer, intent(in) :: m,n
    real(kind=dp), intent(in) :: h
    real(kind=dp), dimension(n), intent(in) :: q
    real(kind=dp), dimension(n,n), intent(in) :: MAT
    real(kind=dp), dimension(n), intent(out) :: projq
    real(kind=dp), dimension(n) :: newq,q0,errq
    real(kind=dp) :: test
    integer :: info
    real(kind=dp), dimension(m,n) :: Jacg1,Jacg2
    real(kind=dp), dimension(m,m) :: J
    real(kind=dp), dimension(m) :: lambda,lam0,g1,g2,b
    integer, dimension(m) :: ipiv
    integer, parameter :: lwork = 20000
    real(kind=dp), dimension(lwork) :: work
    real(kind=dp), parameter :: tol=1.e-15_dp

    errq=0._dp
    call Jg(q,Jacg1)
    lam0=0._dp
```

```
  do
     q0=q+matmul(transpose(Jacg1),lam0)
     call Jg(q0,Jacg2)
     J=matmul(Jacg2,transpose(Jacg1))
     call g(q0,errq,g1)
     b=g1
     call dgetrf(m,m,J,m,ipiv,work,lwork,info)
     call dgetrs('No transpose',m,1,J,m,ipiv,b,m,info)
     lambda=lam0-b
     newq=q+matmul(transpose(Jacg1),lambda)
     call g(newq,errq,g2)
     test=sqrt(dot_product(g2,g2))

     if (test<tol) then
        projq=newq
        exit
     end if

     q0=newq
     lam0=lambda
  end do

  return

end subroutine projectionq


subroutine projectionp(n,m,MAT,invM,ipivn,p,q,Jg,res)

  external Jg,dgetrf,dgetrs
  integer, intent(in) :: m,n
  real(kind=dp), dimension(n,n), intent(in) :: MAT,invM
  integer, dimension(n), intent(in) :: ipivn
  real(kind=dp), dimension(n), intent(in) :: p,q
  real(kind=dp), dimension(n), intent(out) :: res
  real(kind=dp), dimension(n) :: incr
  real(kind=dp), dimension(m,n) :: J
  real(kind=dp), dimension(m,m) :: B
  real(kind=dp), dimension(m) :: c
  integer, dimension(m) :: ipiv
  integer :: info
  integer, parameter :: lwork=20000,lworkn=20000
  real(kind=dp), dimension(lwork) :: work, workn

  call Jg(q,J)
  B=matmul(matmul(J,invM),transpose(J))
  call dgetrf(m,m,B,m,ipiv,work,lwork,info)
  incr=p
  call dgetrs('No transpose',n,1,MAT,n,ipivn,incr,n,info)
  c=matmul(J,incr)
  call dgetrs('No transpose',m,1,B,m,ipiv,c,m,info)
  res=p-matmul(transpose(J),c)
  return

end subroutine projectionp
```

```
subroutine indexreduction(m,n,q,p,MAT,invM,ipivn,f,Jg,indred,&
     lambda)

  external indred,f,Jg,dgetri,dgetrs
  integer, intent(in) :: m,n
  real(kind=dp), dimension(n), intent(in) :: p,q
  real(kind=dp), dimension(n,n), intent(in) :: MAT,invM
  integer, dimension(n), intent(in) :: ipivn
  real(kind=dp), dimension(m), intent(out):: lambda
  real(kind=dp), dimension(n,n) :: invMAT
  real(kind=dp), dimension(m) :: B,bloc
  real(kind=dp), dimension(m,n) :: G
  real(kind=dp), dimension(n) :: dU
  real(kind=dp), dimension(m,m) :: A
  integer :: info
  integer, dimension(m) :: ipiv
  integer, parameter :: lwork=20000,lworkn=20000
  real(kind=dp), dimension(lwork) :: work, workn

  call indred(q,p,MAT,ipivn,bloc)
  call Jg(q,G)
  call f(q,dU)
  A=matmul(G,matmul(invM,transpose(G)))
  call dgetrs('No␣transpose',n,1,MAT,n,ipivn,dU,n,info)
  B=bloc+matmul(G,dU)
  call dgetrf(m,m,A,m,ipiv,work,lwork,info)
  call dgetrs('No␣transpose',m,1,A,m,ipiv,B,m,info)
  lambda=B
  return

end subroutine indexreduction


subroutine field(m,n,x,MAT,invM,ipivn,f,Jg,indred,res)

  external f,Jg,indred
  integer, intent(in) :: m,n
  real(kind=dp), dimension(2*n), intent(in) ::x
  real(kind=dp), dimension(n,n), intent(in) :: MAT,invM
   integer, dimension(n),intent(in) :: ipivn
  real(kind=dp), dimension(2*n), intent(out) :: res
  real(kind=dp), dimension(m) :: lambda
  real(kind=dp), dimension(n) :: Hq,p,q
  real(kind=dp), dimension(m,n) :: G
  integer :: info

  q=x(1:n)
  p=x(n+1:2*n)
  call indexreduction(m,n,q,p,MAT,invM,ipivn,f,Jg,indred,&
       lambda)
  call f(q,Hq)
  call Jg(q,G)
  res(n+1:2*n)=Hq-matmul(transpose(G),lambda)
  call dgetrs('No␣transpose',n,1,MAT,n,ipivn,p,n,info)
  res(1:n)=p
```

```
  return

end subroutine field


subroutine rungekutta(m,n,q,p,MAT,invM,ipivn,h,f,Jg,indred,res)

  external indred,f,Jg
  integer, intent(in) :: m,n
  real(kind=dp), dimension(n),intent(in) :: p,q
  real(kind=dp), dimension(n,n), intent(in) :: MAT,invM
  integer, dimension(n), intent(in) :: ipivn
  real(kind=dp), intent(in) :: h
  real(kind=dp), dimension(2*n), intent(out) :: res
  real(kind=dp), dimension(8*n) :: kold,knew
  real(kind=dp), dimension(2*n) :: f1,f2,f3,f4
  real(kind=qp), dimension(4,4) :: A
  real(kind=qp) :: omega1,omega2,omega3,omega4,omega5,&
       omega1prime,omega2prime,omega3prime,omega4prime,&
       omega5prime

  omega1=1._qp/8._qp-sqrt(30._qp)/144._qp
  omega2=0.5_qp*sqrt((15._qp+2*sqrt(30._qp))/35._qp)
  omega3=omega2*(1._qp/6._qp+sqrt(30._qp)/24._qp)
  omega4=omega2*(1._qp/21._qp+5._qp*sqrt(30._qp)/168._qp)
  omega5=omega2-2*omega3
  omega1prime=1._qp/8._qp+sqrt(30._qp)/144._qp
  omega2prime=0.5_qp*sqrt((15._qp-2*sqrt(30._qp))/35._qp)
  omega3prime=omega2prime*(1._qp/6._qp-sqrt(30._qp)/24._qp)
  omega4prime=omega2prime*(1._qp/21._qp-5._qp*&
       sqrt(30._qp)/168._qp)
  omega5prime=omega2prime-2*omega3prime
  A(1,1)=omega1
  A(1,2)=omega1prime-omega3+omega4prime
  A(1,3)=omega1prime-omega3-omega4prime
  A(1,4)=omega1-omega5
  A(2,1)=omega1-omega3prime+omega4
  A(2,2)=omega1prime
  A(2,3)=omega1prime-omega5prime
  A(2,4)=omega1-omega3prime-omega4
  A(3,1)=omega1+omega3prime+omega4
  A(3,2)=omega1prime+omega5prime
  A(3,3)=omega1prime
  A(3,4)=omega1+omega3prime-omega4
  A(4,1)=omega1+omega5
  A(4,2)=omega1prime+omega3+omega4prime
  A(4,3)=omega1prime+omega3-omega4prime
  A(4,4)=omega1
  kold=(/q,p,q,p,q,p,q,p/)

  do
     kold(1:2*n)=kold(1:2*n)
     kold(2*n+1:4*n)=kold(2*n+1:4*n)
     kold(4*n+1:6*n)=kold(4*n+1:6*n)
     kold(6*n+1:8*n)=kold(6*n+1:8*n)
     f1=(/q,p/)+h*(A(1,1)*kold(1:2*n)+A(1,2)*kold(2*n+1:4*n)&
```

```
           +A(1,3)*kold(4*n+1:6*n)+A(1,4)*kold(6*n+1:8*n))
      f2=(/q,p/)+h*(A(2,1)*kold(1:2*n)+A(2,2)*kold(2*n+1:4*n)&
           +A(2,3)*kold(4*n+1:6*n)+A(2,4)*kold(6*n+1:8*n))
      f3=(/q,p/)+h*(A(3,1)*kold(1:2*n)+A(3,2)*kold(2*n+1:4*n)&
      +A(3,3)*kold(4*n+1:6*n)+A(3,4)*kold(6*n+1:8*n))
      f4=(/q,p/)+h*(A(4,1)*kold(1:2*n)+A(4,2)*kold(2*n+1:4*n)&
           +A(4,3)*kold(4*n+1:6*n)+A(4,4)*kold(6*n+1:8*n))
      call field(m,n,f1,MAT,invM,ipivn,f,Jg,indred,knew(1:2*n))
      call field(m,n,f2,MAT,invM,ipivn,f,Jg,indred,&
           knew(2*n+1:4*n))
      call field(m,n,f3,MAT,invM,ipivn,f,Jg,indred,&
           knew(4*n+1:6*n))
      call field(m,n,f4,MAT,invM,ipivn,f,Jg,indred,&
           knew(6*n+1:8*n))

      if (sqrt(dot_product(knew-kold,knew-kold))<1e-15) then
         res=2*(omega1*(knew(1:2*n)+knew(6*n+1:8*n))&
              +omega1prime*(knew(2*n+1:4*n)+knew(4*n+1:6*n)))
         return
      end if

      kold=knew
   end do

   return

end subroutine rungekutta


subroutine initialdata(k,m,n,p0,q0,MAT,invM,ipivn,h,f,Jg&
      ,indred,p,q,l,evalf,jac)

   external indred,f,Jg
   integer, intent(in) :: k,m,n
   real(kind=dp), dimension(n) :: p0,q0
   real(kind=dp), dimension(n,n), intent(in) :: MAT,invM
   integer, dimension(n), intent(in) :: ipivn
   real(kind=dp), intent(in) :: h
   real(kind=dp), dimension(n,(k-1)),intent(out) :: p,q
   real(kind=dp), dimension(m,k-1), intent(out) :: l
   real(kind=dp), dimension(n,k-1), intent(out) :: evalf
   real(kind=dp), dimension(m,n*(k-1)), intent(out) :: jac
   integer :: s,j,i
   real(kind=dp), dimension(2*n) :: x0,incr,res,dummy
   real(kind=dp), dimension(m) :: ldummy

   s=1
   x0=(/q0,p0/)
   do j=1,k-1

      incr=0.

      do i=1,s
         call rungekutta(m,n,q0,p0,MAT,invM,ipivn,h/s,&
              f,Jg,indred,res)
         q0=q0+(h/s)*res(1:n)
```

```
          p0=p0+(h/s)*res(n+1:2*n)
          incr=incr+res
       end do

       dummy=x0+(h/s)*incr
       q0=dummy(1:n)
       p0=dummy(n+1:2*n)
       x0=dummy
       p(:,j)=h*matmul(MAT,incr(1:n)/s)
       q(:,j)=dummy(1:n)
       call indexreduction(m,n,q0,p0,MAT,invM,ipivn,f,Jg,&
            indred,ldummy)
       l(:,j)=ldummy*h**2
       call f(q(:,j),evalf(:,j))
       call Jg(q(:,j),jac(:,n*(j-1)+1:n*j))
    end do

    return

  end subroutine initialdata


end module initial



module lmmconstrainedmodule

  use doublep
  use precision
  use initial
  implicit none

contains

  subroutine coefficients(k,alpha,beta,den)

    integer, intent(in) :: k
    real(kind=dp), intent(out) :: den
    real(kind=dp), dimension(k/2-1), intent(out) :: alpha
    real(kind=dp), dimension(k/2), intent(out) :: beta
    integer :: a1,a2,a3,s1,s2,s3

    if (k==4) then
       a1=0
       alpha=(/-6+12*a1/)
       beta=(/(7._dp+a1),(-1._dp+5._dp*a1)*2._dp/)
       den=6._dp
    end if

    if (k==6) then
       a1=-7
       a2=4
       s1=a1+a2
       s2=a1*a2
       alpha=(/-1200*s1+6000,-240*s2+1200*s1-12000/)
```

```
      beta=(/(7900+90*s1-s2),(-5600+1040*s1+24*s2),&
           (140*s1+194*s2+19400)/)
      den=6000
   end if

   if (k==8) then
      a1=-8
      a2=-4
      a3=7
      s1=a1+a2+a3
      s2=a1*a2+a2*a3+a3*a1
      s3=a1*a2*a3
      alpha=(/1512000*s1-7560000,302400*s2-1512000*s1+22680000,&
           60480*s3+3024000*s1-302400*s2-22680000/)
      beta=(/10993000+103900*s1-950*s2+31*s3,-13290000&
           +1367400*s1+28380*s2-438*s3,49983000+147300*s1&
           +247830*s2+6513*s3,-34892000+2810800*s1+54280*s2&
           +48268*s3/)
      den=7560000
   end if

   return

end subroutine coefficients


subroutine finitedifferences(n,meth,h,p,newp)

   integer, intent(in) :: n,meth
   real(kind=dp), intent(in):: h
   real(kind=dp), dimension(n,meth), intent(in) :: p
   real(kind=dp), dimension(n), intent(out) :: newp

   if (meth==4) then
      newp=(-(p(:,1)+p(:,4))+7._dp*(p(:,2)+p(:,3)))/(12._dp*h)
   end if

   if (meth==6) then
      newp=(p(:,1)+p(:,6)-8._dp*(p(:,2)+p(:,5))+37._dp*(p(:,3)&
           +p(:,4)))/(60._dp*h)
   end if

   if (meth==8) then
      newp=(-3._dp*(p(:,1)+p(:,8))+29._dp*(p(:,2)+p(:,7))&
           -139._dp*(p(:,3)+p(:,6))+533._dp*(p(:,4)&
           +p(:,5)))/(840._dp*h)
   end if

   return

end subroutine finitedifferences


subroutine lagrangemultipliers(m,n,meth,alpha,beta,den,a,q,&
     roundoffa,roundoffq,lambda,jacobian,h,MAT,invM,ipivn,&
     g,Jg,newl)
```

```
      external g,Jg,dgetrf,dgetrs
      integer, intent(in) :: m,n,meth
      real(kind=dp), dimension(meth/2-1), intent(in) :: alpha
      real(kind=dp), dimension(meth/2), intent(in) :: beta
      real(kind=dp), dimension(n), intent(in):: a,q,roundoffa,&
           roundoffq
      real(kind=dp), dimension(m), intent(in) :: lambda
      real(kind=dp), dimension(m,n), intent(in) :: jacobian
      real(kind=dp), intent(in) :: h,den
      real(kind=dp), dimension(n,n), intent(in) :: MAT,invM
      integer, dimension(n), intent(in) ::ipivn
      real(kind=dp), dimension(m), intent(out) :: newl
      real(kind=dp), dimension(n) ::b,temp,roundoffb,roundoffdummy&
           ,dummy,qdummy,roundoffqdummy,adummy,roundoffadummy
      real(kind=dp), dimension(m):: lam0,lam
      real(kind=dp), dimension(m) :: g1
      real(kind=dp), dimension(m,m) :: J
      integer, dimension(m) ::ipiv
      integer :: info,i
      real(kind=dp) :: norm1,norm2
      integer, parameter :: lwork=20000
      real(kind=dp), dimension(lwork) :: work,workn

      adummy=a
      roundoffadummy=roundoffa
      call dgetrs('No␣transpose',n,1,MAT,n,ipivn,adummy,n,info)
      call dgetrs('No␣transpose',n,1,MAT,n,ipivn,roundoffadummy,&
           n,info)
      temp=q
      roundoffdummy=roundoffq+(adummy+roundoffadummy)
      b=temp+roundoffdummy
      roundoffb=(temp-b)+roundoffdummy
      J=-beta(1)*matmul(jacobian,matmul(invM,&
           transpose(jacobian)))/den
      lam0=lambda
      call dgetrf(m,m,J,m,ipiv,work,lwork,info)
      i=1

      do
         dummy=beta(1)*matmul(transpose(jacobian),lam0)/den
         call dgetrs('No␣transpose',n,1,MAT,n,ipivn,dummy,n,info)
         temp=b
         roundoffdummy=roundoffb-dummy
         qdummy=temp+roundoffdummy
         roundoffqdummy=(temp-qdummy)+roundoffdummy
         call g(qdummy,roundoffqdummy,g1)
         call dgetrs('No␣transpose',m,1,J,m,ipiv,g1,m,info)

         if (info.ne.0) then
            print*,"Error␣in␣the␣inversion␣of␣the␣Jacobian␣during&
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣the␣computation␣of␣the␣Lagrange␣multipliers!",info
            stop
         end if

         lam=lam0
```

```
      lam=lam0-g1
      norm2=sqrt(dot_product(g1,g1))

      if (norm2==0._dp .or. (norm2>=norm1 .and. i>1)) then
         newl=lam0
         exit
      end if

      norm1=norm2
      lam0=lam
      i=i+1
   end do

   return

end subroutine lagrangemultipliers


subroutine lmmconstrained(m,n,k,q0,p0,t0,tf,MAT,h,f,g,Jg,&
     indred,iout,solout)

  external f,g,Jg,indred,solout
  integer, intent(in) :: m,n,k,iout
  real(kind=dp), dimension(n), intent(inout) :: p0,q0
  real(kind=dp), intent(in) :: t0,tf,h
  real(kind=dp), dimension(n,n), intent(in) :: MAT
  real(kind=dp), dimension(n) :: newp,newq,rho,sigma,pzero,d,a,&
       b,c,incr,evalfdummy,pdummy,qdummy,newpdummy,roundoffq,&
       roundoffpdummy,roundoffa,fieldummy
  real(kind=dp), dimension(k/2-1) :: alpha
  real(kind=dp), dimension(k/2) :: beta
  real(kind=dp) :: den
  real(kind=dp), dimension(n,n) :: invM
  real(kind=dp), dimension(m) :: lambda,newlambda
  integer :: i,j,jj,info,grid
  real(kind=dp), dimension(n,k) :: p,q,roundoffp
  real(kind=dp), dimension(m,k-1) :: l
  real(kind=dp), dimension(n,k-1) :: evalf,field
  real(kind=dp), dimension(m,n*(k-1)) :: jac
  real(kind=dp), dimension(m,n) :: jacobian
  real(kind=dp), dimension(m,n) :: jacdummy
  integer, dimension(n) ::ipivn
  integer, parameter :: lwork=20000, lworkn=20000
  real(kind=dp), dimension(lwork) :: work,workn

  call dgetrf(n,n,MAT,n,ipivn,workn,lworkn,info)
  invM=MAT
  call dgetri(n,invM,n,ipivn,work,lwork,info)
  call coefficients(k,alpha,beta,den)
  call initialdata(k,m,n,p0,q0,MAT,invM,ipivn,h,f,Jg,indred,p,&
       q,l,evalf,jac)
  grid=nint((tf-t0)/h)
  roundoffp=0._dp
  roundoffq=0._dp
  roundoffa=0._dp
```

```
do j=1,k-2
   field(:,j)=h**2*evalf(:,j)-&
        matmul(transpose(jac(:,(j-1)*n+1:n*j)),l(:,j))
end do

field(:,k-1)=h**2*evalf(:,k-1)
jacobian=jac(:,(k-2)*n+1:n*(k-1))
lambda=l(:,k-2)

do j=k,grid

   if (k==4) then
      rho=alpha(1)*(p(:,2)-p(:,3))
      sigma=beta(1)*(field(:,3)+field(:,1))+beta(2)*field(:,2)
      d=alpha(1)*(roundoffp(:,2)-roundoffp(:,3))
   end if

   if (k==6) then
      rho=-alpha(2)*(p(:,3)-p(:,4))+alpha(1)*(p(:,5)-p(:,2))
      sigma=beta(1)*(field(:,5)+field(:,1))+&
           beta(2)*(field(:,4)+field(:,2))+beta(3)*field(:,3)
      d=(-alpha(2)*(roundoffp(:,3)-roundoffp(:,4))+alpha(1)*&
           (roundoffp(:,5)-roundoffp(:,2)))
   end if

   if (k==8) then
      rho=alpha(1)*(p(:,2)-p(:,7))+alpha(2)*(p(:,3)-p(:,6))&
      +alpha(3)*(p(:,4)-p(:,5))
      sigma=beta(1)*(field(:,7)+field(:,1))+beta(2)*&
           (field(:,6)+field(:,2))+beta(3)*(field(:,5)+&
           field(:,3))+beta(4)*field(:,4)
      d=(alpha(1)*(roundoffp(:,2)-roundoffp(:,7))+alpha(2)*&
           (roundoffp(:,3)-roundoffp(:,6))+alpha(3)*&
           (roundoffp(:,4)-roundoffp(:,5)))
   end if

   pzero=p(:,1)
   roundoffa=(rho+sigma)/den+(d/den+roundoffp(:,1))
   a=pzero+roundoffa
   roundoffa=(pzero-a)+roundoffa
   call lagrangemultipliers(m,n,k,alpha,beta,den,a,q(:,k-1),&
        roundoffa,roundoffq,lambda,jacobian,h,MAT,invM,&
        ipivn,g,Jg,newlambda)
   incr=-beta(1)*matmul(transpose(jacobian),newlambda)/den
   pzero=a
   roundoffp(:,k)=roundoffa+incr
   p(:,k)=pzero+roundoffp(:,k)
   roundoffp(:,k)=(pzero-p(:,k))+roundoffp(:,k)
   incr=p(:,k)
   c=q(:,k-1)
   call dgetrs('No␣transpose',n,1,MAT,n,ipivn,incr,n,info)
   roundoffq=incr+roundoffq
   q(:,k)=c+roundoffq
   roundoffq=(c-q(:,k))+roundoffq

   if (iout==1) then
```

```
         newq=q(:,k/2)
         call finitedifferences(n,k,h,p,pdummy)
         call projectionp(n,m,MAT,invM,ipivn,pdummy,newq,Jg,newp)
         call solout(j,newp,newq,n)
      end if

      field(:,k-1)=field(:,k-1)-matmul(transpose(jacobian),&
           newlambda)

      do i=1,k-1
         roundoffpdummy=roundoffp(:,i+1)
         pdummy=p(:,i+1)
         qdummy=q(:,i+1)
         p(:,i)=pdummy
         q(:,i)=qdummy
         roundoffp(:,i)=roundoffpdummy
      end do

      do i=1,k-2
         fieldummy=field(:,i+1)
         field(:,i)=fieldummy
      end do

      call f(q(:,k),evalfdummy)
      call Jg(q(:,k),jacobian)
      field(:,k-1)=h**2*evalfdummy
      lambda=newlambda
   end do

   p0=newp
   q0=newq
   return

 end subroutine lmmconstrained


end module lmmconstrainedmodule
```

Appendix B

Maple Scripts

We used the symbolic manipulation package MAPLE to compute the classes of methods described in Chapter 1 and 3 and all the related relevant quantities described in Chapter 2 and 4.
In this Section all these computations are presented.

- Class of linear multistep methods of order 4 for first order Hamiltonian equations and its constant λ_4:

```
> rho := (x-1)*(x^2+2*a1*x+1)*(x^2+2*a2*x+1);
> sigma := collect(expand(convert(taylor(rho/log(x), x = 1, 6),polynom)
   +c*(x-1)^4*(x+1)), x);
> solve(coeff(%, x^5), c);
> c := %;
> sigma;

  (11/6+(5/6)*a2+(5/6)*a1-(1/6)*a1*a2)*x^4+(1/6+(7/6)*a1+(7/6)*a2
  +(13/6)*a1*a2)*x^3+(1/6+(7/6)*a1+(7/6)*a2+(13/6)*a1*a2)*x^2
  +(11/6+(5/6)*a2+(5/6)*a1-(1/6)*a1*a2)*x

> e := exp(1);
> rho1 := eval(rho, x = e^x);
> sigma1 := eval(sigma, x = e^x);
> res := taylor(rho1/(x*sigma1), x = 0, 5);
> simplify(coeff(res, x, 4));

  (1/720)*(131-19*a1-19*a2+11*a1*a2)/(1+a2+a1+a1*a2)
```

- Class of linear multistep methods of order 4 for first order Hamiltonian equations and its constants λ_6 and λ_8:

```
> rho := (x-1)*(x^2+2*a1*x+1)*(x^2+2*a2*x+1)*(x^2+2*a3*x+1);
> sigma := collect(expand(convert(taylor(rho/log(x), x = 1, 8), polynom)
   +c*(x-1)^6*(x+1)), x);
> solve(coeff(%, x^7), c);
> c := %;
```

```
> sigma;

 (461/180-(19/180)*a1*a3+(11/180)*a1*a2*a3-(19/180)*a1*a2
 -(19/180)*a2*a3+(131/180)*a1+(131/180)*a2+(131/180)*a3)*x^6
 +(-(31/60)*a1*a2*a3+(119/60)*a1*a3+(89/60)*a1-121/60+
 +(119/60)*a1*a2+(119/60)*a2*a3+(89/60)*a2+(89/60)*a3)*x^5
 +((161/90)*a2+(191/90)*a1*a2+(401/90)*a1*a2*a3+311/90
 +(191/90)*a2*a3+(191/90)*a1*a3+(161/90)*a1+(161/90)*a3)*x^4
 +((161/90)*a2+(191/90)*a1*a2+(401/90)*a1*a2*a3+311/90
 +(191/90)*a2*a3+(191/90)*a1*a3+(161/90)*a1+(161/90)*a3)*x^3
 +(-(31/60)*a1*a2*a3+(119/60)*a1*a3+(89/60)*a1-121/60
 +(119/60)*a1*a2+(119/60)*a2*a3+(89/60)*a2+(89/60)*a3)*x^2
 +(461/180-(19/180)*a1*a3+(11/180)*a1*a2*a3-(19/180)*a1*a2
 -(19/180)*a2*a3+(131/180)*a1+(131/180)*a2+(131/180)*a3)*x

> e := exp(1);
> rho1 := eval(rho, x = e^x);
> sigma1 := eval(sigma, x = e^x);
> expansion := taylor(rho1/(sigma1*x), x = 0, 10);
> lambda6 := simplify(coeff(expansion, x^6));

 -(1/60480)*(527*a1+527*a2+527*a3+191*a1*a2*a3-271*a1*a2
 -271*a2*a3-271*a1*a3-4975)/(1+a3+a2+a2*a3+a1+a1*a3+a1*a2
 +a1*a2*a3)

> lambda8 := simplify(coeff(expansion, x^8));

 -(1/172800)*(1/(1+a3+a2+a2*a3+a1+a1*a3+a1*a2+a1*a2*a3)^2
 *(17103+8006*a2+8006*a1+8006*a3-4416*a1*a2*a3+1472*a1*a3
 +1472*a1*a2+1472*a2*a3-314*a1^2*a2*a3^2+352*a1*a2^2*a3
 +352*a1^2*a2*a3+352*a1*a2*a3^2-314*a1*a2^2*a3^2
 -314*a1^2*a2^2*a3+63*a1^2*a2^2*a3^2+186*a2^2*a3+186*a1^2*a2
 +186*a1*a2^2+283*a1^2*a2^2+186*a1^2*a3-1237*a1^2-1237*a2^2
 -1237*a3^2+186*a2*a3^2+186*a1*a3^2+283*a2^2*a3^2
 +283*a1^2*a3^2)
```

- Class of linear multistep methods of order 4 for second order Hamiltonian equations and its error constant:

```
> rho := (x-1)^2*(x^2+2*a1*x+1);
> taylor(rho/log(x)^2, x = 1, 4);
> collect(expand(convert(%, polynom)), x);
> sigma := %;

   (7/6+(1/6)*a1)*x^3+(-1/3+(5/3)*a1)*x^2+(7/6+(1/6)*a1)*x

> e := exp(1);
> rho1 := eval(rho, x = e^x);
```

```
> sigma1 := eval(sigma, x = e^x);
> taylor(rho1-x^2*sigma1, x = 0, 7);
> c := simplify(coeff(%, x^6)/(eval(sigma, x = 1)));

  -(1/240)*(-9+a1)/(1+a1)
```

- Class of linear multistep methods of order 6 for second order Hamiltonian equations and its error constant:

```
> rho := (x-1)^2*(x^2+2*a1*x+1)*(x^2+2*a2*x+1);
> taylor(rho/log(x)^2, x = 1, 6);
> sigma := collect(expand(convert(%, polynom)), x);

  (79/60+(3/20)*a2+(3/20)*a1-(1/60)*a1*a2)*x^5+((2/5)*a1*a2
  +(26/15)*a2+(26/15)*a1-14/15)*x^4+((7/30)*a1+(97/30)*a1*a2
  +97/30+(7/30)*a2)*x^3+((2/5)*a1*a2+(26/15)*a2+(26/15)*a1
  -14/15)*x^2+(79/60+(3/20)*a2+(3/20)*a1-(1/60)*a1*a2)*x

> e := exp(1);
> rho1 := eval(rho, x = e^x);
> sigma1 := eval(sigma, x = e^x);
> taylor(rho1-x^2*sigma1, x = 0, 9);
> C := simplify(coeff(%, x^8)/(eval(sigma, x = 1)));

    (1/60480)*(-95*a1-95*a2+31*a1*a2+1039)/(1+a2+a1+a1*a2)
```

- Class of linear multistep methods of order 8 for second order Hamiltonian equations and its error constant:

```
> rho := (x-1)^2*(x^2+2*a1*x+1)*(x^2+2*a2*x+1)*(x^2
   +2*a3*x+1);
> sigma := collect(expand(convert(taylor(rho/log(x)^2,
    x = 1, 8), polynom)), x);

  (10993/7560+(1039/7560)*a2+(1039/7560)*a1-(19/1512)*a1*a2
  +(1039/7560)*a3-(19/1512)*a2*a3-(19/1512)*a1*a3+(31/7560)
  *a1*a2*a3)*x^7+(-(73/1260)*a1*a2*a3+(473/1260)*a1*a2
  +(473/1260)*a2*a3-443/252+(2279/1260)*a1+(2279/1260)*a2
  +(473/1260)*a1*a3+(2279/1260)*a3)*x^6+((491/2520)*a2
  +16661/2520+(8261/2520)*a2*a3+(2171/2520)*a1*a2*a3
  +(8261/2520)*a1*a3+(491/2520)*a1+(8261/2520)*a1*a2
  +(491/2520)*a3)*x^5+((1357/1890)*a1*a2+(12067/1890)
  *a1*a2*a3+(1357/1890)*a1*a3+(1357/1890)*a2*a3
  +(7027/1890)*a1+(7027/1890)*a2+(7027/1890)*a3
  -8723/1890)*x^4+((491/2520)*a2+16661/2520
  +(8261/2520)*a2*a3+(2171/2520)*a1*a2*a3+(8261/2520)*a1*a3
  +(491/2520)*a1+(8261/2520)*a1*a2+(491/2520)*a3)*x^3
```

```
  +(-(73/1260)*a1*a2*a3+(473/1260)*a1*a2+(473/1260)*a2*a3
  -443/252+(2279/1260)*a1+(2279/1260)*a2+(473/1260)*a1*a3
  +(2279/1260)*a3)*x^2+(10993/7560+(1039/7560)*a2
  +(1039/7560)*a1-(19/1512)*a1*a2+(1039/7560)*a3
  -(19/1512)*a2*a3-(19/1512)*a1*a3+(31/7560)*a1*a2*a3)*x

> e := exp(1);
> rho1 := eval(rho, x = e^x);
> sigma1 := eval(sigma, x = e^x);
> taylor(rho1-x^2*sigma1, x = 0, 11);
> c := simplify(coeff(%, x^10)/(eval(sigma, x = 1)));

  -(1/3628800)*(2209*a1+2209*a2+2209*a3+289*a1*a2*a3
  -641*a1*a2-641*a2*a3-641*a1*a3-28961)/(1+a3+a2+a2*a3
  +a1+a1*a3+a1*a2+a1*a2*a3)
```

Annexe C

Résumé de la thèse

Le sujet de cette thèse est l'étude des méthodes multi-pas linéaires et symétriques appliquées aux systèmes hamiltoniens.

Nous montrons que cette classe de méthodes peut posséde bonnes propriétés de conservation approximative de l'énergie lors de l'intégration des systemes hamiltoniens ; de plus les méthodes multi-pas sont faciles à construire, et elles peuvent avoir un ordre arbitraire. Toute l'analyse présentée dans cette thèse est basée sur l'étude des méthodes multi-pas linéaires symétriques appliquées aux systèmes hamiltoniens dans [HL04].

La thèse est divisée en deux parties.

Dans la première partie (Chapitres 1 et 2), nous étudions les méthodes multi-pas linéaires symétriques partitionnées appliquées aux systèmes hamiltoniens du premier ordre : nous nous concentrons principalement sur les Hamiltoniens séparables. Dans cette étude, nous montrons comment obtenir, pour une classe spécifique de systèmes hamiltoniens séparables, une conservation approximative de l'énergie en utilisant des méthodes multi-pas symétriques partitionnées.

Dans la deuxième partie (Chapitres 3, 4 et 5), nous étudions les méthodes multi-pas symétriques appliquées aux équations de Hamilton du second ordre avec contraintes. Dans les Chapitres 3 et 4 nous nous concentrons sur l'analyse théorique de l'excellent comportement que cette classe de méthodes présente sur ce genre de problemes ; l'analyse est complétée par des considérations pratiques et des expériences numériques. Dans le Chapitre 5, nous étudions l'optimisation de l'implémentation de cette classe de méthodes.

Chapitre 1 Nous étudions les méthodes linéaires multi-pas partitionnées appliquées aux systèmes hamiltoniens.

Cette étude théorique, focalisée sur les systèmes hamiltoniens séparables, se base sur l'analyse rétrograde (*backward error analysis*) et sur la technique des développements de Fourier modulés (*modulated Fourier expansions*). Pour la solution lisse, nous pouvons construire une équation modifiée mais il est impossible de construire une intégrale première qui soit proche de l'Hamiltonien ; nous montrons qu'il est possible d'améliorer le comportement de la solution lisse en imposant des conditions d'ordre qui dépendent des coefficients de la méthode. Puis, nous analysons les composants parasites pour des systèmes hamiltoniens séparable, et nous montrons que sur les intervalles où ces composantes sont petites et bornées, la solution numérique se comporte comme une méthode symétrique à un pas.

Chapitre 2 Nous présentons es compléments à l'analyse faite dans le Chapitre 1.

Nous montrons l'optimisation de la stabilité des méthodes multi-pas symétriques partitionnées qui satisfont les conditions d'ordre supplémentaires décrites dans le Chapitre 1. Ce travail est effectué pour les classes de méthodes d'ordre 4 et 6, et est complété par des expériences numériques pour des systèmes hamiltoniens séparables et non-séparables, où le comportement des composants parasites est également montré.

Chapitre 3 Nous exposons l'étude théorique des méthodes linéaires multi-pas symétriques appliquées aux systèmes hamiltoniens du second ordre avec contraintes.

Pour cette classe de méthodes il est possible d'adapter les techniques de [HL04], en construisant l'équation modifiée, puis l'Hamiltonienne modifiée. L'analyse est complétée en démontrant, sous certaines hypothèses pas (peu ?) restrictives, que les composants parasites restent bornées ; ceci explique l'excellent comportement constaté dans les expériences numériques complétant l'analyse théorique. L'étude se termine par une discussion sur la construction de cette classe de méthodes.

Chapitre 4 Nous proposons des compléments à l'analyse faite dans le Chapitre 3.

Nous étudions l'intervalle de périodicité et la constante de l'erreur pour les classes de méthodes d'ordre 4, 6 et 8 présentées dans le Chapitre 3. Cette étude est complétée par des figures représentant ces quantités.

Chapitre 5 Nous montrons l'optimisation de l'implémentation des méthodes étudiées dans le Chapitre 3.

C'est un point important, car, pour des méthodes d'ordre éléve, il est très facile d'obtenir une erreur de discrétisation de l'ordre de grandeur de la précision de la machine. Si tel est le cas, l'erreur d'arrondi devient la source dominante de l'erreur. Il faut donc optimiser l'erreur d'arrondi afin d'éviter la croissance linéaire due aux erreurs déterministes. Nous présentons les techniques utilisées dans ce but, et nous illustrons à l'aide des figures les effets qu'elles ont sur la propagation des erreurs d'arrondi.

Bibliography

[ACM99] G. Akrivis, M. Crouzeix, and C. Makridakis. Implicit-explicit multistep methods for quasilinear parabolic equations. *Numer. Math.*, 82(4):521–541, 1999.

[AFS97] C. Arévalo, C. Führer, and G. Söderlind. β-blocked multistep methods for Euler-Lagrange DAEs: linear analysis. *Z. Angew. Math. Mech.*, 77(8):609–617, 1997.

[AL76] M. J. Ablowitz and J. F. Ladik. A nonlinear difference scheme and inverse scattering. *Studies in Appl. Math.*, 55:213–229, 1976.

[And83] H. C. Andersen. Rattle: a "velocity" version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.*, 52:24–34, 1983.

[AS95] C. Arévalo and G. Söderlind. Convergence of multistep discretizations of DAEs. *BIT*, 35(2):143–168, 1995.

[CFM06] P. Chartier, E. Faou, and A. Murua. An algebraic approach to invariant preserving integrators: the case of quadratic and Hamiltonian invariants. *Numer. Math.*, 103:575–590, 2006.

[CH13a] P. Console and E. Hairer. *Long-term stability of symmetric partitioned linear multistep methods*, volume 2082 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2013. Lectures from the Summer School held in Cetraro, June 27 - July 2, 2011, Edited by Luca Dieci and Nicola Guglielmi Fondazione C.I.M.E.. [C.I.M.E. Foundation].

[CH13b] P. Console and E. Hairer. Reducing round-off errors in symmetric multistep methods. *submitted for publication*, ??:?–?, 2013.

[Chi09] S. A. Chin. Explicit symplectic integrators for solving nonseparable hamiltonians. *Phys. Rev. E*, 80:037701, Sep 2009.

[CHL13] P. Console, E. Hairer, and C. Lubich. Symmetric multistep methods for constrained Hamiltonian systems. *Numer. Math.*, 124(3):517–539, 2013.

[Dah59] G. Dahlquist. Stability and error bounds in the numerical integration of ordinary differential equations. *Trans. of the Royal Inst. of Techn., Stockholm, Sweden*, 130, 1959.

[Fen95] K. Feng. *Collected Works II*. National Defense Industry Press, Beijing, 1995.

[Hai99] E. Hairer. Backward error analysis for multistep methods. *Numer. Math.*, 84:199–232, 1999.

[Hen62] P. Henrici. *Discrete Variable Methods in Ordinary Differential Equations*. John Wiley & Sons Inc., New York, 1962.

[HH03] E. Hairer and M. Hairer. GniCodes – Matlab programs for geometric numerical integration. In *Frontiers in numerical analysis (Durham, 2002)*, pages 199–240. Springer, Berlin, 2003.

[HL01] E. Hairer and C. Lubich. Long-time energy conservation of numerical methods for oscillatory differential equations. *SIAM J. Numer. Anal.*, 38:414–441, 2001.

[HL04] E. Hairer and C. Lubich. Symmetric multistep methods over long times. *Numer. Math.*, 97:699–723, 2004.

[HLW06] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer Series in Computational Mathematics 31. Springer-Verlag, Berlin, 2nd edition, 2006.

[HMS09] E. Hairer, R.I. McLachlan, and R.D. Skeel. On energy conservation of the simplified Takahashi–Imada method. *M2AN Math. Model. Numer. Anal.*, 43(4):631–644, 2009.

[HNW93] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems*. Springer Series in Computational Mathematics 8. Springer, Berlin, 2nd edition, 1993.

[HS00] E. Hairer and C.M. Schober. Corrigendum to: "Symplectic integrators for the Ablowitz-Ladik discrete nonlinear Schrödinger equation" [Phys. Lett. A 259 (1999), 140–151]. *Phys. Lett. A*, 272(5-6):421–422, 2000.

[Jay96] L. Jay. Symplectic partitioned Runge-Kutta methods for constrained Hamiltonian systems. *SIAM J. Numer. Anal.*, 33:368–387, 1996.

[Kir86] U. Kirchgraber. Multi-step methods are essentially one-step methods. *Numer. Math.*, 48:85–90, 1986.

[LS94] B. J. Leimkuhler and R. D. Skeel. Symplectic numerical integrators in constrained Hamiltonian systems. *J. Comput. Phys.*, 112(1):117–125, 1994.

[LW76] J. D. Lambert and I. A. Watson. Symmetric multistep methods for periodic initial value problems. *J. Inst. Maths. Applics.*, 18:189–202, 1976.

[QT90] G. D. Quinlan and S. Tremaine. Symmetric multistep methods for the numerical integration of planetary orbits. *Astron. J.*, 100:1694–1700, 1990.

[RCB77] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–341, 1977.

[Rei96] S. Reich. Symplectic integration of constrained Hamiltonian systems by composition methods. *SIAM J. Numer. Anal.*, 33:475–491, 1996.

[Sch99] C. M. Schober. Symplectic integrators for the Ablowitz-Ladik discrete nonlinear Schrdinger equation. *Phys. Lett. A*, 259:140–151, 1999.

[Tan93] Y.-F. Tang. The symplecticity of multi-step methods. *Computers Math. Applic.*, 25:83–90, 1993.

[Vil08] G. Vilmart. Reducing round-off errors in rigid body dynamics. *J. Comput. Phys.*, 227(15):7083–7088, 2008.

[VS04] D. S. Vlachos and T. E. Simos. Partitioned linear multistep method for long term integration of the N-body problem. *Appl. Numer. Anal. Comput. Math.*, 1(3):540–546, 2004.