ISBN 978-82-326-4140-6 (printed ver.)
ISBN 978-82-326-4141-3 (electronic ver.)
ISSN 1503-8181

Lu Li

Doctoral thesis

Doctoral theses at NTNU, 2019:272

Lu Li

Energy-preserving numerical methods for differential equations: Linearly implicit methods and Krylov subspace methods

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Doctoral theses at NTNU, 2019:272

NTNU

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

Lu Li

Energy-preserving numerical methods for differential equations: Linearly implicit methods and Krylov subspace methods

Thesis for the Degree of Philosophiae Doctor

Trondheim, October 2019

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

ISBN 978-82-326-4140-6 (printed ver.)
ISBN 978-82-326-4141-3 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2019:272

Printed by NTNU Grafisk senter

Preface

This thesis is submitted in partial fulfilment of the requirements for the degree of Philosophiae doctor (PhD) at the Norwegian University of Science and Technology (NTNU). The work was carried out at NTNU and the research was funded by the Department of Mathematical Sciences at NTNU.

This thesis is the conclusion of nearly four years of research work. It would not have been possible without the help from numerous people. First, I would like to thank my supervisor Professor Elena Celledoni for her encouragement and valuable advice during my PhD study. She is always patient with me and provides me help whenever I need. I am also thankful to my co-supervisor Professor Brynjulf Owren for his interesting discussions, considerable help and insightful suggestions. Next, I thank my co-authors, Shun Sato and Sølve Eidnes for their collaboration, interesting discussions and hard work for fulfilling the work. I would also thank Professor Takayasu Matsuo and my colleague Benjamin Taplay, Torbjørn Ringholm and so on for their insightful suggestions and help. Also, I would thank Professor Einar Rønquist for his kind help and encouragement during my PhD life.

Last but not least, I wish to thank my friends Wuyan Wang, Ling Zhang, Wenting Xu, Wanxia Yang and so on for their help during my stay in Trondheim. In particular, I want to thank my two lovely children Weichen Ola Pei and Xiyue Eva Pei, and my husband Dr. Long Pei for their love, patience and support.

Lu Li

Trondheim

September 16, 2019

Contents

Contents

Introduction

Since Newton and Leibniz invented calculus independently, differential equations have been the most important tool for modeling continuous physical systems and an important area of study for mathematicians. For many differential equations, finding a closed form solution is a difficult or impossible task [39]. This necessitates the study of numerical approximations to differential equations, an area of study called numerical analysis, which has a long history [29]. The first and most simple numerical approach for ordinary differential equations was described by Euler (1768) in his "Institutiones Calculi Integralis". Inspired by the idea of Euler, Runge and later Heun and Kutta tried to extend the Euler method to more advanced schemes which provide higher accuracy and are called Runge–Kutta methods, [9]. In the last few decades, general-purpose numerical methods for ordinary differential equations, including mainly Runge–Kutta methods and linear multistep methods, have been well studied [10]. However, recently much attention has been paid to purpose-designed numerical methods, which are tailored to a class of problems possessing geometric properties, for example the preservation of energy, angular momentum, volume or symplecticity. Numerical methods that can preserve one or more of these properties are called geometric numerical integrators, which are shown to produce not only an improved qualitative behaviour, but also a more accurate long-time solution compared to general-purpose methods [32].

One of the most important classes of differential equations are Hamiltonian systems, which have two well-known geometric properties: the preservation of symplecticity and the preservation of energy [32]. Therefore numerical integrators that are symplectic and that are energy-preserving are of particular interest for Hamiltonian systems. Symplectic methods and energy-preserving methods have their own advantages. However, there is no numerical integrator that can be simultaneously symplectic and energy-preserving for a general Hamiltonian system, except for a time-reparametrization of the exact solution[1] [18, 62]. Symplectic integrators are shown to have a bounded energy error and a lin-

[1] In addition the system is assumed to have no other conserved quantities than the Hamiltonian and functions of the Hamiltonian [62].

ear global error growth for integrable systems over long-time integration [32]. Examples of such types of integrators can be found in [37, 56]. In this thesis, we focus on energy-preserving methods. In the continuous time setting, the concept of energy and its conservation has a far-reaching importance throughout the physical sciences [26]; one example is that the exact preservation of energy plays an important role in the study of orbital stability of soliton solutions to certain Hamiltonian partial differential equations [3]. Even from a numerical point of view, the energy-preserving property is found to be crucial in the proof of stability and convergence for some numerical methods, see for example [24, 41]. Some examples of energy-preserving methods are discrete gradient methods [48, 49], the average vector field method [54], and Hamiltonian boundary value methods [7, 8].

Most of the existing energy-preserving integrators for Hamiltonian systems are implicit [12, 61], and so a nonlinear system has to be solved at each time step. This might lead to either a high computational cost or to a loss of the conservation property due to the non-negligible truncation error in the numerical solution of the nonlinear system [21]. An alternative idea is to build a linearly implicit method, for which one linear system is needed to be solved at each time step [21, 46]. In this thesis, we focus on solving Hamiltonian partial differential equations (PDEs) and Hamiltonian ordinary differential equations (ODEs) with numerical integrators that are energy-preserving and linearly implicit. For ODEs it is common to develop rather general structure-preserving frameworks. However it is somewhat different from the usual practice with PDEs where each considered equation normally requires a dedicated scheme [21]. Nevertheless there exist certain fairly general methodologies that can be used for developing structure-preserving methods for PDEs, and we consider two of them in this thesis. The first one is to discretize the Hamiltonian PDE in space and make sure to obtain a Hamiltonian ODE to which geometric numerical integrators are applied, see for example [12]. The other approach is to reformulate the PDE into a multisymplectic form and then apply a scheme that preserves the conservation laws inherent in the multisymplectic structure [44].

The semi-discretization of Hamiltonian PDEs may lead to large and sparse linear Hamiltonian ODEs, for example the discretization of the wave equations [25] and the Maxwell equations [45]. Moreover, large and sparse linear Hamiltonian systems have also been widely considered in engineering, like the models in network dynamics [58]. Another topic of the thesis is to consider energy-preserving methods for such systems. Krylov subspace methods have been extensively used to solve eigenvalue problems [60], linear systems [55] and linear differential equations [34]. However, when applied to Hamiltonian ODE systems, the standard Krylov subspace methods, e.g. the Arnoldi projection method [17], will in general fail to preserve geometric properties, such as

energy and symplecticity, see for example [13]. In this thesis, we focus on constructing the energy-preserving Krylov subspace methods, using a symplectic basis [23, 43]. Besides such methods, we also consider modifying the classical Arnoldi projection method to construct an energy-preserving method.

In the following we will briefly review the basic concepts used in the papers that constitute the main contribution of this thesis, and give a summary of each paper.

1.1 Discrete gradient methods and linearly implicit schemes

We consider an initial value problem

$$\dot{y} = f(y), \quad y(t_0) = y_0, \tag{1.1.1}$$

where $y(t) \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}^n$. A differentiable function $\mathcal{I}(y)$ is said to be a first integral of the ODE (1.1.1) if

$$\frac{d\mathcal{I}(y)}{dt} |_{y=y(t)} = \nabla\mathcal{I}(y)\dot{y} = \nabla\mathcal{I}(y)f(y) = 0, \quad \forall y \in \mathbb{R}^n.$$

Then $\mathcal{I}(y)$ is a conserved quantity along the flow of equation (1.1.1): $\mathcal{I}(y(t)) - \mathcal{I}(y(t_0)) = \int_{t_0}^{t} \nabla\mathcal{I}(y)\dot{y}dt = 0$. It is shown in [48] that an ODE system with a first integral $\mathcal{I}(y)$ can generally be written into the form

$$\dot{y} = S(y)\nabla\mathcal{I}(y),$$

where $S(y)$ is a skew-symmetric matrix, under some mild assumptions. The most well-known example of a first integral is the energy of Hamiltonian ODEs which arise in many areas of physics [38]. Henceforth, we make no distinction between integral-preserving integrators and energy-preserving integrators. Let us restrict to systems (1.1.1) of the form

$$\dot{y} = S\nabla H(y), \quad y(t_0) = y_0, \tag{1.1.2}$$

where S is a constant skew-symmetric matrix and $H(y)$ is a scalar function. A popular class of methods to solve systems of the form (1.1.2) are the discrete gradient methods. The basic idea is to introduce a map $\overline{\nabla}H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ called the discrete gradient of $H(y)$, which is an approximation to $\nabla H(y)$ and

satisfies

$$H(y) - H(x) = \overline{\nabla} H(x, y)^T (y - x),$$
$$\overline{\nabla} H(x, x) = \nabla H(x).$$

Given the discrete gradient $\overline{\nabla} H$, the discrete gradient method for (1.1.2) is defined by

$$\frac{y^{k+1} - y^k}{\Delta t} = S \overline{\nabla} H(y^k, y^{k+1}). \tag{1.1.3}$$

The above scheme (1.1.3) preserves the energy of (1.1.2) since

$$\begin{aligned} H(y^{k+1}) - H(y^k) =& \overline{\nabla} H(y^k, y^{k+1})^T (y^{k+1} - y^k) \\ &= \Delta t \overline{\nabla} H(y^k, y^{k+1})^T S^T \overline{\nabla} H(y^k, y^{k+1}) \\ &= 0. \end{aligned}$$

There are a number of ways to create a discrete gradient and the most used examples are

- the average vector field (AVF) discrete gradient [33]

$$\overline{\nabla} H_{\mathrm{AVF}}(x, y) = \int_0^1 \nabla H(\xi x + (1 - \xi) y) d\xi,$$

- the midpoint (MP) or Gonzalez discrete gradient [31]

$$\overline{\nabla} H_{\mathrm{MP}}(x, y) = \nabla H(\frac{x + y}{2}) + \frac{H(y) - H(x) - \nabla H(\frac{x+y}{2})^T (y - x)}{||y - x||_2} (y - x),$$

- the Itoh–Abe (IA) discrete gradient [35]

$$\overline{\nabla} H_{\mathrm{IA}}(x, y)_l = \begin{cases} \frac{H(\sum_{i=1}^{l} y_i e_i + \sum_{i=l+1}^{n} x_i e_i) - H(\sum_{i=1}^{l-1} y_i e_i + \sum_{i=l}^{n} x_i e_i)}{y_l - x_l}, & \text{if } x_l \neq y_l, \\ \frac{\partial H}{\partial x_l}(\sum_{i=1}^{l-1} y_i e_i + \sum_{i=l}^{n} x_i e_i) & \text{if } x_l = y_l, \end{cases}$$

where e_l denotes the l_{th} Euclidean unit vector field. The AVF and MP discrete gradients are symmetric with respect to x and y, leading to symmetric energy-preserving methods of second order. The Itoh–Abe discrete gradient is not symmetric, however, it can be symmetrized by $\overline{\nabla} H_{\mathrm{SIA}}(x, y) := \frac{1}{2}(\overline{\nabla} H_{\mathrm{IA}}(x, y) + \overline{\nabla} H_{\mathrm{IA}}(y, x))$.

Discrete gradient methods were systematically studied for ODEs in [31,48]. The idea was applied to solve PDEs in [49], in [7, 12] where the AVF method was considered, and in [27] where the discrete variational derivative method

was considered. These methods are normally fully implicit, and one drawback for such kind of methods is that one has to solve a non-linear system at each time step. To avoid this drawback, linearly implicit energy-preserving methods may be considered. One technique to generate linearly implicit methods was proposed by introducing the concept of "multiple points discrete variational derivative", see [46]. Following the idea there, a general framework for constructing linearly implicit methods which allow for an arbitrary number of variables with derivatives of any order, was presented in [21]. In this thesis, we consider a numerical comparison of two linearly implicit energy-preserving methods for Hamiltonian PDEs: one method is achieved by applying the technique introduced in [21, 46] to the semi-discrete Hamiltonian ODE systems, and the other is obtained by applying Kahan's method to the semi-discrete systems, see next section for a definition. In [21], Dahlby and Owren introduced a notion of polarised Hamiltonian and polarised discrete variational derivative (PDVD) for introducing the linearly implicit schemes, see also [46] for related work. In the following, we consider a simplified version of the definitions in [21] adapted to cubic Hamiltonian functions.

Definition 1.1. For a cubic polynomial energy function $H:\mathbb{R}^n \to \mathbb{R}$, consider the polarised energy function $\tilde{H}:\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ satisfying the properties

$$\tilde{H}(x,x) = H(x), \quad \tilde{H}(x,y) = \tilde{H}(y,x),$$

then the polarised discrete gradient (PDG) for $\tilde{H}$ is defined by $\overline{\nabla}\tilde{H}:\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ satisfying

$$\begin{aligned} \tilde{H}(y,z) - \tilde{H}(x,y) &= \frac{1}{2}(z-x)^T \overline{\nabla}\tilde{H}(x,y,z), \\ \overline{\nabla}\tilde{H}(x,x,x) &= \nabla H(x). \end{aligned}$$

The corresponding polarised discrete gradient scheme for system (1.1.2) is given by

$$\frac{y^{k+2}-y^k}{2\Delta t} = S\overline{\nabla}\tilde{H}(y^k, y^{k+1}, y^{k+2}). \tag{1.1.4}$$

There exist many ways to build a PDG and one is based on the work by Matsuo and coauthors [46], where plenty of examples are considered for polynomial functions. Two more examples are the generalizations of the AVF discrete gradient and IA discrete gradient:

- the polarised discrete gradient based on AVF [21]

$$\overline{\nabla}_{\mathrm{AVF}}\tilde{H}(x,y,z) = 2\int_0^1 \nabla_x \tilde{H}(\xi x + (1-\xi)z, y)\,\mathrm{d}\xi,$$

where $\nabla_x \tilde{H}(x,y)$ is the partial derivative of $\tilde{H}(x,y)$ with respect to its first argument;

- the polarised discrete gradient based on Itoh–Abe (IA) discrete gradient [22]

$$\overline{\nabla}_{\mathrm{IA}} \tilde{H}(x,y,z)_l = 2 \begin{cases} \bar{\partial} \tilde{H}(x,y,z)_l & \text{if } x_l \neq z_l, \\ \frac{\partial \tilde{H}}{\partial x_l}(\sum_{i=1}^{l-1} z_i e_i + \sum_{i=l}^{n} x_i e_i, y)) & \text{if } x_l = z_l, \end{cases}$$

where

$$\bar{\partial} \tilde{H}(x,y,z)_l = \frac{\tilde{H}(\sum_{i=1}^{l} z_i e_i + \sum_{i=l+1}^{n} x_i e_i, y) - \tilde{H}(\sum_{i=1}^{l-1} z_i e_i + \sum_{i=l}^{n} x_i e_i, y)}{z_l - x_l}.$$

1.2 Kahan's method

To introduce Kahan's method, we start with the problem in the form

$$\dot{y} = f(y) = A(y) + By + c, \tag{1.2.1}$$

where $A(y)$ is an $\mathbb{R}^n$-valued quadratic form, $B \in \mathbb{R}^{n \times n}$ is a symmetric constant matrix, c is a constant vector. A system of the type (1.2.1) looks quite simple and restrictive, however it appears often in applications, for example air pollution models [63] and ordinary differential equations that arise after semi-discretisation of a PDE, like the Korteweg–de Vries equation [36] or the Camassa–Holm equation [22]. For the above problem (1.2.1), consider the following discretization

$$\frac{y^{k+1} - y^k}{\Delta t} = A(y^k, y^{k+1}) + B\frac{y^k + y^{k+1}}{2} + c, \tag{1.2.2}$$

where

$$A(y^k, y^{k+1}) = \frac{1}{2}(A(y^k, y^{k+1}) - A(y^{k+1}) - A(y^k)),$$

is a symmetric bilinear form which is obtained by the polarization of the quadratic form A, [16]. The scheme (1.2.2) is symmetric and linearly implicit. We will call this scheme Kahan's method as in [16]. Kahan's method can be shown to coincide with the following Runge–Kutta method

$$\frac{y^{k+1} - y^k}{\Delta t} = -\frac{1}{2} f(y^k) + 2 f(\frac{y^k + y^{k+1}}{2}) - \frac{1}{2} f(y^{k+1}). \tag{1.2.3}$$

when applied to quadratic vector fields, see [16]. Using the Runge-Kutta form (1.2.3), it is shown in [16] that Kahan's method applied to a Hamiltonian ODE with quadratic vector field has a conserved modified Hamiltonian and an invariant measure, a combination previously unknown amongst Runge–Kutta methods applied to nonlinear vector fields. Inspired by this property, large classes of integrable rational mappings are found, examples including [14, 53]. Kahan's method was generalized to higher degree polynomial vector fields and the discretization was shown to preserve modified versions of the measure and energy when Hamiltonian vector fields are considered in [15].

Suppose we restrict the problem (1.2.1) to be a Hamiltonian system on either a symplectic vector space or a Poisson vector space with constant Poisson structure:

$$\dot{y} = S\nabla H(y), \tag{1.2.4}$$

where $S \in \mathbb{R}^{n\times n}$ is a constant skew-symmetric matrix and $H : \mathbb{R}^n \to \mathbb{R}$ is a cubic polynomial Hamiltonian function. We first consider the Hamiltonian H to be homogeneous, and according to [15], Kahan's method applied to (1.2.4) can be rewritten by

$$\frac{y^{k+1} - y^k}{\Delta t} = \frac{1}{2} S\bar{H}(y^k, y^{k+1}, \cdot), \tag{1.2.5}$$

where $\bar{H}(\cdot,\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a symmetric 3-tensor and satisfies $\bar{H}(x,x,x) = H(x)$. We now consider the 3-tensor[2] $\bar{H}(x,y,z) = x^T Q(y) z$, where $Q(y) = \frac{1}{6}H''(y)$ with $H''(y)$ the Hessian of H, and we can rewrite Kahan's method (1.2.5) as

$$\frac{y^{k+1} - y^k}{\Delta t} = 3S\frac{\partial \bar{H}(x,y,z)}{\partial x}|_{(y^k, y^{k+1})}, \tag{1.2.6}$$

which we will use in Paper 2 in this thesis.

Consider a nonhomogeneous cubic Hamiltonian

$$H(y) = y^T Q(y) y + y^T B y + c^T y + d,$$

where $y = [y_1, \ldots, y_n]^T$, $Q(y)$ is an $n \times n$ symmetric matrix whose elements are homogeneous linear polynomials, B is an $n \times n$ symmetric matrix, $c \in \mathbb{R}^n$ is a

[2] Denote the elements in $Q(y)$ by $q_{ij}y = \sum_k q_{ij}^k y_k$, where q_{ij}^k, $i,j,k = 1,\cdots,n$, are scalars and y_k is the kth element of y. We observe that q_{ij}^k satisfies $q_{ij}^k = q_{ki}^j = q_{jk}^i$ since $q_{ij}^k = \frac{1}{6}\frac{\partial^3 H(y)}{\partial y_i \partial y_j \partial y_k}$, which is unchanged under any permutation of i,j,k. This provides the symmetry of the three tensor $\bar{H}(x,y,z)$.

vector and d is a number, with B, c, and d constant. We follow the method in [16]: adding one variable $\tilde{y} = [y_0, y_1, \ldots, y_n]^T$, extending S to $\tilde{S}$ by adding a zero initial row and column, extending the nonhomogeneous Hamiltonian $H(y)$ to a homogeneous function $\tilde{H}(\tilde{y})$ so that $\tilde{H}(\tilde{y})|_{y_0=1} = H(y)$, and finally solving instead a Hamiltonian system with homogeneous cubic Hamiltonian problem $\dot{\tilde{y}} = \tilde{S}\nabla\tilde{H}(\tilde{y})$ with $y_0 = 1$. In such way we can also get the reformulation of Kahan's method as (1.2.6) with

$$H(x, y, z) = x^T Q(y)z + \frac{1}{3}(x^T By + y^T Bz + z^T Bx) + \frac{1}{3}c^T(x + y + z) + d.$$

1.3 Multisymplectic Hamiltonian PDEs

The classical concept of a finite-dimensional Hamiltonian system has been generalized to an infinite-dimensional form for PDEs, which results in the Hamiltonian PDEs of the form

$$\frac{\partial u}{\partial t} = \mathcal{D}\frac{\delta \mathcal{H}}{\delta u}, \tag{1.3.1}$$

where $\mathcal{D}$ is a skew-adjoint differential operator with constant coefficients, $\mathcal{H}$ is an energy function and $\frac{\delta \mathcal{H}}{\delta u}$ is the variational derivative of $\mathcal{H}$ [51]. In a finite-dimensional system, the Hamiltonian formulation is obtained by applying a Legendre transform to the Lagrangian equation. When it comes to an infinite-dimensional system, taking Klein–Gordon equation as an example; considering the variational formulation with Lagrangian density $L(u, u_t, u_x)$, in [6] it is shown that the Hamiltonian formulation (1.3.1) is obtained by a partial Legendre transform using a new variable $v = \frac{\partial L(u,u_t,u_x)}{\partial u_t}$. A complete Legendre transform introduces also a variable $w = \frac{\partial L(u,u_t,u_x)}{\partial u_x}$ in addition to v, leading to the multi-symplectic formulation [6],

$$Mz_t + Kz_x = \nabla_z S(z), \quad z \in \mathbb{R}^n, \quad (x, t) \in \mathbb{R}^2, \tag{1.3.2}$$

where n depends on the number of variables needed to put the differential equations into multisymplectic form (it is 3 for Klein–Gordon equation), M and K are two $n \times n$ constant skew-symmetric matrices and $S : \mathbb{R}^n \to \mathbb{R}$ is a scalar-valued function [6]. The finite-dimensional Hamiltonian formulation (1.1.2) treats time as a preferred direction compared with space and it is most useful when the spatial domain is finite. On the other hand the multisymplectic formulation (1.3.2) puts space and time on an equal footing, and it is a natural framework for analysing and proving particular properties of dispersive wave

propagation in conservative systems [6], one example is the rigorous proof of the instability of periodic travelling waves that is predicted by the Whitham modulation equation [4].

Following the analysis in [5, 6], it can be shown that multisymplectic PDEs have the following local conservation laws [38, 50]:

- the multisymplectic conservation law

$$\partial_t \omega + \partial_x \kappa = 0,$$

  where $\omega = dz \wedge M_+ dz$, $\kappa = dz \wedge K_+ dz$, with $\wedge$ the exterior product of two differential forms, and M_+ and K_+ are splittings of M and K satisfying $M = M_+ - M_+^T$, $K = K_+ - K_+^T$;

- the local energy conservation law

$$E_t + F_x = 0,$$

  where $E(z) = S(z) + z_x^T K_+ z$ is the energy density and $F(z) = -z_t^T K_+ z$ is the energy flux;

- the local momentum conservation law

$$I_t + G_x = 0,$$

  with $I(z) = -z_x^T M_+ z$ the momentum density and $G(z) = S(z) + z_t^T M_+ z$ the momentum flux.

Example The Klein–Gordon equation [6]

$$u_{tt} - u_{xx} = V'(u), \quad x \in \mathbb{R}, \quad t > 0, \tag{1.3.3}$$

where $V(u)$ is a smooth nonlinear function of u. The Lagrangian functional for (1.3.3) is

$$\mathcal{L} = \int\int L(u, u_t, u_x) dx dt \quad \text{with} \quad L(u, u_t, u_x) = \frac{1}{2} u_t^2 - \frac{1}{2} u_x^2 + V(u).$$

Taking the change of variables $v = u_t$ and $w = u_x$, equation (1.3.3) has the multisymplectic form (1.3.2) with $S(z) = \frac{w^2 - v^2}{2} + V(u)$,

$$M = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad K = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

The multisymplectic form (1.3.2) can be generalized to problems in a high-dimensional spatial domain. Consider a Hamiltonian PDE on a d-dimensional domain, according to [6], the corresponding multisymplectic form can be written as

$$Mz_t + \sum_{\alpha=1}^{d} K^\alpha z_{x_\alpha} = \nabla_z S(z), \quad z \in \mathbb{R}^n, \quad (x,t) \in \mathbb{R}^d \times \mathbb{R}, \tag{1.3.4}$$

where M and $K^\alpha, \alpha = 1, \cdots, d$ are constant $n \times n$ skew-symmetric matrices and $S : \mathbb{R}^n \to \mathbb{R}$ is a smooth functional. Equation (1.3.4) has the following local energy conservation law

$$E_t + \sum_{\alpha=1}^{d} F^\alpha_{x_\alpha} = 0,$$

where $E(z) = S(z) + \sum_{\alpha=1}^{d} z^T_{x_\alpha} K^\alpha_+ z$, $F^\alpha = -z_t^T K^\alpha_+ z$, and K^α_+ are splittings of K^α satisfying $K^\alpha = K^\alpha_+ - K^{\alpha\,T}_+$.

The multisymplectic structure motivates the study of geometric numerical integration for the Hamiltonian PDEs from a new perspective. There is numerical evidence that the methods, which preserve discrete approximations to at least one of the above local conservation laws, perform well when applied to multistymplectic PDEs. Examples include multisymplectic integrators [2, 19, 47, 57], and integrators preserving energy conservation laws, [11, 20, 30, 42].

1.4 Krylov subspace methods and linear Hamiltonian systems

Krylov subspace methods have been widely used for solving large problems. A Krylov subspace of dimension r based on a given matrix $A \in \mathbb{R}^n \times \mathbb{R}^n$ and vector $b \in \mathbb{R}^n$ is defined by

$$\mathcal{K}_r(A, b) := \text{span}\{b, Ab, \cdots, A^{r-1}b\}. \tag{1.4.1}$$

The columns of Krylov subspace $\mathcal{K}_r$ are normally linearly independent for small values r. However, in practice, orthonormal basis have computational advantages. One well-known technique to generate an orthonormal basis is using the Arnoldi algorithm [1].

Methods that use Krylov subspace are called Krylov subspace methods or Krylov projection methods and have a long history [59]. Such methods are mostly used in solving linear systems [55], eigenvalue problems [52] and

computing matrix exponentials [28]. In this thesis, we focus on using the Krylov projection method to solve linear Hamiltonian ODEs and explore the geometric properties behind it. Given a linear Hamiltonian ODE

$$\dot{y} = f(y) = Ay = JHy, \quad y(t_0) = y_0, \qquad J = J_{2m} = \begin{bmatrix} 0 & I_m \\ -I_m & 0 \end{bmatrix}, \tag{1.4.2}$$

where $y(t) \in \mathbb{R}^{2m}$, $H \in \mathbb{R}^{2m\times 2m}$ is symmetric, $y_0 \in \mathbb{R}^{2m}$, and I_m is the $m \times m$ identity matrix. The idea of Krylov projection methods is to build numerical approximations for (1.4.2) in the Krylov subspace $\mathcal{K}_r$ of dimension $r << 2m$. Let us consider even dimension $r = 2n$. The Krylov projection method based on Arnoldi algorithm gives a $2m \times 2n$ matrix V_{2n} with orthonormal columns, and an upper Hessenberg $2n \times 2n$ matrix T_{2n} such that

$$I_{2n} = {V_{2n}}^T V_{2n}, \quad T_{2n} = V_{2n}^T A V_{2n}.$$

The approximation of $y(t)$ is

$$y_A(t) := V_{2n} z(t), \quad \text{where} \quad \dot{z} = T_{2n} z, \quad z(0) = z_0 = V_{2n}^T y_0.$$

The Krylov projection method based on the Arnoldi algorithm for solving general ODE systems has been studied in [17, 34]. However, this method fails in general to preserve energy for Hamiltonian systems [13], except for the case when A is skew-symmetric and the case when

$$A = \begin{bmatrix} 0 & I \\ -H_{11} & 0 \end{bmatrix}, \qquad \text{and} \quad y_0 = (0, p_0^T)^T.$$

It is shown in [40] that the Arnoldi method preserves a certain number of first integrals and has a bounded energy error over a long integration time in the former case, and the latter case ensures the Arnoldi method to be energy-preserving. Consider a symmetric and positive definite matrix H; one way to improve the behaviour of the Arnoldi algorithm applied to the general Hamiltonian system is to modify the classical Arnoldi algorithm by replacing the Euclidean inner product by a new inner product $\langle \cdot, \cdot \rangle_H := \langle \cdot, H \cdot \rangle$. It is shown in [40] that the modified Arnoldi projection method turns out to be energy-preserving. The modified Arnoldi projection method leads to a $2m \times 2n$ matrix V_{2n} with orthonormal columns, and an upper Hessenberg $2n \times 2n$ matrix T_{2n} such that

$$I_{2n} = {V_{2n}}^T H V_{2n}, \quad T_{2n} = V_{2n}^T H A V_{2n}.$$

The approximation of $y(t)$ is

$$y_A(t) := V_{2n} z(t), \quad \text{where} \quad \dot{z} = T_{2n} z, \quad z(0) = z_0 = V_{2n}^T H y_0.$$

Another preferable alternative could be to build a symplectic basis, for which the projected system inherits the structure of a Hamiltonian system. The Krylov projection methods using a symplectic basis give rise to a $2m \times 2n$ matrix S_{2n} with symplectic columns, and an upper Hessenberg $2n \times 2n$ matrix T_{2n} such that

$$J_{2n} = {S_{2n}}^T J_{2m} S_{2n}, \quad T_{2n} = J_{2n} S_{2n}^T H S_{2n}. \tag{1.4.3}$$

The approximation of $y(t)$ is

$$y_A(t) := S_{2n} z(t), \quad \text{where} \quad \dot{z} = T_{2n}\, z, \quad z(0) = z_0 = J_{2n}^{-1} S_{2n}^T J y_0. \tag{1.4.4}$$

If the matrix S_{2n} in the equation (1.4.3) does not depend on y_0 and a symplectic map is used to solve the projected system (1.4.4), then the projection method (1.4.4) is symplectic. However, concerning the efficiency of the projection method itself, S_{2n} normally depends on y_0. Nevertheless this projection method preserves the energy and gives good numerical behaviour over long time [23, 40].

1.5 Summary of papers

This thesis consists of four papers and one appendix. Paper 1 and 2 concern linearly implicit energy-preserving numerical methods for Hamiltonian ODEs and PDEs. In papers 3 and 4, we consider the energy preservation property of Krylov subspace methods for linear Hamiltonian systems. The appendix considers the analysis of the propagation of the roundoff errors in the energy preservation of one method considered in Paper 4. To fit the thesis format, we reformulate the layout and typography of the published papers.

Paper 1: Linearly implicit structure-preserving schemes for Hamiltonian systems

Sølve Eidnes, Lu Li and Shun Sato
Submitted

In this paper, we consider a comparative study of two linearly implicit energy-preserving methods for Hamiltonian PDEs with cubic Hamiltonian. One method is based on using a two-step generalization of the discrete gradient [46] and is called a polarised discrete gradient (PDG) method. The other is based on Kahan's method. We present a general class of two-step methods encompassing both methods considered in the paper. We apply these two methods to the Hamiltonian ODEs from the semi-discretization of Hamiltonian PDEs, such as the KdV equation and the Camassa–Holm equation. A number

of numerical experiments have been performed here, for the KdV equation with one and two solitons, and the Camassa–Holm equation with one and two peakons. The experiments show that Kahan's method is more stable; it allows for a larger time step-size than the PDG method when the same spatial step-size is considered. Kahan's method also yields more accurate results, as we have observed in the energy error and the shape and phase error with respect to the analytical solutions.

Paper 2: Linearly implicit local and global energy-preserving methods for Hamiltonian PDEs

Sølve Eidnes and Lu Li
Submitted

In this paper, we present a new linearly implicit local energy-preserving algorithm and a class of linearly implicit global energy-preserving methods for multisymplectic PDEs, of the form $Mz_t + Kz_x = \nabla_z S(z)$ with the scalar function S a cubic polynomial. The construction of linearly implicit local energy-preserving method follows two steps. First we apply the midpoint scheme in space to get a semi-discrete system. Then we use Kahan's method for the semi-discrete system to get the full discretization. The linearly implicit global energy-preserving method is created by semi-discretizing the spatial operator ∂_x with a skew-symmetric differentiation matrix and then applying Kahan's method to the semi-discrete system to get the full discretization. We prove that the new local energy-preserving method has a discrete local energy conservation law, from which a global preservation of discrete energy can be deduced. We also show that the global energy-preserving methods preserve a discrete global energy conservation law. In addition, all the results can be generalized for multisymplectic forms in a high-dimensional spatial domain. We test our methods on Hamiltonian PDEs, such as the KdV equation and Zakharov–Kuznetsov equation. The proposed methods give good approximations to the exact wave profiles for both systems over long integration times and they are comparable to the implicit energy-preserving methods in [30], however with much less computational cost.

Paper 3: Symplectic Lanczos and Arnoldi Method for Solving Linear Hamiltonian Systems: Preservation of Energy and Other Invariants

Elena Celledoni and Lu Li
Published in: *Progress in Industrial Mathematics at ECMI 2016*

In this paper we report several numerical experiments for different Krylov subspace methods applied to linear Hamiltonian systems: the projection method based on the classical Arnoldi algorithm; the projection method based on the

symplectic Lanczos algorithm. The Arnoldi projection method, which computes an orthonormal basis of the Krylov subspace, fails in general to preserve the energy or symplecticity of the exact solution under numerical discretization. However, we find that for some special Hamiltonian systems the Arnoldi projection method preserves the energy and even some other invariants, for example the case when the Hamiltonian matrix is skew-symmetric. The Symplectic Lanczos projection method constructed by using a J-orthogonal basis of the Krylov subspace is shown to be energy-preserving with a good numerical behaviour for long-time integration.

Paper 4: Krylov projection methods for linear Hamiltonian systems

Elena Celledoni and Lu Li
Published in: *Numerical Algorithms*

Krylov subspace methods are popular for the approximation of solutions of large and sparse linear systems of ordinary differential equations. One well known technique is based on the method of Arnoldi which computes an orthonormal basis of the Krylov subspace. However, when applied to Hamiltonian linear systems of ODEs, this method fails in general to preserve the symplecticity or energy. In this work, we show that the Arnoldi projection method preserves the energy and even a number of invariants when the Hamiltonian vector field has a special structure. Moreover, we modify the classical Arnoldi algorithm by using a new inner product and get a new energy-preserving method that is shown to preserve several other invariants. We also consider methods based on the use of the Symplectic Lanczos algorithm, and on model reduction techniques and other new strategies, like combining the Arnoldi algorithm and QR factorization to construct a J-orthogonal basis of the Krylov subspace. We test our methods on randomly generated Hamiltonian matrices and Hamiltonian systems corresponding to semi-discretization of Hamiltonian PDEs.

Bibliography

[1] W. E. ARNOLDI, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] U. M. ASCHER AND R. I. MCLACHLAN, *Multisymplectic box schemes and the Korteweg-de Vries equation*, Appl. Numer. Math., 48 (2004), pp. 255–269. Workshop on Innovative Time Integrators for PDEs.

[3] T. B. BENJAMIN, *The stability of solitary waves*, Proc. Roy. Soc. (London) Ser. A, 328 (1972), pp. 153–183.

[4] T. J. BRIDGES, *Errata: "Periodic patterns, linear instability, symplectic structure and mean-flow dynamics for three-dimensional surface waves"*, Philos. Trans. Roy. Soc. London Ser. A, 354 (1996), pp. 2591–2592.

[5] T. J. BRIDGES, *A geometric formulation of the conservation of wave action and its implications for signature and the classification of instabilities*, Proc. Roy. Soc. London Ser. A, 453 (1997), pp. 1365–1395.

[6] T. J. BRIDGES, *Multi-symplectic structures and wave propagation*, Math. Proc. Cambridge Philos. Soc., 121 (1997), pp. 147–190.

[7] L. BRUGNANO, G. FRASCA-CACCIA, AND F. IAVERNARO, *Line integral solution of Hamiltonian PDEs*, 2019.

[8] L. BRUGNANO, F. IAVERNARO, AND D. TRIGIANTE, *Hamiltonian boundary value methods (energy preserving discrete line integral methods)*, J. Numer. Anal. Ind. Appl. Math, 5 (2010), pp. 17–37.

[9] J. C. BUTCHER, *A history of Runge-Kutta methods*, Appl. Numer. Math., 20 (1996), pp. 247–260.

[10] J. C. BUTCHER, *Numerical methods for ordinary differential equations in the 20th century*, J. Comput. Appl. Math., 125 (2000), pp. 1–29. Numerical analysis 2000, Vol. VI, Ordinary differential equations and integral equations.

[11] J. CAI, Y. WANG, AND C. JIANG, *Local structure-preserving algorithms for general multi-symplectic Hamiltonian PDEs*, Comput. Phys. Commun., 235 (2019), pp. 210–220.

[12] E. CELLEDONI, V. GRIMM, R. I. MCLACHLAN, D. I. MCLAREN, D. O'NEALE, B. OWREN, AND G. R. W. QUISPEL, *Preserving energy resp. dissipation in numerical PDEs using the "Average Vector Field" method*, J. Comput. Phys., 231 (2012), pp. 6770–6789.

[13] E. CELLEDONI AND L. LI, *Symplectic Lanczos Arnoldi method for solving linear Hamiltonian systems: preservation of energy and other invariants*, in Progress in industrial mathematics at ECMI 2016, vol. 26 of Math. Ind., Springer, Cham, 2017, pp. 553–559.

[14] E. CELLEDONI, R. I. MCLACHLAN, D. I. MCLAREN, B. OWREN, AND G. R. W. QUISPEL, *Integrability properties of Kahan's method*, J. Phys. A, 47 (2014), pp. 365202, 20.

[15] E. CELLEDONI, R. I. MCLACHLAN, D. I. MCLAREN, B. OWREN, AND G. R. W. QUISPEL, *Discretization of polynomial vector fields by polarization*, Proc. A., 471 (2015), pp. 20150390, 10.

[16] E. CELLEDONI, R. I. MCLACHLAN, B. OWREN, AND G. R. W. QUISPEL, *Geometric properties of Kahan's method*, J. Phys. A, 46 (2013), pp. 025201, 12.

[17] E. CELLEDONI AND I. MORET, *A Krylov projection method for systems of ODEs*, Appl. Numer. Math., 24 (1997), pp. 365–378.

[18] P. CHARTIER, E. FAOU, AND A. MURUA, *An algebraic approach to invariant preserving integrators: the case of quadratic and Hamiltonian invariants*, Numer. Math., 103 (2006), pp. 575–590.

[19] J.-B. CHEN AND M.-Z. QIN, *A multisymplectic variational integrator for the nonlinear Schrödinger equation*, Numer. Methods Partial Differential Equations, 18 (2002), pp. 523–536.

[20] Y. CHEN, Y. SUN, AND Y. TANG, *Energy-preserving numerical methods for Landau-Lifshitz equation*, J. Phys. A, 44 (2011), pp. 295207, 16.

[21] M. DAHLBY AND B. OWREN, *A general framework for deriving integral preserving numerical methods for PDEs*, SIAM J. Sci. Comput., 33 (2011), pp. 2318–2340.

[22] S. EIDNES, L. LI, AND S. SATO, *Linearly implicit structure-preserving schemes for Hamiltonian systems*, arXiv preprint arXiv:1901.03573, (2019).

[23] T. EIROLA AND A. KOSKELA, *Krylov integrators for Hamiltonian systems*, BIT, 59 (2019), pp. 57–76.

[24] Z. FEI, V. M. PÉREZ-GARCÍA, AND L. VÁZQUEZ, *Numerical simulation of nonlinear Schrödinger systems: a new conservative scheme*, Appl. Math. Comput., 71 (1995), pp. 165–177.

[25] K. FENG AND M. Z. QIN, *The symplectic methods for the computation of Hamiltonian equations*, in Numerical methods for partial differential equations (Shanghai, 1987), vol. 1297 of Lecture Notes in Math., Springer, Berlin, 1987, pp. 1–37.

[26] R. P. FEYNMAN, R. B. LEIGHTON, AND M. SANDS, *The Feynman lectures on physics, Vol. I: The new millennium edition: mainly mechanics, radiation, and heat*, vol. 1, Basic books, 2011.

[27] D. FURIHATA AND T. MATSUO, *Discrete variational derivative method: a structure-preserving numerical method for partial differential equations*, Chapman and Hall/CRC, 2010.

[28] E. GALLOPOULOS AND Y. SAAD, *Efficient solution of parabolic equations by Krylov approximation methods*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1236–1264.

[29] H. H. GOLDSTINE, *A history of numerical analysis from the 16th through the 19th century*, vol. 2, Springer Science & Business Media, 2012.

[30] Y. GONG, J. CAI, AND Y. WANG, *Some new structure-preserving algorithms for general multi-symplectic formulations of Hamiltonian PDEs*, J. Comput. Phys., 279 (2014), pp. 80–102.

[31] O. GONZALEZ, *Time integration and discrete Hamiltonian systems*, J. Nonlinear Sci., 6 (1996), pp. 449–467.

[32] E. HAIRER, C. LUBICH, AND G. WANNER, *Geometric numerical integration*, vol. 31 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 2006.

[33] A. HARTEN, P. D. LAX, AND B. VAN LEER, *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws*, SIAM Rev., 25 (1983), pp. 35–61.

[34] M. HOCHBRUCK, C. LUBICH, AND H. SELHOFER, *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comput., 19 (1998), pp. 1552–1574.

[35] T. ITOH AND K. ABE, *Hamiltonian-conserving discrete canonical equations based on variational difference quotients*, J. Comput. Phys., 76 (1988), pp. 85–102.

[36] W. KAHAN AND R.-C. LI, *Unconventional schemes for a class of ordinary differential equations with applications to the Korteweg–de Vries equation*, J. Comput. Phys., 134 (1997), pp. 316–331.

[37] B. LEIMKUHLER AND G. W. PATRICK, *A symplectic integrator for riemannian manifolds*, J. Nonlinear Sci., 6 (1996), pp. 367–384.

[38] B. LEIMKUHLER AND S. REICH, *Simulating Hamiltonian dynamics*, vol. 14, Cambridge university press, 2004.

[39] H. LEWY, *An example of a smooth linear partial differential equation without solution*, Ann. of Math. (2), 66 (1957), pp. 155–158.

[40] L. LI AND E. CELLEDONI, *Krylov projection methods for linear hamiltonian systems*, Numer. Algorithms, (2019), pp. 1–18.

[41] S. LI AND L. VU-QUOC, *Finite difference calculus invariant structure of a class of algorithms for the nonlinear Klein–Gordon equation*, SIAM J. Numer. Anal., 32 (1995), pp. 1839–1875.

[42] Y. LI AND X. WU, *General local energy-preserving integrators for solving multi-symplectic Hamiltonian PDEs*, J. Comput. Phys., 301 (2015), pp. 141–166.

[43] L. LOPEZ AND V. SIMONCINI, *Preserving geometric properties of the exponential matrix by block Krylov subspace methods*, BIT, 46 (2006), pp. 813–830.

[44] J. E. MARSDEN, G. W. PATRICK, AND S. SHKOLLER, *Multisymplectic geometry, variational integrators, and nonlinear PDEs*, Comm. Math. Phys, 199 (1998), pp. 351–395.

[45] J. E. MARSDEN AND A. WEINSTEIN, *The Hamiltonian structure of the Maxwell-Vlasov equations*, Phys. D, 4 (1981/82), pp. 394–406.

[46] T. MATSUO, M. SUGIHARA, D. FURIHATA, AND M. MORI, *Linearly implicit finite difference schemes derived by the discrete variational method*, RIMS Kokyuroku, (2000), pp. 121–129.

[47] F. MCDONALD, R. I. MCLACHLAN, B. E. MOORE, AND G. R. W. QUISPEL, *Travelling wave solutions of multisymplectic discretizations of semi-linear wave equations*, J. Difference Equ. Appl., 22 (2016), pp. 913–940.

[48] R. I. MCLACHLAN, G. QUISPEL, AND N. ROBIDOUX, *Geometric integration using discrete gradients*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 357 (1999), pp. 1021–1045.

[49] R. I. MCLACHLAN AND G. R. W. QUISPEL, *Discrete gradient methods have an energy conservation law*, Discrete Contin. Dyn. Syst., 34 (2014), pp. 1099–1104.

[50] B. MOORE AND S. REICH, *Backward error analysis for multi-symplectic integration methods*, Numer. Math., 95 (2003), pp. 625–652.

[51] P. J. OLVER, *Applications of Lie groups to differential equations*, vol. 107, Springer Science & Business Media, 2000.

[52] C. C. PAIGE, B. N. PARLETT, AND H. A. VAN DER VORST, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Numer. Linear Algebra Appl, 2 (1995), pp. 115–133.

[53] M. PETRERA AND Y. B. SURIS, *A construction of a large family of commuting pairs of integrable symplectic birational four-dimensional maps*, Proc. A., 473 (2017), pp. 20160535, 16.

[54] G. R. W. QUISPEL AND D. I. MCLAREN, *A new class of energy-preserving numerical integration methods*, J. Phys. A, 41 (2008), pp. 045206, 7.

[55] Y. SAAD, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[56] J. M. SANZ-SERNA, *Symplectic integrators for Hamiltonian problems: an overview*, in Acta numerica, 1992, Acta Numer., Cambridge Univ. Press, Cambridge, 1992, pp. 243–286.

[57] Y. SUN AND P. S. P. TSE, *Symplectic and multisymplectic numerical methods for Maxwell's equations*, J. Comput. Phys., 230 (2011), pp. 2076–2094.

[58] A. VAN DER SCHAFT AND D. JELTSEMA, *Port-Hamiltonian systems theory: An introductory overview*, Found. and Trends in Syst. and Control, 1 (2014), pp. 173–378.

[59] H. A. VAN DER VORST, *Iterative Krylov methods for large linear systems*, vol. 13 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2003.

[60] D. S. WATKINS, *The matrix eigenvalue problem: GR and Krylov subspace methods*, vol. 101, SIAM Rev., 2007.

[61] X. WU, B. WANG, AND W. SHI, *Efficient energy-preserving integrators for oscillatory Hamiltonian systems*, J. Comput. Phys., 235 (2013), pp. 587–605.

[62] G. ZHONG AND J. E. MARSDEN, *Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators*, Phys. Lett. A, 133 (1988), pp. 134–139.

[63] Z. ZLATEV AND J. WAŚNIEWSKI, *Running air pollution models on the connection machine*, Math. Comput. Modelling, 20 (1994), pp. 1–17.

Linearly implicit structure-preserving schemes for Hamiltonian systems

Sølve Eidnes, Lu Li and Shun Sato

Submitted

Linearly implicit structure-preserving schemes for Hamiltonian systems

Abstract. Kahan's method and a two-step generalization of the discrete gradient method are both linearly implicit methods that can preserve a modified energy for Hamiltonian systems with a cubic Hamiltonian. These methods are here investigated and compared. The schemes are applied to the Korteweg–de Vries equation and the Camassa–Holm equation, and the numerical results are presented and analysed.

2.1 Introduction

The field of geometric numerical integration (GNI) has garnered increased attention over the last three decades. It considers the design and analysis of numerical methods that can capture geometric properties of the flow of the differential equation to be modelled. These geometric properties are mainly invariants over time; they are conserved quantities such as Hamiltonian energy, angular momentum, volume or symplecticity. Numerical schemes inheriting such properties from the continuous dynamical system have been shown in many cases to be advantageous, especially when integration over long time intervals is considered [12].

For general non-linear differential equations, most of the geometric numerical integrators are fully implicit schemes [3, 5, 22]. Then a non-linear system must be solved at each time step. Typically this is done by the use of an iterative solver where a linear system is to be solved at each iteration. This quickly becomes a computationally expensive procedure, especially since the number of iterations needed in general increases with the size of the system. A fully explicit method on the other hand, may over-simplify the problem and lead to the loss of important information, and will often have inferior stability properties. The golden middle way may be found in linearly implicit schemes, i.e. schemes where the non-linear terms are discretized such that the solution at the next time step is found from solving one linear system; see a numerical example comparing the computational cost for implicit and linearly implicit methods in [7].

This paper focuses on a study of linearly implicit geometric numerical integrators for differential equations. We consider ordinary differential equations

(ODEs) that can be written in the form

$$\begin{aligned} \dot{x} &= f(x) = S\nabla H(x), \quad x \in \mathbb{R}^d, \\ x(0) &= x_0, \end{aligned} \tag{2.1.1}$$

where S is a constant skew-symmetric matrix and H is a cubic Hamiltonian function. The well-known geometric characteristic for equations like (2.1.1) is that the exact flow is energy-preserving,

$$\frac{d}{dt}H(x) = \nabla H(x)^T \frac{dx}{dt} = \nabla H(x)^T S \nabla H(x) = 0,$$

and symplectic if S is a canonical[1] skew-symmetric matrix,

$$\Psi_{x_0}(t)^T S \Psi_{x_0}(t) = S,$$

where $\Psi_{x_0}(t) := \frac{\partial \varphi_t(x_0)}{\partial x_0}$, with $\varphi_t : \mathbb{R}^d \to \mathbb{R}^d$, $\varphi_t(x_0) = x(t)$ the flow map of (2.1.1) [12]. A numerical one-step method is said to be energy-preserving if H is constant along the numerical solution, and symplectic if the numerical flow map is symplectic [12]. Both the energy-preserving methods and the symplectic methods, the latter of which has the ability to preserve a perturbation of the Hamiltonian H of (2.1.1), have their own advantages. However, there is no numerical integration method that can be simultaneously symplectic and energy-preserving for a general Hamiltonian system, which has no other conserved quantities than the Hamiltonian and functions of the Hamiltonian [23]. We will focus on the energy preservation property. In the continuous setting, it is shown in [21] that the conservation of energy plays an important role in proving the existence and uniqueness of solutions for partial differential equations (PDEs). From the numerical view, the energy-preserving property has been found to be crucial in the proof of stability for several such numerical methods, see e.g [8]. Some examples of energy-preserving methods are [2, 18, 19].

In this paper, we give a study of two types of existing linearly implicit methods with energy-preseving property. The first one is Kahan's method for quadratic ODE vector fields [14], which by construction is linearly implicit, and for which the geometric properties have been studied in [4]. This is a one-step method, but we will also give its formulation as a two-step method, for the convenience of comparison with other methods. The other method to be studied here, is based on the linearly implicit method for PDEs presented by Furihata, Matsuo and coauthors in the papers [15–17] and the monograph [10]. A generalization of this method, from two-step schemes to general multistep

[1] Here the $2n \times 2n$ canonical skew-symmetric matrix is a matrix of the form $\begin{bmatrix} 0_{n\times n} & I_{n\times n} \\ -I_{n\times n} & 0_{n\times n} \end{bmatrix}$, with $I_{n\times n}$ the identity matrix.

schemes, is given by Dahlby and Owren in [7]. We present here the two-step method as it looks for ODEs of the form (2.1.1), from which the schemes of the aforementioned references may arise after semi-discretizing the Hamiltonian PDE in space to obtain a system of Hamiltonian ODEs.

This paper is divided into two main parts. In the next chapter, we present the methods in consideration, and give some theoretical results on their geometric properties. In Chapter 3, we present numerical results for the Camassa–Holm equation and the Korteweg–de Vries equation, including analysis of stability and dispersion, comparing the methods.

2.2 Linearly implicit schemes

We will present the ODE formulation of the linearly implicit schemes presented by Furihata, Matsuo and coauthors in [10, 15–17] and by Dahlby and Owren in [7]. Following the nomenclature of the latter reference, we call these schemes polarised discrete gradient (PDG) methods. Then we present a special case of this polarization method in the same framework as Kahan's method, with the goal of obtaining more clarity in comparison of the methods.

2.2.1 Polarised discrete gradient methods

The idea behind the PDG methods is to generalize the discrete gradient method in such a way that a relaxed variant of the preservation property is intact, while nonlinear terms are discretized over consecutive time steps to ensure linearity in the scheme. Let us first recall the concept of discrete gradient methods. A discrete gradient is a continuous map $\overline{\nabla} H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ such that for any $x, y \in \mathbb{R}^d$

$$H(y) - H(x) = (y - x)^T \overline{\nabla} H(x, y). \tag{2.2.1}$$

The discrete gradient method for (2.1.1) is then given by

$$\frac{x^{n+1} - x^n}{\Delta t} = S \overline{\nabla} H(x^n, x^{n+1}),$$

which will preserve the energy of the system (2.1.1) at any time step. Here and in what follows, x^n is the numerical approximation for x at $t = t_n$ and x_k^n is the numerical approximation for the kth component of x at $t = t_n$. Restricting ourselves to two-step methods, we define the PDG methods as follows.

Definition 2.1. For the energy H of (2.1.1), consider the polarised energy as a function $\tilde{H}:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ satisfying the properties

$$\begin{aligned}\tilde{H}(x,x) &= H(x),\\ \tilde{H}(x,y) &= \tilde{H}(y,x).\end{aligned}$$

A polarised discrete gradient (PDG) for $\tilde{H}$ is a function $\overline{\nabla}\tilde{H}:\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}^d$ satisfying

$$\begin{aligned}\tilde{H}(y,z)-\tilde{H}(x,y) &= \frac{1}{2}(z-x)^T\overline{\nabla}\tilde{H}(x,y,z),\\ \overline{\nabla}\tilde{H}(x,x,x) &= \nabla H(x),\end{aligned} \tag{2.2.2}$$

and the corresponding polarised discrete gradient scheme is given by

$$\frac{x^{n+2}-x^n}{2\Delta t} = S\overline{\nabla}\tilde{H}(x^n,x^{n+1},x^{n+2}). \tag{2.2.3}$$

Proposition 2.1. *The numerical scheme* (2.2.3) *preserves the polarised invariant* $\tilde{H}$ *in the sense that* $\tilde{H}(x^n,x^{n+1})=\tilde{H}(x^0,x^1)$ *for all* $n\geq 0$.

Proof.

$$\begin{aligned}\tilde{H}(x^{n+1},x^{n+2})-\tilde{H}(x^n,x^{n+1}) &= \frac{1}{2}(x^{n+2}-x^n)^T\overline{\nabla}\tilde{H}(x^n,x^{n+1},x^{n+2})\\ &= \Delta t\overline{\nabla}\tilde{H}(x^n,x^{n+1},x^{n+2})^T S^T\overline{\nabla}\tilde{H}(x^n,x^{n+1},x^{n+2})\\ &= 0,\end{aligned}$$

where the last equality follows from the skew-symmetry of S. □

We remark here that in the cases where we seek a time-stepping scheme for the system of Hamiltonian ODEs resulting from discretizing a Hamiltonian PDE in space in an appropriate manner, e.g. as described in [3], H will be a discrete approximation to an integral $\mathcal{H}$. Thus a two-step PDG method and a standard one-step discrete gradient method, the latter in general fully implicit, will preserve two different discrete approximations separately to the same $\mathcal{H}$.

The task of finding a PDG satisfying (2.2.2) is approached differently in our two main references, [10,15–17] and [7]. Furihata, Matsuo and coauthors apply a generalization of the approach introduced by Furihata in [9] for finding discrete variational derivatives, while Dahlby and Owren suggest a generalization

of the average vector field (AVF) discrete gradient [18], given by

$$\overline{\nabla}_{\mathrm{AVF}}\tilde{H}(x,y,z) = 2\int_0^1 \nabla_x \tilde{H}(\xi x + (1-\xi)z, y)\,\mathrm{d}\xi,$$

where $\nabla_x \tilde{H}(x,y)$ is the gradient of $\tilde{H}(x,y)$ with respect to its first argument. Provided that the spatial discretization is performed in the same way, these two approaches lead to the same scheme for an $\tilde{H}$ quadratic in each of its arguments, as does a generalization of the midpoint discrete gradient of Gonzalez [11]. Based on this, we now propose a new, straightforward approach for finding this specific PDG:

Proposition 2.2. *Given a polarised energy function $\tilde{H}(x,y)$ which is at most quadratic in each of its arguments, define $\nabla_x \tilde{H}(x,y)$ as the partial derivative of $\tilde{H}$ with respect to its first argument. Then a PDG for $\tilde{H}$ is given by*

$$\overline{\nabla}\tilde{H}(x,y,z) = 2\nabla_x \tilde{H}(\frac{x+z}{2}, y). \tag{2.2.4}$$

Proof. We may write

$$\tilde{H}(x,y) = x^T A(y)x + b(y)^T x + c(y),$$

for some symmetric $A:\mathbb{R}^d \to \mathbb{R}^{d\times d}$, $b:\mathbb{R}^d \to \mathbb{R}^d$ and $c:\mathbb{R}^d \to \mathbb{R}$. Then

$$\nabla_x \tilde{H}(x,y) = 2A(y)x + b(y),$$

and

$$\begin{aligned}\nabla_x \tilde{H}(\frac{x+z}{2}, y)^T (z-x) &= (2A(y)\frac{x+z}{2} + b(y))^T (z-x)\\ &= z^T A(y)z + b(y)^T z - x^T A(y)x - b(y)^T x\\ &= \tilde{H}(y,z) - \tilde{H}(x,y).\end{aligned}$$

The last equation follows from definition 2.1 for a polarised energy. Furthermore, we have

$$\overline{\nabla}\tilde{H}(x,x,x) = 2\nabla_x \tilde{H}(x,x) = \nabla H(x).$$

□

As remarked in Theorem 4.5 of [7]: if the polarised energy $\tilde{H}(x,y)$ is at most quadratic in each of its arguments, the scheme (2.2.3) with the PDG (2.2.4) is linearly implicit.

An alternative to (2.2.4) could be a generalization of the Itoh–Abe discrete

gradient [13], defined by its i-th component

$$\overline{\nabla}_{\mathrm{IA}}\tilde{H}(x,y,z)_i = 2\begin{cases}\bar{\partial}\tilde{H}(x,y,z)_i & \text{if } x_i \neq z_i,\\ \frac{\partial\tilde{H}}{\partial x_i}(\sum_{j=1}^{i-1} z_j e_j + \sum_{j=i}^{n} x_j e_j, y)) & \text{if } x_i = z_i,\end{cases}$$

where e_j denotes the j_{th} Euclidean unit vector field, and

$$\bar{\partial}\tilde{H}(x,y,z)_i = \frac{\tilde{H}(\sum_{j=1}^{i} z_j e_j + \sum_{j=i+1}^{n} x_j e_j, y) - \tilde{H}(\sum_{j=1}^{i-1} z_j e_j + \sum_{j=i}^{n} x_j e_j, y)}{z_i - x_i}.$$

A symmetrized variant of this, given by $\overline{\nabla}_{\mathrm{SIA}}\tilde{H}(x,y,z) := \frac{1}{2}(\overline{\nabla}_{\mathrm{IA}}\tilde{H}(x,y,z) + \overline{\nabla}_{\mathrm{IA}}\tilde{H}(z,y,x))$ is again identical to (2.2.4), whenever $\tilde{H}$ is quadratic in each of its arguments.

2.2.2 A general framework and Kahan's method

For ODEs of the form (2.1.1), consider the two-step schemes of the form

$$\frac{x^{k+2}-x^k}{2\Delta t} = S\sum_{i,j=1}^{3}\alpha_{ij}(H''(x^{k-1+i})x^{k-1+j} + \beta(x^{k-1+i})), \tag{2.2.5}$$

where $H'' : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ is the Hessian matrix of H and $\beta(x) := 2\nabla H(x) - H''(x)x$. For cubic H, this scheme is linearly implicit if and only if $\alpha_{33} = 0$. It can be shown that many well known Runge–Kutta methods composed with themselves over two consecutive steps are methods in the class (2.2.5) when applied to (2.1.1) with cubic H. As two examples, the implicit midpoint method over two steps is (2.2.5) with $\alpha_{11} = \alpha_{33} = \frac{1}{16}, \alpha_{21} = \alpha_{22} = \alpha_{23} = \frac{1}{8}$, $\alpha_{ij} = 0$ otherwise, while the trapezoidal rule is (2.2.5) with $\alpha_{11} = \alpha_{33} = \frac{1}{8}, \alpha_{22} = \frac{1}{4}$, $\alpha_{ij} = 0$ otherwise. The integral-preserving average vector field method [20] over two steps is (2.2.5) with $\alpha_{11} = \alpha_{21} = \alpha_{23} = \alpha_{33} = \frac{1}{12}, \alpha_{22} = \frac{1}{6}$, $\alpha_{ij} = 0$ otherwise.

In this section, we first consider the case when the Hamiltonian is a cubic homogeneous polynomial, in which case the term $\beta(x)$ in (2.2.5) will disappear, and then we get the following results.

Theorem 2.1. *The scheme* (2.2.5) *with* $\alpha_{21} = \alpha_{23} = \frac{1}{4}$, $\alpha_{ij} = 0$ *otherwise, i.e.*

$$\frac{x^{n+2}-x^n}{2\Delta t} = \frac{1}{4}SH''(x^{n+1})(x^n + x^{n+2}), \tag{2.2.6}$$

is Kahan's method composed with itself over two consecutive steps when ap-

plied to ODEs of the form (2.1.1) *with homogeneous cubic H.*

Proof. As shown in [4], Kahan's method can be written into a Runge–Kutta form

$$\frac{x^{n+1}-x^n}{\Delta t} = -\frac{1}{2}f(x^n) + 2f(\frac{x^n+x^{n+1}}{2}) - \frac{1}{2}f(x^{n+1}).$$

Two steps of this can be written as

$$\begin{aligned}\frac{x^{n+2}-x^n}{2\Delta t} = &-\frac{1}{4}f(x^n) - \frac{1}{2}f(x^{n+1}) - \frac{1}{4}f(x^{n+2}) \\ &+ f(\frac{x^n+x^{n+1}}{2}) + f(\frac{x^{n+1}+x^{n+2}}{2}).\end{aligned} \tag{2.2.7}$$

Inserting $f(x) = S\nabla H(x)$ in (2.2.7) and noting that for homogeneous cubic H we have $\nabla H(x) = \frac{1}{2}H''(x)x$, $H''(x)y = H''(y)x$ and $H''(x+y) = H''(x) + H''(y)$, we can obtain (2.2.6). □

Remark 2.1. *From [4], it can be easily deduced that Kahan's method preserves the polarised energy* $\tilde{H}(x,y) = \frac{1}{3}\nabla H(x)y = \frac{1}{3}\nabla H(y)x = \frac{1}{6}x^T H''(\frac{x+y}{2})y$.

A special case of the PDG method which preserves the same polarised Hamiltonian as Kahan's method, can also be written on the form (2.2.5):

Theorem 2.2. *For a homogeneous cubic H and the polarised energy given by* $\tilde{H}(x,y) = \frac{1}{6}x^T H''(\frac{x+y}{2})y$, *the scheme* (2.2.3) *with the PDG* (2.2.4) *applied to* (2.1.1) *is equivalent to* (2.2.5) *with* $\alpha_{21} = \alpha_{22} = \alpha_{23} = \frac{1}{6}$, $\alpha_{ij} = 0$ *otherwise, i.e.*

$$\frac{x^{n+2}-x^n}{2\Delta t} = \frac{1}{6}SH''(x^{n+1})(x^n + x^{n+1} + x^{n+2}).$$

Proof.

$$\nabla_x \tilde{H}(x,y) = \frac{1}{6}H''(\frac{x+y}{2})y + \frac{1}{6}H''(\frac{y}{2})x = \frac{1}{12}H''(2x+y)y,$$

and thus

$$\overline{\nabla}\tilde{H}(x,y,z) = 2\nabla_x \tilde{H}(\frac{x+z}{2}, y) = \frac{1}{6}H''(x+y+z)y = \frac{1}{6}H''(y)(x+y+z).$$

□

When a non-homogeneous H is considered, one can use the technique employed in [4], adding one variable x_0 to generate an equivalent problem to

the original one, for a homogeneous Hamiltonian $\bar{H}:\mathbb{R}^{d+1}\to\mathbb{R}$ defined such that $\bar{H}(1,x_1,\dots,x_d)=H(x_1,\dots,x_d)$. Also constructing the $(d+1)\times(d+1)$ skew-symmetric matrix $\bar{S}$ by adding a zero initial row and a zero initial column to S, we get that solving the system

$$\begin{aligned}\dot{\bar{x}} &= \bar{S}\nabla\bar{H}(\bar{x}), \quad \bar{x}\in\mathbb{R}^{d+1}\\ \bar{x}(0) &= (1,x^0),\end{aligned} \tag{2.2.8}$$

is equivalent to solving (2.1.1). Following the above results for the homogeneous $\bar{H}$ and (2.2.8), we can generalize Theorem 2.1 and Theorem 2.2 for all cubic H. Generalizations of the preservation properties follow directly; e.g., Kahan's method and the PDG method can preserve the perturbed energy $\tilde{H}(x^n,x^{n+1}):=\frac{1}{6}(\bar{x}^n)^T\bar{H}''(\frac{\bar{x}^n+\bar{x}^{n+1}}{2})\bar{x}^{n+1}$ also for non-homogeneous cubic H.

2.3 Numerical experiments

To have a better understanding of the above methods, we will apply them to systems of two different PDEs: the Korteweg–de Vries (KdV) equation and the Camassa–Holm equation. We will compare our methods to the midpoint method, which is a symplectic, fully implicit method. We solve the two PDEs by discretizing in space to obtain a Hamiltonian ODE system of the type (2.1.1) and then apply the PDG method (denoted by PDGM), Kahan's method (denoted by Kahan) and the midpoint method (denoted by MP) to this.

2.3.1 Camassa–Holm equation

In this section, we consider the Camassa–Holm equation

$$u_t - u_{xxt} + 3uu_x = 2u_xu_{xx} + uu_{xxx}$$

defined on the periodic domain $\mathbb{S}:=\mathbb{R}/L\mathbb{Z}$. It has the conserved quantities

$$\mathcal{H}_1[u]=\frac{1}{2}\int_{\mathbb{S}}(u^2+u_x^2)\,\mathrm{d}x, \qquad \mathcal{H}_2[u]=\frac{1}{2}\int_{\mathbb{S}}\left(u^3+uu_x^2\right)\mathrm{d}x.$$

Here we consider the variational form of the Hamiltonian $\mathcal{H}_2$:

$$(1-\partial_x^2)u_t=-\partial_x\frac{\delta\mathcal{H}_2}{\delta u}, \qquad \frac{\delta\mathcal{H}_2}{\delta u}=\frac{3}{2}u^2+\frac{1}{2}u_x^2-(uu_x)_x. \tag{2.3.1}$$

We follow the approach presented in [3] and semi-discretize the energy $\mathcal{H}_2$

of (2.3.1) as

$$H_2(u)\Delta x = \frac{1}{2}\sum_{k=1}^{K}\left(u_k^3 + u_k\frac{(\delta_x^+ u_k)^2 + (\delta_x^- u_k)^2}{2}\right)\Delta x,$$

where the difference operators $\delta_x^+ u_k := \frac{u_{k+1}-u_k}{\Delta x}$, $\delta_x^- u_k := \frac{u_k - u_{k-1}}{\Delta x}$. For later use, we here also introduce the notation $\delta_x^{\langle 1\rangle} u_k := \frac{u_{k+1}-u_{k-1}}{2\Delta x}$, $\delta_x^{\langle 2\rangle} u_k := \frac{u_{k+1}-2u_k+u_{k-1}}{(\Delta x)^2}$, $\mu_x^+ u_k := \frac{u_{k+1}+u_k}{2}$, $\mu_x^- u_k := \frac{u_k+u_{k-1}}{2}$, and the matrices corresponding to the difference operators δ_x^+, δ_x^-, $\delta_x^{\langle 1\rangle}$, $\delta_x^{\langle 2\rangle}$, μ_x^+ and μ_x^- are denoted by D^+, D^-, $D^{\langle 1\rangle}$, $D^{\langle 2\rangle}$, M^+ and M^-. Denoting the numerical solution $U = [u_1, ...u_K]^T$, and by using the properties of the above difference operators, we thus get

$$\nabla H_2(U) = \frac{3}{2}U_\cdot^2 + \frac{1}{2}M^-(D^+U)_\cdot^2 - \frac{1}{2}D^{\langle 2\rangle}U_\cdot^2,$$

where $U_\cdot^2$ is the elementwise square of U. Then the semi-discretized system for the Camassa–Holm equation becomes

$$\dot{U} = S\nabla H_2(U) = -(I - D^{\langle 2\rangle})^{-1}D^{\langle 1\rangle}\nabla H_2(U). \tag{2.3.2}$$

The above-mentioned schemes applied to (2.3.2) give us

$$(I - D^{\langle 2\rangle})\frac{U^{n+1}-U^n}{\Delta t} = -D^{\langle 1\rangle}\nabla H_2(\frac{U^{n+1}+U^n}{2}), \qquad \text{(MP)}$$

$$(I - D^{\langle 2\rangle})\frac{U^{n+1}-U^n}{\Delta t} = -\frac{1}{2}D^{\langle 1\rangle}H_2''(U^n)U^{n+1}, \qquad \text{(Kahan)}$$

$$(I - D^{\langle 2\rangle})\frac{U^{n+2}-U^n}{2\Delta t} = -D^{\langle 1\rangle}\overline{\nabla}\tilde{H}_2(U^n, U^{n+1}, U^{n+2}), \qquad \text{(PDGM)}$$

where $H_2''(U) = 3\,\mathrm{diag}(U) + M^-\,\mathrm{diag}(D^+U)D^+ - D^{\langle 2\rangle}\,\mathrm{diag}(U)$ is the Hessian of $H_2(U)$ and $\overline{\nabla}\tilde{H}_2(U^n, U^{n+1}, U^{n+2})$ is the PDG of Proposition 2.2 with polarised discrete energy

$$\tilde{H}_2(U^n, U^{n+1}) := \frac{1}{2}\sum_{k=1}^{K}\left(u_k^n u_k^{n+1}\frac{u_k^n + u_k^{n+1}}{2} + a(\mu_x^+\frac{u_k^n + u_k^{n+1}}{2})(\delta_x^+ u_k^n)(\delta_x^+ u_k^{n+1})\right.$$
$$\left. + (1-a)\frac{(\mu_x^+ u_k^n)(\delta_x^+ u_k^{n+1})^2 + (\mu_x^+ u_k^{n+1})(\delta_x^+ u_k^n)^2}{2}\right)$$

for some $a \in \mathbb{R}$, typically between -1 and 2.

Remark 2.2. *We performed numerical experiments for finding a good choice of the parameter a in PDGM and based on these set $a = \frac{1}{2}$ in the following.*

Numerical tests for the Camassa–Holm equation

Example 1 (Single peakon solution): In this numerical test, we consider the same experiment as in [6], where multisymplectic schemes are considered for the Camassa–Holm equation with

$$u(x,0) = \frac{\cosh(|x-\frac{L}{2}|-\frac{L}{2})}{\cosh(L/2)},$$

$x \in [0,L]$, $L = 40$, $t \in [0,T]$, $T = 5$, spatial step size $\Delta x = 0.04$ and time step size $\Delta t = 0.0002$. From Figure 2.2 (the right two), we observe that all considered methods keep a shape close to the exact solution except some small oscillatory tails, resulting from the semi-discretization, as observed in [6]. The numerical simulations show that the global error is mainly due to the shape error[2], see Figure 2.1. In Figure 2.2 (the left one), we can see that the numerical energy for all the methods oscillate, but appears to be bounded. Here we consider also coarser grids. We observe that there appear some small wiggles for both PDGM and Kahan's method for $\Delta t = 0.02$ and long time integration $T = 100$. However, the wiggles in the solution by PDGM are much more evident than those in the solution of Kahan's method, see Figure 2.3 (the left two plots). We keep on increasing Δt to 0.15 and 0.2; we observe that the numerical solution obtained by the PDGM suffers from evident numerical dispersion when $\Delta t = 0.15$, while Kahan's method seems to keep the shape well when comparing to the exact wave. Spurious oscillations appear also in Kahan's method when the time-step is increased to the value $\Delta t = 0.2$, see Figure 2.3 (right).

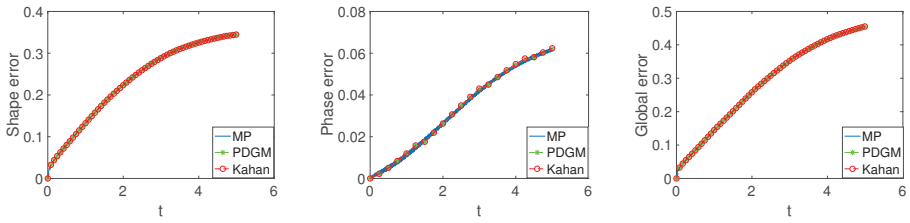


Figure 2.1: In this experiment, space step size $\Delta x = 0.04$ and time step size $\Delta t = 0.0002$. **Left:** shape error, **middle:** phase error, **right:** global error.

[2] Shape error is defined by $\epsilon_{\text{shape}} := \min_{\tau} \| U^n - u(\cdot - \tau) \|_2^2$, and phase error is defined by $\epsilon_{\text{phase}} := |\operatorname{argmin}_{\tau} \| U^n - u(\cdot - \tau) \|_2^2 - ct_n|$, [7].

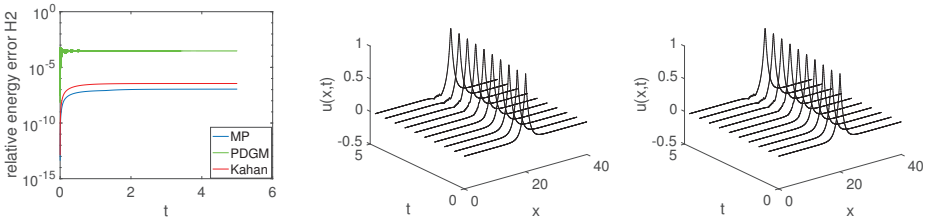


Figure 2.2: In this experiment, $\Delta x = 0.04$, $\Delta t = 0.0002$. **Left:** relative energy errors. **middle:** propagation of the wave by PDGM. **right:** propagation of the wave by Kahan's method.

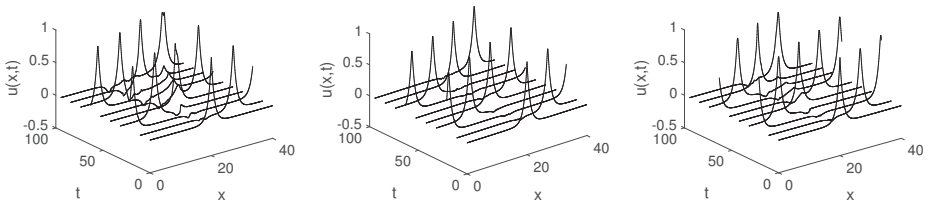


Figure 2.3: In this experiment, space step size $\Delta x = 0.04$. **Left:** propagation of the wave by PDGM, $\Delta t = 0.02$, **middle:** propagation of the wave by Kahan's method, $\Delta t = 0.02$, **right:** propagation of the wave by Kahan's method, $\Delta t = 0.15$.

Example 2 (Two peakons solution): Now we consider the initial condition

$$u(x,0) = \frac{\cosh(|x-\frac{L}{4}|-\frac{L}{2})}{\cosh(L/2)} + \frac{3}{2}\frac{\cosh(|x-\frac{3L}{4}|-\frac{L}{2})}{\cosh(L/2)},$$

where $x \in [0,L]$, $L = 40$, $t \in [0,T]$, $T = 5$, and $\Delta x = 0.04$, $\Delta t = 0.0002$. We observe that all the methods keep the shape of the exact solution very well and the numerical energy appears bounded, see Figure 2.5. The numerical simulation shows that the global error is mainly due to the shape error, see Figure 2.4. When a coarser time grid and longer integration time is considered, $\Delta t = 0.02$ and $T = 100$, small wiggles appear in the solution of PDGM and Kahan's method, see Figure 2.6 (the left two figures). We increase Δt to 0.2, and observe that PDGM fails to preserve the shape of the solution, while Kahan's method can still keep a shape close to the exact solution even though also for this method the numerical dispersion increases, see Figure 2.6 (right).

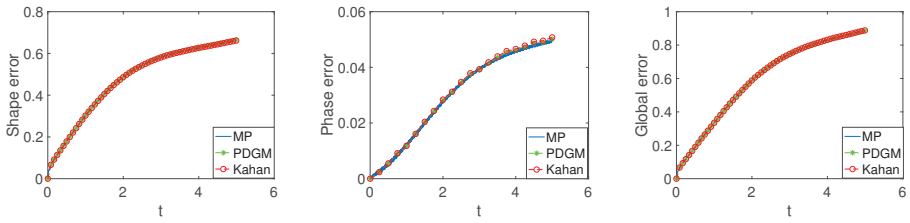


Figure 2.4: In this experiment, space step size $\Delta x = 0.04$, time step size $\Delta t = 0.0002$. **Left:** shape error, **middle:** phase error, **right:** global error.

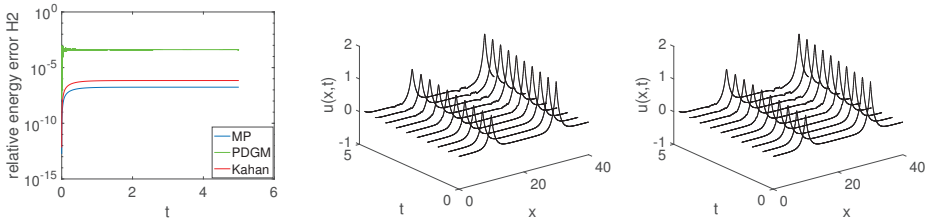


Figure 2.5: In this experiment, $\Delta x = 0.04$, $\Delta t = 0.0002$. **Left:** relative energy errors, **middle:** propagation of the wave by PDGM, **right:** propagation of the wave by Kahan's method.

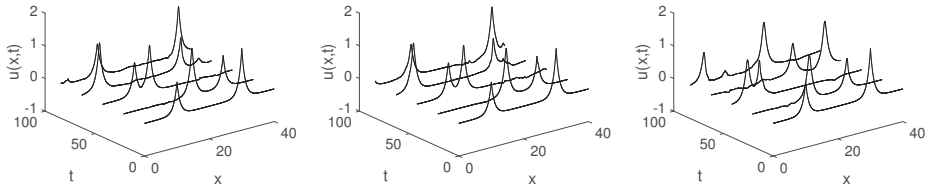


Figure 2.6: In this experiment, $\Delta x = 0.04$. **Left:** propagation of the wave by PDGM, $\Delta t = 0.02$, **middle:** propagation of the wave by Kahan's method, $\Delta t = 0.02$, **right:** propagation of the wave by Kahan's method, $\Delta t = 0.2$.

2.3.2 Korteweg–de Vries equation

In the previous example, the vector field of the semi-discretized system based on the Camassa–Holm equation is a homogeneous cubic polynomial. In this section, we deal with the KdV equation, for which the vector field of the semi-discretized equation is a non-homogeneous cubic polynomial.

The KdV equation

$$u_t + 6uu_x + u_{xxx} = 0 \tag{2.3.3}$$

on the periodic domain $\mathbb{S} := \mathbb{R}/L\mathbb{Z}$ has the conserved Hamiltonians

$$\mathcal{H}_1(u(t)) = \frac{1}{2}\int_{\mathbb{S}} u^2 \,\mathrm{d}x, \qquad \mathcal{H}_2(u(t)) = \int_{\mathbb{S}} \left(-u^3 + \frac{1}{2}u_x^2\right)\mathrm{d}x.$$

In the following we consider the variational form based on the Hamiltonian $\mathcal{H}_2$:

$$u_t = \partial_x \frac{\delta \mathcal{H}_2}{\delta u}, \qquad \frac{\delta \mathcal{H}_2}{\delta u} = -3u^2 - u_{xx}. \tag{2.3.4}$$

Numerical schemes for the KdV equation

We discretize the energy $\mathcal{H}_2$ for the KdV equation (2.3.4) as

$$H_2(U)\Delta x = \sum_{k=1}^{K}\left(-u_k^3 + \frac{(\delta_x^+ u_k)^2 + (\delta_x^- u_k)^2}{4}\right)\Delta x.$$

From simple calculations, the corresponding gradient is given by

$$\nabla H_2(U) = \left(-3U_\cdot^2 - D^{\langle 2\rangle}U\right),$$

and thus we have the semi-discretized form for (2.3.4):

$$\dot{U} = D^{\langle 1\rangle}\left(-3U_\cdot^2 - D^{\langle 2\rangle}U\right). \tag{2.3.5}$$

Applying the schemes under consideration to (2.3.5) gives

$$\begin{aligned}
&\frac{U^{n+1}-U^n}{\Delta t} = D^{\langle 1\rangle}\nabla H_2\left(\frac{U^n+U^{n+1}}{2}\right), && \text{(MP)}\\
&\frac{U^{n+1}-U^n}{\Delta t} = -\frac{1}{2}D^{\langle 1\rangle}\left(\nabla H(U^n) + \nabla H(U^{n+1})\right)\\
&\qquad\qquad + 2D^{\langle 1\rangle}\nabla H\left(\frac{U^n+U^{n+1}}{2}\right), && \text{(Kahan)}\\
&\frac{U^{n+2}-U^n}{2\Delta t} = D^{\langle 1\rangle}\overline{\nabla}\tilde{H}_2(U^n, U^{n+1}, U^{n+2}), && \text{(PDGM)}
\end{aligned}$$

where $H_2''(U) = -6\,\mathrm{diag}(U) - D^{\langle 2\rangle}$ is the Hessian of $H_2(U)$, and the polarised discrete gradient $\overline{\nabla}\tilde{H}_2(U^n, U^{n+1}, U^{n+2})$ is found as in Proposition 2.2, with

polarised discrete energy

$$\tilde{H}_2(u_k^n, u_k^{n+1}) := \sum_{k=1}^{d} (-u_k^n u_k^{n+1} \frac{u_k^n + u_k^{n+1}}{2} + \frac{a}{2}(\delta_x^+ u_k^n)(\delta_x^+ u_k^{n+1}) + \frac{1-a}{2} \frac{(\delta_x^+ u_k^n)^2 + (\delta_x^+ u_k^{n+1})^2}{2}) \Delta x.$$

Remark 2.3. *We perform several numerical simulations to find a good choice of parameter a, and we take* $a = -\frac{1}{2}$ *for PDGM in the following numerical examples for KdV equation.*

Stability analysis of the schemes

To analyse the stability of the above methods, we perform the von Neumann stability analysis for the Kahan and PDGM schemes applied to the linearized form of the KdV equation (2.3.3)

$$u_t + u_{xxx} = 0.$$

The equation for the amplification factor for Kahan's method is

$$(1 + i\lambda(\cos\theta - 1)\sin\theta)g + i\lambda(\cos\theta - 1)\sin\theta - 1 = 0,$$

and its root is

$$g = \frac{1 - i\lambda(\cos\theta - 1)\sin\theta}{1 + i\lambda(\cos\theta - 1)\sin\theta},$$

where $\lambda := \frac{\Delta t}{\Delta x^3}$. Since g is a simple root on the unit circle, Kahan's method is unconditionally stable for the linearized KdV equation.

The equation for the amplification factor for PDGM is

$$g^2 - 1 + i\lambda(3g^2 - 2g + 3)(\cos\theta - 1)\sin\theta = 0.$$

The two roots of the above equation are thus

$$g_1 = \frac{3b^2 + \sqrt{1+8b^2} + ib(3\sqrt{1+8b^2} - 1)}{1+9b^2},$$
$$g_2 = \frac{3b^2 - \sqrt{1+8b^2} - ib(3\sqrt{1+8b^2} + 1)}{1+9b^2},$$

where $b = \lambda(1 - \cos\theta)\sin\theta$. We observe that $|g_1| = |g_2| = 1$, and $g_1 \neq g_2$, therefore the PDGM is unconditionally stable for the linearized KdV equation.

Numerical tests for the KdV equation

Example 1 (One soliton solution): Consider the initial value

$$u(x,0) = 2\,\mathrm{sech}^2(x - L/2),$$

where $x \in [0, L]$, $L = 40$. We apply our schemes over the time interval $[0, T]$, $T = 100$, with step sizes $\Delta x = 0.05$, $\Delta t = 0.0125$. From our observations, all the methods behave well. The shape of the wave is well kept by all the methods, also for long integration time, see Figure 2.7. The energy errors of all the methods are rather small and do not increase over long time integration, see Figure 2.8 (left). We then use a coarser time grid, $\Delta t = 0.035$, and both methods are still stable, see Figure 2.9 (left two). However we observe that the global error of PDGM becomes much bigger than that of Kahan's method. When an even larger time step-size, $\Delta t = 0.04$, is considered, the solution for PDGM blows up while the solution for Kahan's method is rather stable. In this case, the PDGM applied to the nonlinear KdV equation is unstable and the numerical solution blows up at around t=8. Even if we increase the time step-size to $\Delta t = 0.1$, Kahan's method still works well, see Figure 2.9 (middle). When $\Delta t = 0.15$ is considered, we observe evident signs of instability in the solution of Kahan's method. The solution will blow up rapidly when $\Delta t = 0.2 \gg \Delta x$.

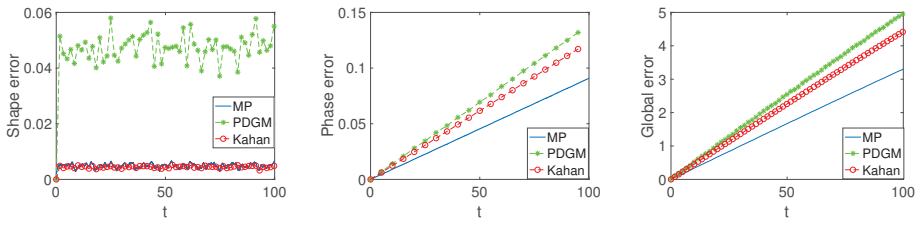


Figure 2.7: Space step size $\Delta x = 0.05$, time step size $\Delta t = 0.0125$. **Left:** shape error, **middle:** phase error, **right:** global error.

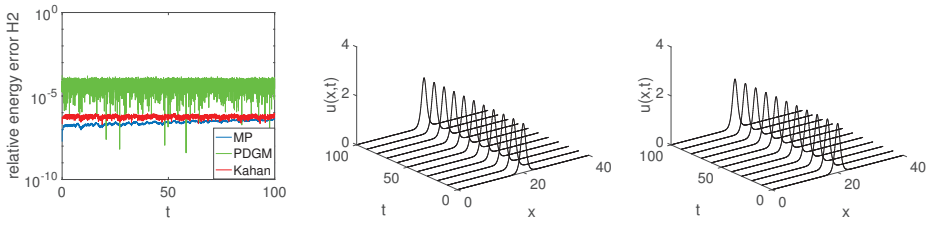


Figure 2.8: With $\Delta x = 0.05$, $\Delta t = 0.0125$. **Left:** relative energy errors, **right two:** propagation of the wave by PDGM and Kahan's method.

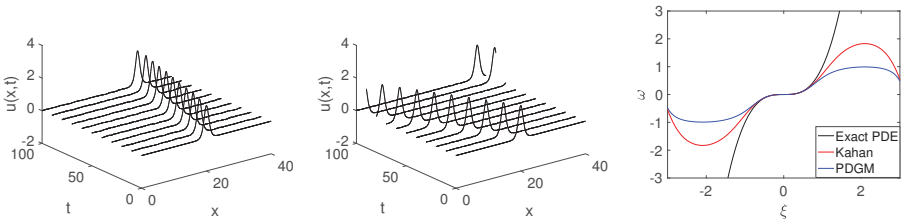


Figure 2.9: With $\Delta x = 0.05$. **Left:** $\Delta t = 0.035$, propagation of the wave by PDGM, **middle:** $\Delta t = 0.1$, propagation of the wave by Kahan's method, **right:** dispersion relation for $\lambda = 1$.

Example 2 (Two solitons solution): We choose initial value

$$u(x,0) = 6\,\mathrm{sech}^2 x,$$

and consider periodic boundary conditions $u(0,t) = u(L,t)$, where $x \in [0,L]$, $L = 40$. We set the space step size $\Delta x = 0.05$ and apply the aforementioned schemes on time interval $[0,T]$ with $T = 100$, $\Delta t = 0.001$. All the methods behave stably. The profiles of Kahan's method and the midpoint method are almost indistinguishable, and the profiles for the midpoint method are thus not presented here. Kahan's method and PDGM preserve the modified energy, and accordingly the energy error of all the methods are rather small over long time integration, see Figure 2.10 (left). After a short while the solution has two solitons; one is tall and the other is shorter, see Figure 2.10 (the right two plots).

When we consider a coarser time grid, $\Delta t = 0.00375$, both methods are still stable, see Figure 2.11 (the left two). However, there appear more small wiggles in the solution by PDGM and we observe that the solution of PDGM will blow up soon, around $t = 1$, for an even coarser time grid $\Delta t = 0.005$. When we increase the time step size to $\Delta t = 0.0125$ and consider $T = 100$, the shape of the exact solution is still well preserved by Kahan's method, even though there appear some small wiggles in the solution at around $t = 100$. We observe that the solution of Kahan's method will blow up when $\Delta t = 0.05$ is considered. Similar experiments as in this subsection, but for the multisymplectic box schemes, can be found in a paper by Ascher and McLachlan [1]. However, here we consider even coarser time grid than there, and the numerical results show that Kahan's method is quite stable, even though it is linearly implicit.

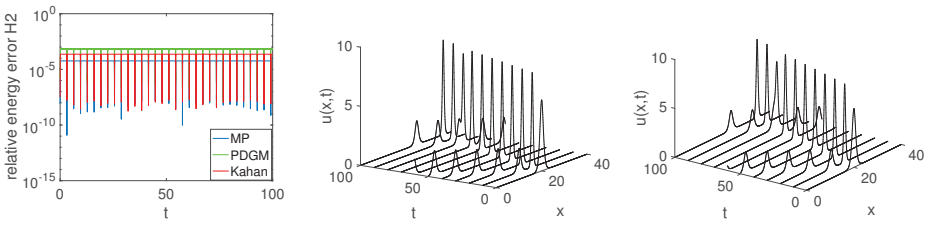


Figure 2.10: In this experiment, $\Delta x = 0.05$, $\Delta t = 0.001$. **Left:** relative energy errors, **right two:** propagation of the wave by PDGM and Kahan's method.

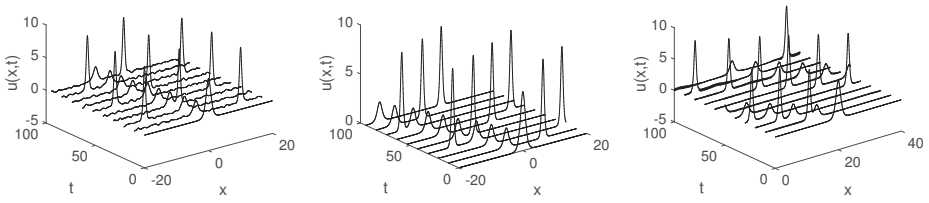


Figure 2.11: $\Delta x = 0.05$. **Left:** Propagations of the wave by PDGM, $\Delta t = 0.00375$, **middle:** propagations of the wave by Kahan's method, $\Delta t = 0.00375$, **right:** Propagations of the wave by Kahan's method, $\Delta t = 0.0125$.

Dispersion analysis

We consider the traditional linear analysis of numerical dispersion relations for the numerical schemes applied to the KdV equation, getting the dispersion relation of frequency ω and wave number ξ to be

$$\omega = \xi^3, \qquad \text{(exact solution)} \tag{2.3.6}$$

$$\sin\omega = \lambda(1-\cos\xi)(3\cos\omega - 1)\sin\xi, \qquad \text{(PDGM)} \tag{2.3.7}$$

$$\frac{\sin\omega}{1+\cos\omega} = \lambda(1-\cos\xi)\sin\xi, \qquad \text{(Kahan)} \tag{2.3.8}$$

where $\lambda = \frac{\Delta t}{\Delta x^3}$. The dispersion curve is displayed in Figure (2.9) (right). We observe that Kahan's method is better than PDGM at preserving the exact dispersion relation. This coincides with the behaviour of the methods applied to the nonlinear KdV equation shown in Section 2.3.2.

2.4 Conclusion

In this paper we perform a comparative study of Kahan's method and what we call the polarised discrete gradient method (PDGM). To that end, we present a general form encompassing a class of two-step methods that includes both a specific case of the PDGM and Kahan's method composed with itself. We also compare the methods for completely integrable Hamiltonian PDEs, the KdV equation and the Camassa–Holm equation. Both Kahan's method and the PDGM are linearly implicit methods, which will save the computational cost. A series of numerical experiments has been performed here, for the KdV equation with one and two solitons, and the Camassa–Holm equation with one and two peakons. These experiments show that Kahan's method is more stable than the PDGM. They also indicate that Kahan's method yields more accurate results, as we have witnessed in the energy error and the shape and phase error when comparing to analytical solutions. Based on our results, we would recommend the use of Kahan's method if one seeks a linearly implicit scheme for a Hamiltonian system with cubic H.

Bibliography

[1] U. M. ASCHER AND R. I. MCLACHLAN, *On symplectic and multisymplectic schemes for the KdV equation*, J. Sci. Comput., 25 (2005), pp. 83–104.

[2] L. BRUGNANO, F. IAVERNARO, AND D. TRIGIANTE, *Hamiltonian boundary value methods (energy preserving discrete line integral methods)*, J. Numer. Anal. Ind. Appl. Math, 5 (2010), pp. 17–37.

[3] E. CELLEDONI, V. GRIMM, R. I. MCLACHLAN, D. I. MCLAREN, D. O'NEALE, B. OWREN, AND G. R. W. QUISPEL, *Preserving energy resp. dissipation in numerical PDEs using the "average vector field" method*, J. Comput. Phys., 231 (2012), pp. 6770–6789.

[4] E. CELLEDONI, R. I. MCLACHLAN, B. OWREN, AND G. R. W. QUISPEL, *Geometric properties of Kahan's method*, J. Phys. A, 46 (2013), pp. 025201, 12.

[5] J.-B. CHEN AND M.-Z. QIN, *Multi-symplectic Fourier pseudospectral method for the nonlinear Schrödinger equation*, Electron. Trans. Numer. Anal., 12 (2001), pp. 193–204.

[6] D. COHEN, B. OWREN, AND X. RAYNAUD, *Multi-symplectic integration of the Camassa-Holm equation*, J. Comput. Phys., 227 (2008), pp. 5492–5512.

[7] M. DAHLBY AND B. OWREN, *A general framework for deriving integral preserving numerical methods for PDEs*, SIAM J. Sci. Comput., 33 (2011), pp. 2318–2340.

[8] Z. FEI, V. M. PÉREZ-GARCÍA, AND L. VÁZQUEZ, *Numerical simulation of nonlinear Schrödinger systems: a new conservative scheme*, Appl. Math. Comput., 71 (1995), pp. 165–177.

[9] D. FURIHATA, *Finite difference schemes for $\partial u/\partial t = (\partial/\partial x)^{\alpha}\delta G/\delta u$ that inherit energy conservation or dissipation property*, J. Comput. Phys., 156 (1999), pp. 181–205.

[10] D. FURIHATA AND T. MATSUO, *Discrete variational derivative method*, Chapman & Hall/CRC Numerical Analysis and Scientific Computing, CRC Press, Boca Raton, FL, 2011. A structure-preserving numerical method for partial differential equations.

[11] O. GONZALEZ, *Time integration and discrete Hamiltonian systems*, J. Nonlinear Sci., 6 (1996), pp. 449–467.

[12] E. HAIRER, C. LUBICH, AND G. WANNER, *Geometric numerical integration*, vol. 31 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 2006.

[13] T. ITOH AND K. ABE, *Hamiltonian-conserving discrete canonical equations based on variational difference quotients*, J. Comput. Phys., 76 (1988), pp. 85–102.

[14] W. KAHAN, *Unconventional numerical methods for trajectory calculations*, Unpublished lecture notes, (1993).

[15] T. MATSUO, *New conservative schemes with discrete variational derivatives for nonlinear wave equations*, J. Comput. Appl. Math., 203 (2007), pp. 32–56.

[16] T. MATSUO AND D. FURIHATA, *Dissipative or conservative finite-difference schemes for complex-valued nonlinear partial differential equations*, J. Comput. Phys., 171 (2001), pp. 425–447.

[17] T. MATSUO, M. SUGIHARA, D. FURIHATA, AND M. MORI, *Spatially accurate dissipative or conservative finite difference schemes derived by*

the discrete variational method, Japan J. Indust. Appl. Math., 19 (2002), pp. 311–330.

[18] R. I. MCLACHLAN, G. QUISPEL, AND N. ROBIDOUX, *Geometric integration using discrete gradients*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 357 (1999), pp. 1021–1045.

[19] G. R. W. QUISPEL AND D. I. MCLAREN, *A new class of energy-preserving numerical integration methods*, J. Phys. A, 41 (2008), pp. 045206, 7.

[20] G. R. W. QUISPEL AND D. I. MCLAREN, *A new class of energy-preserving numerical integration methods*, J. Phys. A, 41 (2008), pp. 045206, 7.

[21] M. E. TAYLOR, *Partial differential equations I. Basic theory*, vol. 115 of Applied Mathematical Sciences, Springer, New York, second ed., 2011.

[22] X. WU, B. WANG, AND W. SHI, *Efficient energy-preserving integrators for oscillatory Hamiltonian systems*, J. Comput. Phys., 235 (2013), pp. 587–605.

[23] G. ZHONG AND J. E. MARSDEN, *Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators*, Phys. Lett. A, 133 (1988), pp. 134–139.

Linearly implicit local and global energy-preserving methods for Hamiltonian PDEs

Sølve Eidnes and Lu Li

Submitted

Linearly implicit local and global energy-preserving methods for Hamiltonian PDEs

Abstract. We present linearly implicit methods that preserve discrete approximations to local and global energy conservation laws for multi-symplectic PDEs with cubic invariants. The methods are tested on the one-dimensional Korteweg–de Vries equation and the two-dimensional Zakharov–Kuznetsov equation; the numerical simulations confirm the conservative properties of the methods, and demonstrate their good stability properties and superior running speed when compared to fully implicit schemes.

3.1 Introduction

In recent years, much attention has been given to the design and analysis of numerical methods for differential equations that can capture geometric properties of the exact flow. The increased interest in this subject can mainly be attributed to the superior qualitative behaviour over long time integration of such structure-preserving methods, see [13, 19, 21]. A popular class of structure-preserving methods are energy-preserving methods. In particular, the energy preservation property has been found to be crucial in the proof of stability for several of these numerical methods, see e.g [16–18].

Energy-preserving methods are well studied for finite-dimensional Hamiltonian systems [5,7,27,32]. It is also highly conceivable that the ideas behind the finite-dimensional setting can be extended to the infinite-dimensional Hamiltonian systems or Hamiltonian partial differential equations (PDEs) [4]. There are two popular ways to construct energy-preserving methods for Hamiltonian PDEs. One approach is to semi-discretize the PDE in space so that one obtains a system of Hamiltonian ordinary differential equations (ODEs), and then apply an energy-preserving method to this semi-discrete system, see for example [7]. In this way, it is straightforward to generalise the energy-preserving methods for finite-dimensional Hamiltonian systems to Hamiltonian PDEs. However, such methods conserve only a global energy that relies on a proper boundary condition, such as a periodic boundary condition. If this is not present, the energy-preserving property will be destroyed. The other approach is based on a reformulation of the Hamiltonian PDE into a multi-symplectic form, which provides the PDE with three local conservation laws: the multi-symplectic conservation law, the energy conservation and the momentum conservation law [2, 3, 28]. Then one may consider methods that preserve the local con-

servation laws, see for example [36]. These locally defined properties are not dependent on the choice of boundary conditions, giving the methods that preserve local energy an advantage over methods that preserve a global energy, especially since local conservation laws will always lead to global conservation laws whenever periodic boundary conditions are considered. The concept of a multi-symplectic structure for PDEs was introduced by Bridges in [2, 3], see also [30] for a framework based on a Lagrangian formulation of the Cartan form. Local energy-preserving methods were first studied in [35], and have garnered much interest recently, see for example [20, 29, 36].

Most of the local energy-preserving methods proposed so far are fully implicit methods, for which a non-linear system must be solved at each time step. This is normally done by using an iterative solver where a linear system is solved at each iteration, which can lead to computationally expensive procedures, especially since the number of iterations needed in general increases with the size of the system. A fully explicit method on the other hand, may over-simplify the problem and often has inferior stability properties, so that a strong restriction on the grid ratio is needed. A good alternative may therefore be to develop linearly implicit schemes, where the solution at the next time step is found by solving only one linear system.

One example of linearly implicit methods for Hamiltonian ODEs is Kahan's method, which was designed for solving quadratic ODEs [26] and whose geometric properties have been studied in a series of papers by Celledoni et al. [8, 10, 11]. For Hamiltonian PDEs, Matsuo and Furihata proposed the idea of using multiple points to discretize the variational derivative and thus design linearly implicit energy-preserving schemes [31]. Dahlby and Owren generalised this concept and developed a framework for deriving linearly implicit energy-preserving multi-step methods for Hamiltonian PDEs with polynomial invariants [14]. A comparison of this approach and Kahan's method applied to PDEs is given in [15]. Recently, more work has been put into developing linearly implicit energy-preserving schemes for Hamiltonian PDEs, e.g. the partitioned averaged vector field (PAVF) method [6] and schemes based on the invariant energy quadratization (IEQ) approach [37] or the multiple scalar auxiliary variables (MSAV) approach [25]. However, little attention has been given to linearly implicit local energy-preserving methods. To the best of the authors' knowledge, the only existing method is one based on the IEQ approach, specific for the sine-Gordon equation [24]. In this paper, we use Kahan's method to construct a linearly implicit method that preserves a discrete approximation to the local energy for multi-symplectic PDEs with a cubic energy function.

The rest of this paper is organized as follows. First, we give an overview of Kahan's method and formulate it by using a polarised energy function.

A brief introduction to multi-symplectic PDEs and their conservation laws are presented in Section 3.3. In Section 3.4, new linearly implicit local and global energy-preserving schemes are presented. Numerical examples for the Korteweg–de Vries (KdV) and Zakharov–Kuznetsov equations are given in Section 3.5, before we end the paper with some concluding remarks.

3.2 Kahan's method

Consider an ODE system

$$\dot{y} = f(y) = \hat{Q}(y) + \hat{B}y + \hat{c}, \quad y \in \mathbb{R}^M, \tag{3.2.1}$$

where $\hat{Q}(y)$ is an $\mathbb{R}^M$ valued quadratic form, $\hat{B} \in \mathbb{R}^{M\times M}$ is a symmetric constant matrix, and $\hat{c} \in \mathbb{R}^M$ is a constant vector. Kahan's method is then given by

$$\frac{y^{n+1} - y^n}{\Delta t} = \bar{Q}(y^n, y^{n+1}) + \hat{B}\frac{y^n + y^{n+1}}{2} + \hat{c},$$

where

$$\bar{Q}(y^n, y^{n+1}) = \frac{1}{2}\big(\hat{Q}(y^n + y^{n+1}) - \hat{Q}(y^n) - \hat{Q}(y^{n+1})\big)$$

is the symmetric bilinear form obtained by polarisation of the quadratic form $\hat{Q}$ [10]. Polarisation, which maps a homogeneous polynomial function to a symmetric multi-linear form in more variables, was used to generalise Kahan's method to higher degree polynomial vector fields in [9].

Suppose we restrict the problem (3.2.1) to be a Hamiltonian system on a Poisson vector space with a constant Poisson structure:

$$\dot{y} = A\nabla H(y), \tag{3.2.2}$$

where A is a constant skew-symmetric matrix, and $H : \mathbb{R}^M \to \mathbb{R}$ is a cubic polynomial function. We first consider the Hamiltonian H to be homogeneous. Then, following the result in Proposition 2.1 of [9], Kahan's method can be reformulated as

$$\frac{y^{n+1} - y^n}{\Delta t} = 3A\bar{H}(y^n, y^{n+1}, \cdot), \tag{3.2.3}$$

where $\bar{H}(\cdot,\cdot,\cdot) : \mathbb{R}^M \times \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ is a symmetric 3-tensor satisfying $\bar{H}(x, x, x) = H(x)$. Consider the 3-tensor $\bar{H}(x, y, z) = x^T Q(y)z$, where $Q(y) = \frac{1}{6}\nabla^2 H(y)$, with $\nabla^2 H$ being the Hessian of H; then we can rewrite Kahan's method (3.2.3)

as

$$\frac{y^{n+1}-y^n}{\Delta t} = 3A\frac{\partial \bar{H}}{\partial x}\bigg|_{(y^n, y^{n+1})}, \tag{3.2.4}$$

where $\frac{\partial \bar{H}}{\partial x}$ denotes the partial derivative with respect to the first argument of $\bar{H}$.

Consider then the cases where the Hamiltonian in problem (3.2.2) is non-homogeneous, i.e. of the general form

$$H(y) = y^T Q(y) y + y^T B y + c^T y + d, \tag{3.2.5}$$

where $Q(y)$ is the linear part of $\nabla^2 H(y)$ and thus a symmetric matrix whose elements are homogeneous linear polynomials, B is the constant part of $\nabla^2 H(y)$ and thus a symmetric constant matrix, c is a constant vector and d is a constant scalar. We follow the technique in [10], adding one variable to $y = (y_1, \ldots, y_M)^T$ to get $\tilde{y} = (y_0, y_1, \ldots, y_M)^T$, extending A to $\tilde{A}$ by adding a zero initial row and a zero initial column, considering a homogeneous function $\tilde{H}(\tilde{y})$ based on the non-homogeneous Hamiltonian $H(y)$ such that $\tilde{H}(\tilde{y})|_{y_0=1} = H(y)$, and finally solving instead of (3.2.2) the equivalent, homogeneous cubic Hamiltonian problem

$$\dot{\tilde{y}} = \tilde{A}\nabla \tilde{H}(\tilde{y})$$

with $y_0 = 1$. In this way we can still get the reformulation of Kahan's method as (3.2.4) with

$$\bar{H}(x, y, z) = x^T Q(y) z + \frac{1}{3}(x^T B y + y^T B z + z^T B x) + \frac{1}{3} c^T (x + y + z) + d. \tag{3.2.6}$$

Remark 3.1. *The $\mathbb{R}$-valued function $\bar{H}(x, y, z)$ in* (3.2.6) *has the following properties:*

1. *$\bar{H}(x, y, z)$ is symmetric*[1] *w.r.t. x, y and z,*
2. *$\bar{H}(x, x, x) = H(x)$,*
3. *$\frac{\partial \bar{H}(x,y,z)}{\partial x} = Q(y)z + \frac{B(y+z)}{3} + \frac{c}{3}$ is symmetric w.r.t. y and z.*

[1] Denote the elements in $Q(y)$ by $q_{ij}y = \sum_k q_{ij}^k y_k$, where q_{ij}^k, $i, j, k = 1, \cdots, M$, are scalars and y_k is the kth element of y. We have that q_{ij}^k satisfies $q_{ij}^k = q_{ki}^j = q_{jk}^i$ since $q_{ij}^k = \frac{1}{6}\frac{\partial^3 H}{\partial y_i \partial y_j \partial y_k}$, which is unchanged under any permutation of i, j, k. This provides the symmetry of $\bar{H}(x, y, z)$.

In this paper, we will use the form of Kahan's method in (3.2.4) to prove the energy preservation of the proposed methods.

3.3 Conservation laws for multi-symplectic PDEs

Many PDEs, including all one-dimensional Hamiltonian PDEs, can be written on the multi-symplectic form

$$Kz_t + Lz_x = \nabla S(z), \quad z \in \mathbb{R}^l, \quad (x,t) \in \mathbb{R} \times \mathbb{R}, \tag{3.3.1}$$

where K, $L \in \mathbb{R}^{l\times l}$ are two constant skew-symmetric matrices and $S : \mathbb{R}^l \mapsto \mathbb{R}$ is a scalar-valued function. Following the results about multi-symplectic structure in [3], it can be shown that multi-symplectic PDEs satisfy the following local conservation laws [33]: the multi-symplectic conservation law

$$\partial_t \omega + \partial_x \kappa = 0, \quad \omega = dz \wedge K_+ dz, \quad \kappa = dz \wedge L_+ dz,$$

the local energy conservation law (LECL)

$$E_t + F_x = 0, \quad E = S(z) + z_x^T L_+ z, \quad F = -z_t^T L_+ z, \tag{3.3.2}$$

and the local momentum conservation law (LMCL)

$$I_t + G_x = 0, \quad G = S(z) + z_t^T K_+ z, \quad I = -z_x^T K_+ z,$$

where K_+ and L_+ satisfy

$$K = K_+ - K_+^T, \quad L = L_+ - L_+^T.$$

Decomposition of the matrices is done to make deduction of the conservations laws for energy and momentum more efficient [28, Section 12.3.1].

The multi-symplectic form (3.3.1) can also be generalised to problems in higher dimensional spaces. Consider d spatial dimensions; based on the work by Bridges [3], a multi-symplectic PDE can then be written as

$$Kz_t + \sum_{\alpha=1}^{d} L^\alpha z_{x_\alpha} = \nabla S(z), \quad z \in \mathbb{R}^l, \quad (x,t) \in \mathbb{R}^d \times \mathbb{R}, \tag{3.3.3}$$

where K, $L^\alpha \in \mathbb{R}^{l\times l}$ $(\alpha = 1,\dots,d)$ are constant skew-symmetric matrices and $S : \mathbb{R}^l \to \mathbb{R}$ is a smooth functional. Equation (3.3.3) has the following local

energy conservation law:

$$E_t + \sum_{\alpha=1}^{d} F^{\alpha}_{x_\alpha} = 0, \tag{3.3.4}$$

where $E(z) = S(z) + \sum_{\alpha=1}^{d} z_\alpha^T L_+^\alpha z$, $F^\alpha = -z_t^T L_+^\alpha z$, and L_+^α are splittings of L^α satisfying $L^\alpha = L_+^\alpha - (L_+^\alpha)^T$.

Say we have (3.3.3) defined on the spatial domain $\Omega \in \mathbb{R}^d$ with periodic boundary conditions. Integrating over the domain Ω on both sides of the equation (3.3.4) and using the periodic boundary condition then leads to the global energy conservation law for the multi-symplectic PDEs,

$$\frac{d}{dt}\mathcal{E}(z) = 0, \tag{3.3.5}$$

where $\mathcal{E}(z) = \int_\Omega E(z) d\Omega$.

Example 3.1. ***Korteweg–de Vries equation.*** *Consider the KdV equation for modeling shallow water waves,*

$$u_t + \eta u u_x + \gamma^2 u_{xxx} = 0, \tag{3.3.6}$$

where $\eta, \gamma \in \mathbb{R}$. Introducing the potential $\phi_x = u$, momenta $v = \gamma u_x$ and the variable $w = \gamma v_x \phi_t + \frac{\gamma^2 u^2}{2}$ by the covariant Legendre transform from the Lagrangian, we obtain

$$\begin{aligned} \frac{1}{2} u_t + w_x &= 0, \\ -\frac{1}{2}\phi_t - \gamma v_x &= -w + \frac{\eta}{2} u^2, \\ \gamma u_x &= v, \\ -\phi_x &= -u, \end{aligned} \tag{3.3.7}$$

from which we find the multi-symplectic formulation (3.3.1) *for the KdV equation with $z = (\phi, u, v, w)^T$, the Hamiltonian $S(z) = \frac{v^2}{2} - uw + \frac{\eta u^3}{6}$, and*

$$K = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad L = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -\gamma & 0 \\ 0 & \gamma & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}.$$

As for the conservation laws, there are many choices of K_+ and L_+, for example $K_+ = \frac{K}{2}, L_+ = \frac{L}{2}$, or K_+ and L_+ being the upper triangular parts of K and L, respectively.

Example 3.2. ***Zakharov–Kuznetsov equation.*** *Zakharov and Kuznetsov intro-*

duced in [39] a (2+1)-dimensional generalisation of the KdV equation which includes weak transverse variation,

$$u_t + uu_x + u_{xxx} + u_{xyy} = 0. \tag{3.3.8}$$

A multi-symplectification of this leads to a system (3.3.3) for two spatial dimensions,

$$Kz_t + L^1 z_x + L^2 z_y = \nabla S(z), \quad z \in \mathbb{R}^6, \quad (x, y, t) \in \mathbb{R}^2 \times \mathbb{R}. \tag{3.3.9}$$

Following [4], we have that (3.3.8) *is equivalent to a system of first-order PDEs,*

$$\begin{aligned} \phi_x &= u, \\ \frac{1}{2}\phi_t + v_x + w_y &= p - \frac{1}{2}u^2, \\ w_x - v_y &= 0, \\ -\frac{1}{2}u_t - p_x &= 0, \\ -u_x + q_y &= -v, \\ -q_x - u_y &= -w, \end{aligned} \tag{3.3.10}$$

which is (3.3.9) *with* $z = (p, u, q, \phi, v, w)^T$*, the Hamiltonian* $S(z) = up - \frac{1}{2}(v^2 + w^2) - \frac{1}{6}u^3$*, and the skew-symmetric matrices*

$$K = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad L^1 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}, \quad L^2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

3.4 New linearly implicit energy-preserving schemes

In [20], Gong, Cai and Wang present a scheme that preserves the local energy conservation law (3.3.2) of a one-dimensional multi-symplectic PDE, obtained by applying the midpoint rule in space and the averaged vetor field (AVF) method in time. They also present schemes that preserve the global energy, but not (3.3.2), obtained by considering spatial discretizations that preserve the skew-symmetric property of the difference operator ∂_x. We build on their work by considering Kahan's method for the discretization in time, ensuring linearly implicit schemes and also energy preservation.

To introduce our new schemes, we begin with some basic difference opera-

tors:

$$\delta_t v_j^n := \frac{v_j^{n+1} - v_j^n}{\Delta t}, \qquad \delta_x v_j^n := \frac{v_{j+1}^n - v_j^n}{\Delta x}$$
$$\mu_t v_j^n := \frac{v_j^{n+1} + v_j^n}{2}, \qquad \mu_x v_j^n := \frac{v_{j+1}^n + v_j^n}{2}.$$

The operators satisfy the following properties [36]:

1. All the operators commute with each other, e.g.
   $\delta_t\delta_x v_j^n = \delta_x\delta_t v_j^n, \quad \delta_t\mu_x v_j^n = \mu_x\delta_t v_j^n, \quad \mu_t\delta_x v_j^n = \delta_x\mu_t v_j^n.$
2. They satisfy the discrete Leibniz rule

   $$\delta_t(uv)_j^n = (\varepsilon u_j^{n+1} + (1-\varepsilon)u_j^n)\delta_t v_j^n + \delta_t u_j^n((1-\varepsilon)v_j^{n+1} + \varepsilon v_j^n), \quad 0 \le \varepsilon \le 1.$$

   Specifically,

   $$\delta_t(uv)_j^n = u_j^n\delta_t v_j^n + \delta_t u_j^n v_j^{n+1}, \qquad \text{for} \quad \varepsilon = 0,$$
   $$\delta_t(uv)_j^n = \mu_t u_j^n\delta_t v_j^n + \delta_t u_j^n\mu_t v_j^n, \quad \text{for} \quad \varepsilon = \frac{1}{2},$$
   $$\delta_t(uv)_j^n = u_j^{n+1}\delta_t v_j^n + \delta_t u_j^n v_j^n, \qquad \text{for} \quad \varepsilon = 1.$$

One can obtain a series of similar commutative equations and discrete Leibniz rules that are not presented here, but which are also crucial in the proofs of the preservation properties of the schemes to be introduced in the remainder of this section.

3.4.1 A local energy-preserving scheme for multi-symplectic PDEs

In this section, we apply the midpoint rule in space and Kahan's method in time to construct a local energy-preserving method for multi-symplectic PDEs. Introducing the concept by first considering the one-dimensional system (3.3.1), we apply the midpoint rule in space to get

$$K\partial_t\mu_x z_j + L\delta_x z_j = \nabla S(\mu_x z_j), \quad j = 0, \dots, M-1.$$

Then applying Kahan's method gives us the linearly implicit local energy-preserving (LILEP) scheme

$$K\delta_t\mu_x z_j^n + L\delta_x\mu_t z_j^n = 3\frac{\partial\bar{S}}{\partial x}\Big|_{(\mu_x z_j^n, \mu_x z_j^{n+1})}. \tag{3.4.1}$$

Here we consider S of the form $S(y) = y^T Q(y)y + y^T By + c^T y + d$, as in (3.2.5), and accordingly $\bar{S}(x, y, z)$ of the form (3.2.6).

Theorem 3.1. *The scheme* (3.4.1) *satisfies the discrete local energy conservation law*

$$\delta_t(\bar{E}_L)_j^n + \delta_x(\bar{F}_L)_j^n = 0, \tag{3.4.2}$$

where

$$\begin{aligned}(\bar{E}_L)_j^n &= \bar{S}(\mu_x z_j^n, \mu_x z_j^n, \mu_x z_j^{n+1}) \\ &\quad + \frac{1}{3}(\delta_x z_j^n)^T L_+ \mu_x z_j^n + \frac{1}{3}(\delta_x z_j^n)^T L_+ \mu_x z_j^{n+1} + \frac{1}{3}(\delta_x z_j^{n+1})^T L_+ \mu_x z_j^n, \\ (\bar{F}_L)_j^n &= -\frac{1}{3}(\delta_t z_j^n)^T L_+ \mu_t z_j^n - \frac{1}{3}(\delta_t z_j^n)^T L_+ \mu_t z_j^{n+1} - \frac{1}{3}(\delta_t z_j^{n+1})^T L_+ \mu_t z_j^n.\end{aligned} \tag{3.4.3}$$

Proof. Taking the inner product with $\frac{1}{3}\delta_t \mu_x z_j^n$ on both sides of (3.4.1) and using the skew-symmetry of matrix K, we have

$$\frac{1}{3}(\delta_t \mu_x z_j^n)^T L \delta_x \mu_t z_j^n = (\delta_t \mu_x z_j^n)^T \frac{\partial \bar{S}}{\partial x}\bigg|_{(\mu_x z_j^n, \mu_x z_j^{n+1})}. \tag{3.4.4}$$

Taking the inner product with $\frac{1}{3}\delta_t \mu_x z_j^{n+1}$ on both sides of (3.4.1), we get

$$\frac{1}{3}(\delta_t \mu_x z_j^{n+1})^T K \delta_t \mu_x z_j^n + \frac{1}{3}(\delta_t \mu_x z_j^{n+1})^T L \delta_x \mu_t z_j^n = (\delta_t \mu_x z_j^{n+1})^T \frac{\partial \bar{S}}{\partial x}\bigg|_{(\mu_x z_j^n, \mu_x z_j^{n+1})}. \tag{3.4.5}$$

Taking the inner product with $\frac{1}{3}\delta_t \mu_x z_j^n$ on both sides of the scheme (3.4.1) for the next time step, we get

$$\frac{1}{3}(\delta_t \mu_x z_j^n)^T K \delta_t \mu_x z_j^{n+1} + \frac{1}{3}(\delta_t \mu_x z_j^n)^T L \delta_x \mu_t z_j^{n+1} = (\delta_t \mu_x z_j^n)^T \frac{\partial \bar{S}}{\partial x}\bigg|_{(\mu_x z_j^{n+1}, \mu_x z_j^{n+2})}. \tag{3.4.6}$$

Adding equations (3.4.4), (3.4.5) and (3.4.6) and using the skew-symmetry of

matrix K, we obtain

$$\begin{aligned}
&\frac{1}{3}\Big((\delta_t\mu_x z_j^n)^T L\delta_x\mu_t z_j^n + (\delta_t\mu_x z_j^{n+1})^T L\delta_x\mu_t z_j^n + (\delta_t\mu_x z_j^n)^T L\delta_x\mu_t z_j^{n+1}\Big)\\
&= (\delta_t\mu_x z_j^n)^T \frac{\partial \bar{S}}{\partial x}\,|_{(\mu_x z_j^n,\mu_x z_j^{n+1})} + (\delta_t\mu_x z_j^{n+1})^T \frac{\partial \bar{S}}{\partial x}\Big|_{(\mu_x z_j^n,\mu_x z_j^{n+1})}\\
&\quad + (\delta_t\mu_x z_j^n)^T \frac{\partial \bar{S}}{\partial x}\Big|_{(\mu_x z_j^{n+1},\mu_x z_j^{n+2})},\\
&= \frac{1}{\Delta t}\big(\bar{S}(\mu_x z_j^{n+1},\mu_x z_j^{n+1},\mu_x z_j^{n+2}) - \bar{S}(\mu_x z_j^n,\mu_x z_j^n,\mu_x z_j^{n+1})\big),\\
&= \delta_t\bar{S}(\mu_x z_j^n,\mu_x z_j^n,\mu_x z_j^{n+1}).
\end{aligned} \tag{3.4.7}$$

On the other hand, using the aforementioned commutative laws and discrete Leibniz rules for the operators, we can deduce

$$\begin{aligned}
\delta_t((\delta_x z_j^n)^T L_+\mu_x z_j^n) &= (\delta_t\delta_x z_j^n)^T L_+\mu_t\mu_x z_j^n + (\delta_x\mu_t z_j^n)^T L_+\delta_t\mu_x z_j^n,\\
\delta_x((\delta_t z_j^n)^T L_+\mu_t z_j^n) &= (\delta_t\delta_x z_j^n)^T L_+\mu_t\mu_x z_j^n + (\delta_t\mu_x z_j^n)^T L_+\delta_x\mu_t z_j^n,\\
\delta_t((\delta_x z_j^{n+1})^T L_+\mu_x z_j^n) &= (\delta_t\delta_x z_j^{n+1})^T L_+\mu_t\mu_x z_j^n + (\delta_x\mu_t z_j^{n+1})^T L_+\delta_t\mu_x z_j^n,\\
\delta_x((\delta_t z_j^{n+1})^T L_+\mu_t z_j^n) &= (\delta_t\delta_x z_j^{n+1})^T L_+\mu_t\mu_x z_j^n + (\delta_t\mu_x z_j^{n+1})^T L_+\delta_x\mu_t z_j^n,\\
\delta_t((\delta_x z_j^n)^T L_+\mu_x z_j^{n+1}) &= (\delta_t\delta_x z_j^n)^T L_+\mu_t\mu_x z_j^{n+1} + (\delta_x\mu_t z_j^n)^T L_+\delta_t\mu_x z_j^{n+1},\\
\delta_x((\delta_t z_j^n)^T L_+\mu_t z_j^{n+1}) &= (\delta_t\delta_x z_j^n)^T L_+\mu_t\mu_x z_j^{n+1} + (\delta_t\mu_x z_j^n)^T L_+\delta_x\mu_t z_j^{n+1}.
\end{aligned} \tag{3.4.8}$$

Using the above relations (3.4.8), the fact that $L = L_+ - L_+^T$ and the result (3.4.7), we obtain

$$\begin{aligned}
\delta_t E_j^n + \delta_x F_j^n =& \delta_t\bar{S}(\mu_x z_j^n,\mu_x z_j^n,\mu_x z_j^{n+1})\\
&+ \frac{1}{3}\big(\delta_t((\delta_x z_j^n)^T L_+\mu_x z_j^n) + \delta_t((\delta_x z_j^n)^T L_+\mu_x z_j^{n+1})\\
&+ \delta_t((\delta_x z_j^{n+1})^T L_+\mu_x z_j^n)\big) - \frac{1}{3}\big(\delta_x((\delta_t z_j^n)^T L_+\mu_t z_j^n)\\
&+ \delta_x((\delta_t z_j^n)^T L_+\mu_t z_j^{n+1}) + \delta_x((\delta_t z_j^{n+1})^T L_+\mu_t z_j^n)\big)\\
=& \delta_t\bar{S}(\mu_x z_j^n,\mu_x z_j^n,\mu_x z_j^{n+1}) - \frac{1}{3}\big((\delta_t\mu_x z_j^n)^T L\delta_x\mu_t z_j^n\\
&+ (\delta_t\mu_x z_j^{n+1})^T L\delta_x\mu_t z_j^n + (\delta_t\mu_x z_j^n)^T L\delta_x\mu_t z_j^{n+1}\big)\\
=& 0.
\end{aligned}$$

□

Corollary 1. For periodic boundary conditions $z(x+P,t) = z(x,t)$, the scheme (3.4.1) satisfies the discrete global energy conservation law

$$\bar{\mathcal{E}}_L^{n+1} = \bar{\mathcal{E}}_L^n, \quad \bar{\mathcal{E}}_L^n := \Delta x \sum_{j=0}^{M-1} (\bar{E}_L)_j^n, \tag{3.4.9}$$

where $\Delta x = P/M$ and $(\bar{E}_L)_j^n$ is given by (3.4.3).

Proof. With periodic boundary conditions, we get $\sum_{j=0}^{M-1} \delta_x (\bar{F}_L)_j^n = 0$, and thus (3.4.9) follows from (3.4.2). □

The polarised global energy $\bar{\mathcal{E}}_L^n$ may be considered as a function of the solution in time step n only, similarly to the modified Hamiltonian defined in Proposition 3 of [10].

Proposition 3.1. *With the solution z^{n+1} found from z^n by* (3.4.1)*, the discrete global energy $\bar{\mathcal{E}}_L^n$ of* (3.4.9) *satisfies*

$$\bar{\mathcal{E}}_L^n = \mathcal{E}_L^n + \Delta x \sum_{j=0}^{M-1} \frac{1}{3} (\nabla E_L(z_j^n))^T (z_j^{n+1} - z_j^n), \tag{3.4.10}$$

where

$$\mathcal{E}_L^n := \Delta x \sum_{j=0}^{M-1} E_L(z_j^n), \quad E_L(z_j^n) := S(\mu_x z_j^n) + (\delta_x z_j^n)^T L_+ \mu_x z_j^n, \tag{3.4.11}$$

while $z_j^{n+1} - z_j^n$ satisfies

$$R_L(z_j^n)(z_j^{n+1} - z_j^n) = \Delta t\, g_L(z_j^n), \tag{3.4.12}$$

with $g_L(z_j^n) = \nabla S(\mu_x z_j^n) - L\delta_x z_j^n$ and $R_L(z_j^n) = K\mu_x - \frac{\Delta t}{2} \nabla g_L(z_j^n)$.

Proof. Note that

$$\begin{aligned} \bar{S}(\mu_x z_j^n, \mu_x z_j^n, \mu_x z_j^{n+1}) &= S(\mu_x z_j^n) + \frac{1}{3} \nabla S(\mu_x z_j^n)^T (\mu_x z_j^{n+1} - \mu_x z_j^n) \\ &= S(\mu_x z_j^n) + \frac{1}{3} \nabla_{z_j^n} (S(\mu_x z_j^n))^T (z_j^{n+1} - z_j^n), \end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{3}(\delta_x z_j^n)^T L_+ \mu_x z_j^n + \frac{1}{3}(\delta_x z_j^n)^T L_+ \mu_x z_j^{n+1} + \frac{1}{3}(\delta_x z_j^{n+1})^T L_+ \mu_x z_j^n \\
&= (\delta_x z_j^n)^T L_+ \mu_x z_j^n + \frac{1}{3}\big((\delta_x z_j^n)^T L_+ \mu_x (z_j^{n+1} - z_j^n) + (\delta_x (z_j^{n+1} - z_j^n))^T L_+ \mu_x z_j^n\big) \\
&= (\delta_x z_j^n)^T L_+ \mu_x z_j^n + \frac{1}{3}\big((\mu_x z_j^n)^T L_x^T \delta_x + (\delta_x z_j^n)^T L_x \mu_x\big)(z_j^{n+1} - z_j^n) \\
&= (\delta_x z_j^n)^T L_+ \mu_x z_j^n + \frac{1}{3}\Big(\nabla_{z_j^n}\big((\delta_x z_j^n)^T L_+ \mu_x z_j^n\big)\Big)^T (z_j^{n+1} - z_j^n).
\end{aligned}$$

Inserting this in (3.4.3), we get (3.4.10) from (3.4.9). Furthermore, observing that

$$3\frac{\partial \bar{S}}{\partial x}\bigg|_{(\mu_x z_j^n, \mu_x z_j^{n+1})} = \nabla S(\mu_x z_j^n) + \frac{1}{2}\nabla^2 S(\mu_x z_j^n)(\mu_x z_j^{n+1} - \mu_x z_j^n),$$

we may rewrite (3.4.1) as

$$\Big(K\mu_x + \frac{\Delta t}{2} L\delta_x - \frac{\Delta t}{2}\nabla^2 S(\mu_x z_j^n)\mu_x\Big)(z_j^{n+1} - z_j^n) = \Delta t\big(\nabla S(\mu_x z_j^n) - L\delta_x z_j^n\big),$$

which is (3.4.12). □

Note that (3.4.11) is the discrete energy preserved by the fully implicit local energy-preserving method of [20]. Also, for methods based on the multi-symplectic structure, instead of solving for z directly, the normal procedure is to eliminate the auxiliary variables from the scheme and get an equation for one variable u. Therefore we do not give an explicit expression for the modified energy in z^n. However, in Section 3.5, we present an explicit expression for the modified energy in u^n when our scheme is applied to the KdV equation.

The results about the energy conservation for the LILEP method applied to one-dimensional multi-symplectic PDEs can be generalised to problems in spatial dimensions of any finite degree. Consider for example a 2-dimensional multi-symplectic PDE

$$Kz_t + L^1 z_x + L^2 z_y = \nabla S(z), \quad z \in \mathbb{R}^l, \quad (x, y, t) \in \mathbb{R}^3, \tag{3.4.13}$$

for which we have the following corollary. This is presented without its proof, which is rather technical but similar to the proof of Theorem 3.1.

Corollary 2. The scheme obtained by applying the midpoint rule in space and

Kahan's method in time to equation (3.4.13),

$$K\delta_t\mu_x\mu_y z_{j,k}^n + L^1\delta_x\mu_t\mu_y z_{j,k}^n + L^2\delta_y\mu_t\mu_x z_{j,k}^n = 3\frac{\partial\bar{S}}{\partial x}\Big|_{(\mu_x\mu_y z_{j,k}^n, \mu_x\mu_y z_{j,k}^{n+1})}, \tag{3.4.14}$$

where $j = 0,\dots,M_x-1$ and $k = 0,\dots,M_y-1$, satisfies the discrete local energy conservation law

$$\delta_t(\bar{E}_L)_{j,k}^n + \delta_x(\bar{F}_L^1)_{j,k}^n + \delta_y(\bar{F}_L^2)_{j,k}^n = 0,$$

where

$$\begin{aligned}
(\bar{E}_L)_{j,k}^n =& \bar{S}(\mu_x\mu_y z_{j,k}^n, \mu_x\mu_y z_{j,k}^n, \mu_x\mu_y z_{j,k}^{n+1}) \\
&+ \frac{1}{3}(\delta_x\mu_y z_{j,k}^n)^T L_+^1\mu_x\mu_y z_{j,k}^n + \frac{1}{3}(\delta_x\mu_y z_{j,k}^n)^T L_+^1\mu_x\mu_y z_{j,k}^{n+1} \\
&+ \frac{1}{3}(\delta_x\mu_y z_{j,k}^{n+1})^T L_+^1\mu_x\mu_y z_{j,k}^n + \frac{1}{3}(\delta_y\mu_x z_{j,k}^n)^T L_+^2\mu_x\mu_y z_{j,k}^n \\
&+ \frac{1}{3}(\delta_y\mu_x z_{j,k}^n)^T L_+^2\mu_x\mu_y z_{j,k}^{n+1} + \frac{1}{3}(\delta_y\mu_x z_{j,k}^{n+1})^T L_+^2\mu_x\mu_y z_{j,k}^n, \\
(\bar{F}_L^1)_{j,k}^n =& -\frac{1}{3}(\delta_t\mu_y z_{j,k}^n)^T L_+^1\mu_t\mu_y z_{j,k}^n - \frac{1}{3}(\delta_t\mu_y z_{j,k}^n)^T L_+^1\mu_t\mu_y z_{j,k}^{n+1} \\
&- \frac{1}{3}(\delta_t\mu_y z_{j,k}^{n+1})^T L_+^1\mu_t\mu_y z_{j,k}^n, \\
(\bar{F}_L^2)_{j,k}^n =& -\frac{1}{3}(\delta_t\mu_x z_{j,k}^n)^T L_+^2\mu_t\mu_x z_{j,k}^n - \frac{1}{3}(\delta_t\mu_x z_{j,k}^n)^T L_+^2\mu_t\mu_x z_{j,k}^{n+1} \\
&- \frac{1}{3}(\delta_t\mu_x z_{j,k}^{n+1})^T L_+^2\mu_t\mu_x z_{j,k}^n.
\end{aligned}$$

3.4.2 Global energy-preserving methods for multi-symplectic PDEs

As shown in Section 3.3, Hamiltonian PDEs of the form (3.3.1) with periodic boundary conditions have global energy conservation which can be deduced from the local conservation law. On the other hand, the local conservation law is not inherent in the global conservation law. In this section, we will focus on giving a systematic method that preserves the global energy conservation law directly. We discretize ∂_x with an antisymmetric differential matrix D and get the semi-discretized variant of (3.3.1),

$$K\partial_t z_j + L(Dz)_j = \nabla S(z_j), \quad j = 0,1,\dots,M-1, \tag{3.4.15}$$

where $z := (z_0, z_1, \dots, z_{M-1})^T \in \mathbb{R}^{M\times l}$ and $(Dz)_j = \sum_{k=0}^{M-1} D_{j,k}z_k$. We then apply Kahan's method to (3.4.15) and obtain the linearly implicit global energy-

preserving (LIGEP) scheme

$$K\delta_t z_j^n + L(D\mu_t z^n)_j = 3\frac{\partial \bar{S}}{\partial x}\Big|_{(z_j^n, z_j^{n+1})}. \tag{3.4.16}$$

Define the polarised energy density by

$$\bar{E}_j^n = \bar{S}(z_j^n, z_j^n, z_j^{n+1}) + \frac{1}{3}(Dz^n)_j^T L_+ z_j^n + \frac{1}{3}(Dz^n)_j^T L_+ z_j^{n+1} + \frac{1}{3}(Dz^{n+1})_j^T L_+ z_j^n, \tag{3.4.17}$$

and we get the following result.

Theorem 3.2. *For periodic boundary conditions $z(x+P,t) = z(x,t)$, the scheme* (3.4.16) *satisfies the discrete global energy conservation law*

$$\bar{\mathcal{E}}^{n+1} = \bar{\mathcal{E}}^n, \quad \bar{\mathcal{E}}^n := \Delta x \sum_{j=0}^{M-1} \bar{E}_j^n, \quad \Delta x = P/M. \tag{3.4.18}$$

Proof. Taking the inner product with $\frac{1}{3}\delta_t z_j^n$ on both sides of equation (3.4.16) and using the skew-symmetry of the matrix K, we get

$$\frac{1}{3}(\delta_t z_j^n)^T L(D\mu_t z^n)_j = (\delta_t z_j^n)^T \frac{\partial \bar{S}}{\partial x}\Big|_{(z_j^n, z_j^{n+1})}. \tag{3.4.19}$$

Taking the inner product with $\frac{1}{3}\delta_t z_j^{n+1}$ on both sides of (3.4.16), we get

$$\frac{1}{3}(\delta_t z_j^{n+1})^T K\delta_t z_j^n + \frac{1}{3}(\delta_t z_j^{n+1})^T L(D\mu_t z^n)_j = (\delta_t z_j^{n+1})^T \frac{\partial \bar{S}}{\partial x}\Big|_{(z_j^n, z_j^{n+1})}. \tag{3.4.20}$$

Furthermore, taking the inner product with $\frac{1}{3}\delta_t z_j^n$ on both sides of (3.4.16) for the next time step, we have

$$\frac{1}{3}(\delta_t z_j^n)^T K\delta_t z_j^{n+1} + \frac{1}{3}(\delta_t z_j^n)^T L(D\mu_t z^{n+1})_j = (\delta_t z_j^n)^T \frac{\partial \bar{S}}{\partial x}\Big|_{(z_j^{n+1}, z_j^{n+2})}. \tag{3.4.21}$$

Adding equations (3.4.19), (3.4.20) and (3.4.21), we get

$$\begin{aligned}\frac{1}{3}\big((\delta_t z_j^n)^T L(D\mu_t z^n)_j + (\delta_t z_j^n)^T L(D\mu_t z^{n+1})_j \\ + (\delta_t z_j^{n+1})^T L(D\mu_t z^n)_j\big) = \delta_t \bar{S}(z_j^n, z_j^n, z_j^{n+1}).\end{aligned} \tag{3.4.22}$$

By using the commutative laws and discrete Leibniz rules,

$$\begin{aligned}
\delta_t((Dz^n)_j^T L_+ z_j^n) &= (D\delta_t z^n)_j^T L_+ \mu_t z_j^n + (D\mu_t z^n)_j L_+ \delta_t z_j^n, \\
\delta_t((Dz^n)_j^T L_+ z_j^{n+1}) &= (D\delta_t z^n)_j^T L_+ \mu_t z_j^{n+1} + (D\mu_t z^n)_j L_+ \delta_t z_j^{n+1}, \\
\delta_t((Dz^{n+1})_j^T L_+ z_j^n) &= (D\delta_t z^{n+1})_j^T L_+ \mu_t z_j^n + (D\mu_t z^{n+1})_j L_+ \delta_t z_j^n.
\end{aligned} \tag{3.4.23}$$

Based on the above equations (3.4.22) and (3.4.23), we obtain

$$\begin{aligned}
&\delta_t E_j^n \\
&= \delta_t \bar{S}(z_j^n, z_j^n, z_j^{n+1}) + \frac{1}{3}\big(\delta_t((Dz^n)_j^T L_+ z_j^n) + (Dz^n)_j^T L_+ z_j^{n+1} + (Dz^{n+1})_j^T L_+ z_j^n\big) \\
&= \frac{1}{3}\big((\delta_t z_j^n)^T L_+ (D\mu_t z^n)_j + (D\delta_t z^n)_j^T L_+ \mu_t z_j^n\big) \\
&\quad + \frac{1}{3}\big((\delta_t z_j^{n+1})^T L_+ (D\mu_t z^n)_j + (D\delta_t z^{n+1})_j^T L_+ \mu_t z_j^n\big) \\
&\quad + \frac{1}{3}\big((\delta_t z_j^n)^T L_+ (D\mu_t z^{n+1})_j + (D\delta_t z^n)_j^T L_+ \mu_t z_j^{n+1}\big) \\
&= \sum_{k=0}^{N-1} (D)_{j,k} G_{j,k},
\end{aligned}$$

where

$$\begin{aligned}
G_{j,k} := &\frac{1}{3}\big((\delta_t z^n)_j^T L_+ \mu_t z_L^n + (\delta_t z^n)_L^T L_+ \mu_t z_j^n\big) \\
&+ \frac{1}{3}\big((\delta_t z^{n+1})_j^T L_+ \mu_t z_L^n + (\delta_t z^{n+1})_L^T L_+ \mu_t z_j^n\big) \\
&+ \frac{1}{3}\big((\delta_t z^n)_j^T L_+ \mu_t z_L^{n+1} + (\delta_t z^n)_L^T L_+ \mu_t z_j^{n+1}\big).
\end{aligned}$$

Since D is skew-symmetric and $G_{j,k} = G_{k,j}$, we get

$$\sum_{j=0}^{M-1} \delta_t \bar{E}_j^n = 0,$$

which implies that the discrete global energy conservation law $\bar{\mathcal{E}}^{n+1} = \bar{\mathcal{E}}^n$ is satisfied. □

The polarised energy $\bar{\mathcal{E}}$ preserved by (3.4.16) may also be expressed as a modification of the discrete energy

$$\mathcal{E}^n := \Delta x \sum_{j=0}^{M-1} E(z_j^n), \quad E(z_j^n) = S(z_j^n) + (Dz^n)_j^T L_+ z_j^n, \tag{3.4.24}$$

which is preserved by the fully implicit global energy-preserving scheme of [20]. The proof of the following proposition is similar to the proof of Proposition 3.1, and hence omitted.

Proposition 3.2. *If the solution z^{n+1} is found from z^n by* (3.4.16)*, the discrete global energy $\bar{\mathcal{E}}^n$ of* (3.4.18) *satisfies*

$$\bar{\mathcal{E}}^n = \mathcal{E}^n + \Delta x \sum_{j=0}^{M-1} \frac{1}{3} (\nabla E(z_j^n))^T (z_j^{n+1} - z_j^n),$$

and $z_j^{n+1} - z_j^n$ satisfies

$$R(z_j^n)(z_j^{n+1} - z_j^n) = \Delta t g(z_j^n),$$

where $g(z_j^n) = \nabla S(z_j^n) - L(Dz)_j^n$ and $R(z_j^n) = K + \frac{\Delta t}{2} \nabla g(z_j^n)$.

The above global conservation results can be generalised to multi-symplectic formulations in higher spatial dimensions, as demonstrated by the following corollary for the two-dimensional case, whose omitted proof is in the same vein as the proof of Theorem 3.2.

Corollary 3. Discretizing ∂_x and ∂_y by skew-symmetric differential matrices D_x and D_y in equation (3.4.13) and then applying Kahan's method to the semi-discrete system, one obtains the linearly implicit global energy-preserving (LIGEP) scheme

$$K\delta_t z_{j,k}^n + L^1 \mu_t (D_x z^n)_{j,k} + L^2 \mu_t (D_y z^n)_{j,k} = 3 \frac{\partial \bar{S}}{\partial x}\bigg|_{(z_{j,k}^n, z_{j,k}^{n+1})}, \tag{3.4.25}$$

where $j = 0, \ldots, M_x - 1$ and $k = 0, \ldots, M_y - 1$. For periodic boundary conditions $z(x + P_x, y, t) = z(x, y, t)$, $z(x, y + P_y, t) = z(x, y, t)$, the scheme (3.4.25) satisfies the discrete global energy conservation law

$$\bar{\mathcal{E}}^{n+1} = \bar{\mathcal{E}}^n,$$

where

$$\begin{aligned}
\bar{\mathcal{E}}^n &:= \Delta x \Delta y \sum_{j=0}^{M_x-1} \sum_{k=0}^{M_y-1} \bar{E}_{j,k}^n, \quad \Delta x = P_x / M_x, \quad \Delta y = P_y / M_y, \\
\bar{E}_{j,k}^n &= \bar{S}(z_{j,k}^n, z_{j,k}^n, z_{j,k}^{n+1}) \\
&\quad + \frac{1}{3} (D_x z^n)_{j,k}^T L_+^1 z_{j,k}^n + \frac{1}{3} (D_x z^n)_{j,k}^T L_+^1 z_{j,k}^{n+1} + \frac{1}{3} (D_x z^{n+1})_{j,k}^T L_+^1 z_{j,k}^n, \\
&\quad + \frac{1}{3} (D_y z^n)_{j,k}^T L_+^2 z_{j,k}^n + \frac{1}{3} (D_y z^n)_{j,k}^T L_+^2 z_{j,k}^{n+1} + \frac{1}{3} (D_y z^{n+1})_{j,k}^T L_+^2 z_{j,k}^n.
\end{aligned}$$

3.5 Numerical examples

In this section, we apply our proposed new linearly implicit energy-preserving schemes to the KdV equation and Zakharov–Kuznetsov equation, and compare them with fully implicit schemes. Among our reference methods are the methods introduced in [20], for which the local energy-preserving method is denoted by LEP, and the global energy-preserving method by GEP. For the GEP and LIGEP schemes, two different choices are considered for approximating the spatial derivative: the central difference operator δ_x^c defined by $\delta_x^c v_j^n := \frac{1}{2}(\delta_x v_{j-1}^n + \delta_x v_j^n)$ and the first order Fourier pseudospectral operator [4]. The latter results in the $M \times M$ matrix D, given explicitly by its elements

$$D_{i,j} = \begin{cases} \frac{\pi}{P}(-1)^{i+j}\cot\big(\pi(i-j)/M\big), & \text{if} \quad i \neq j, \\ 0, & \text{if} \quad i = j, \end{cases}$$

evaluated on the domain $\big[0, P\big]$, where we assume M even and periodic boundary conditions [12]. If M is odd, we have instead

$$D_{i,j} = \begin{cases} \frac{\pi}{P}(-1)^{i+j}\cot\big(\pi(i-j)/M\big), & \text{if} \quad |i-j| < M/2, \\ \frac{\pi}{P}(-1)^{i+j}\cot\big(\pi(j-i)/M\big), & \text{if} \quad |i-j| > M/2, \\ 0, & \text{if} \quad i = j. \end{cases}$$

3.5.1 Korteweg–de Vries equation

Consider the multi-symplectic structure of the KdV equation as presented in Example 3.1. Applying the LILEP scheme (3.4.1) to (3.3.7), we obtain

$$\begin{aligned} \frac{1}{2}\delta_t\mu_x u_j^n + \delta_x\mu_t w_j^n &= 0, \\ -\frac{1}{2}\delta_t\mu_x\phi_j^n - \gamma\delta_x\mu_t v_j^n &= -\mu_t\mu_x w_j^n + \frac{\eta}{2}\mu_x u_j^n \mu_x u_j^{n+1}, \\ \gamma\delta_x\mu_t u_j^n &= \mu_t\mu_x v_j^n, \\ \delta_x\mu_t\phi_j^n &= \mu_t\mu_x u_j^n. \end{aligned}$$

By eliminating the auxiliary varibles ϕ, v and w, we see that this is equivalent to

$$\delta_t\mu_t\mu_x^3 u_j^n + \frac{\eta}{2}\delta_x\mu_t\mu_x(\mu_x u_j^n \mu_x u_j^{n+1}) + \gamma^2\delta_x^3\mu_t^2 u_j^n = 0.$$

Omitting the average operator μ_t gives us

$$\delta_t \mu_x^3 u_j^n + \frac{\eta}{2}\delta_x \mu_x (\mu_x u_j^n \mu_x u_j^{n+1}) + \gamma^2 \delta_x^3 \mu_t u_j^n = 0. \tag{3.5.1}$$

The polarised discrete energy preserved by this scheme is

$$\bar{\mathcal{E}}_L^n = \Delta x \sum_{j=0}^{M-1} \Big(-\frac{1}{6}\gamma^2 \big((\delta_x u_j^n)^2 + 2\delta_x u_j^n \delta_x u_j^{n+1}\big) + \frac{1}{6}\eta \big(\mu_x u_j^n\big)^2 \mu_x u_j^{n+1} \Big). \tag{3.5.2}$$

On the other hand, the discrete energy preserved by the LEP method of [20] is

$$\mathcal{E}_L^n = \Delta x \sum_{j=0}^{M-1} \Big(-\frac{1}{2}\gamma^2 (\delta_x u_j^n)^2 + \frac{1}{6}\eta (\mu_x u_j^n)^3 \Big). \tag{3.5.3}$$

By Proposition 3.1 and elimination of the variables ϕ, v and w, (3.5.2) can be expressed as a modification of (3.5.3): we may rewrite (3.5.1) as

$$u_j^{n+1} - u_j^n = -\Delta t \big(\mu_x^3 + \frac{\Delta t}{2}\gamma^2 \delta_x^3 + \frac{\Delta t}{2}\eta \delta_x \mu_x \mathrm{diag}(\mu_x u_n)\mu_x\big)^{-1} \big(\gamma^2 \delta_x^3 u^n + \frac{\eta}{2}\delta_x \mu_x (\mu_x u^n)^2\big),$$

where $(\mu_x u^n)^2$ denotes the element-wise square of $\mu_x u^n$. Inserting this in (3.5.2), we get

$$\begin{aligned}
\bar{\mathcal{E}}_L^n &= \mathcal{E}_L^n - \frac{\Delta t \Delta x}{3}\big(-\gamma^2 \delta_x^T \delta_x u^n + \frac{\eta}{2}\mu_x^T (\mu_x u^n)^2\big)^T \\
&\quad \big(\mu_x^3 + \frac{\Delta t}{2}\gamma^2 \delta_x^3 + \frac{\Delta t}{2}\eta \delta_x \mu_x \mathrm{diag}(\mu_x u^n)\mu_x\big)^{-1} \big(\gamma^2 \delta_x^3 u^n + \frac{\eta}{2}\delta_x \mu_x (\mu_x u^n)^2\big) \\
&= \mathcal{E}_L^n + \frac{\Delta t}{3}(\nabla \mathcal{E}_L^n)^T \big(\mu_x^3 - \frac{\Delta t}{2}\zeta_L'(u^n)\big)^{-1} \zeta_L(u^n),
\end{aligned}$$

with

$$\zeta_L(u^n) = -\gamma^2 \delta_x^3 u^n - \frac{\eta}{2}\delta_x \mu_x (\mu_x u^n)^2,$$

where $\nabla \mathcal{E}_L^n$ means the gradient of $\mathcal{E}_L^n$ with respect to u^n, and $\zeta_L'(u^n)$ denotes the Jacobian matrix of $\zeta_L(u^n)$.

Similarly for the LIGEP method (3.4.16); applying it to the the multi-symplectic KdV equations (3.3.7) and eliminating the auxiliary varibles ϕ, v and w, we obtain

$$\delta_t \mu_t u_j^n + \frac{\eta}{2}\mu_t (D(u^n u^{n+1}))_j + \gamma^2 \mu_t^2 (D^3 u^n)_j = 0,$$

where $u^n u^{n+1}$ denotes element-wise multiplication of the vectors. Omitting

the average operator μ_t, we get

$$\delta_t u_j^n + \frac{\eta}{2}(D(u^n u^{n+1}))_j + \gamma^2 \mu_t (D^3 u^n)_j = 0. \tag{3.5.4}$$

The discrete global energy preserved by the GEP method is

$$\mathcal{E}^n = \Delta x \sum_{j=0}^{M-1} \Big(-\frac{1}{2}\gamma^2 (Du^n)_j^2 + \frac{1}{6}\eta (u_j^n)^3 \Big), \tag{3.5.5}$$

while the polarised discrete energy preserved by (3.5.4) is

$$\begin{aligned} \bar{\mathcal{E}}^n &= \Delta x \sum_{j=0}^{M-1} \Big(-\frac{1}{6}\gamma^2 \big((Du^n)_j^2 + 2(Du^n)_j (Du^{n+1})_j \big) + \frac{1}{6}\eta \big(u_j^n \big)^2 u_j^{n+1} \Big) \\ &= \mathcal{E}^n - \frac{\Delta t \Delta x}{3} \big(-\gamma^2 D^T D u^n + \frac{\eta}{2}(u^n)^2 \big)^T \\ &\quad \big(I + \frac{\Delta t}{2}\gamma^2 D^3 + \frac{\Delta t}{2}\eta D \operatorname{diag}(u^n) \big)^{-1} \big(\gamma^2 D^3 u^n + \frac{\eta}{2} D(u^n)^2 \big) \\ &= \mathcal{E}^n + \frac{\Delta t}{3} (\nabla \mathcal{E}^n)^T \big(I - \frac{\Delta t}{2} \zeta'(u^n) \big)^{-1} \zeta(u^n), \end{aligned} \tag{3.5.6}$$

where $\zeta(u^n) = -\gamma^2 D^3 u^n - \frac{\eta}{2} D(u^n)^2$.

Test problem 1

In the first numerical experiment, we consider the problem introduced in [38] and then used by Zhao and Qin [40] and Ascher and McLachlan [1] to test various symplectic and multi-symplectic schemes: the KdV equation with $\gamma = 0.022$, $\eta = 1$, and initial value

$$u_0(x) = \cos(\pi x),$$

with $x \in [0, P]$, $P = 2$. This problem is also considered in Example 3 of [20], where it is solved by implicit schemes that preserve local and/or global energy. As observed by Gong et al., the global energy-preserving scheme (GEP) with the central difference operator used to approximate ∂_x gives unsatisfactory results for this problem; we observed that the same is true for the LIGEP scheme. Therefore, the Fourier pseudospectral operator is used to approximate the spatial derivatives in the GEP and LIGEP schemes. This seems to result in more accurate solutions than the LEP and LILEP schemes for the same number of discretization points, but at a considerably higher computational cost, as seen from Table 3.1. From Figure 3.1, we can conclude that our linearly implicit schemes give results close to their fully implicit counterparts introduced in [20],

and that the different schemes converge to the same solution. Here and in the following test problem, we have solved the fully implicit schemes in each step by Newton's method until $\|F(u^n)\|_2 < 10^{-10}$.

M	200	400	600	800	1000	1500	2000
LEP	1.87	3.16	4.43	11.18	13.81	21.53	28.54
LILEP	4.24e-1	7.40e-1	1.07	1.39	1.73	2.67	3.58
GEP	12.29	78.11	242.48	1016.57	1888.69	5793.18	13154.20
LIGEP	2.16	11.15	33.50	73.94	136.93	398.53	894.52

Table 3.1: Computational time, in seconds, for finding the solution of the first test problem at time $t = 5$ by a temporal step size $\Delta t = 0.005$ and various number of discretization points in space, M.

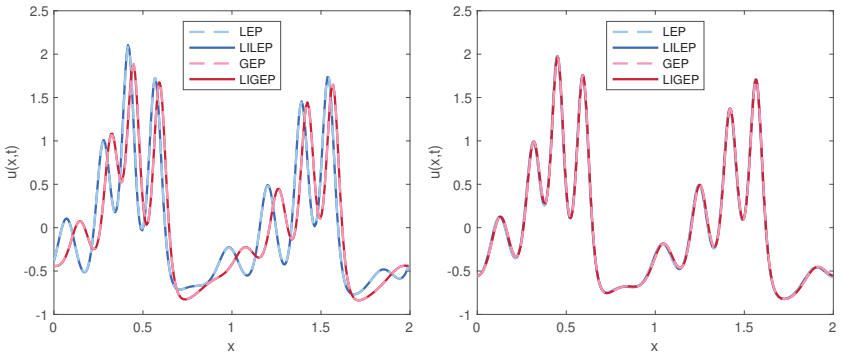
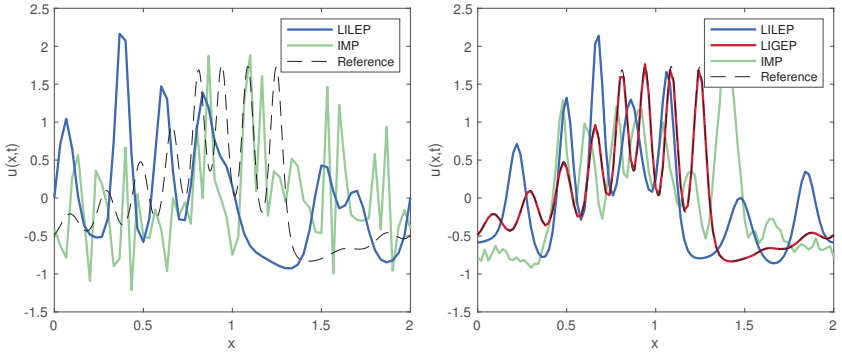


Figure 3.1: Solution of test problem 1 at time $t = 5$ by our schemes and the fully implicit schemes of Gong et al. *Left:* $M = 250$, $\Delta t = 0.02$. *Right:* $M = 1000$, $\Delta t = 0.002$.

Compared to the schemes tested in [1, 40], our schemes do also perform well; see Figure 3.2, where we have plotted solutions by our schemes for the same discretization parameters used in Example 5.3 of [1]. The reference solution is found by the implicit midpoint scheme of [1] with $\theta = 1$ and very fine discretization in space and time: $M = 2000$ and $\Delta t = 0.0001$. We observe that the LILEP scheme behaves similarly to the multi-symplectic box scheme of Arscher and McLachlan (see figures 3 and 4 in [1]), seemingly with the same superior stability for rough discretization in space and time. The LIGEP scheme, on the other hand, starts to blow up at around $t = 1$ when $M = 60$, $\Delta t = 1/150$, but produces for $M = 100$, $\Delta t = 0.004$ a solution that is much closer to the correct solution than any of the schemes tested in [1] (see Figure 3 in that paper for comparison).



(a) **(b)**

Figure 3.2: Solutions of test problem 1 at time $t = 10$ by our schemes and the implicit midpoint scheme (IMP) as given in [1] (with $\theta = 2/3$ in the left figure and $\theta = 1$ in the right figure). *Left:* $M = 60$, $\Delta t = 1/150$. *Right:* $M = 100$, $\Delta t = 0.004$.

Test problem 2

To get quantitative results on the performance of our methods, we wish to study a problem with a known solution. For the KdV equation with $\gamma = 1$, $\eta = 6$, initial value $u_0(x) = \frac{1}{2}c\,\mathrm{sech}^2(-x + P/2)$ and periodic boundary conditions $u(x+P,t) = u(x,t)$, the exact solution is a soliton moving with a constant speed c in the positive x-direction while keeping its initial shape. That is,

$$u(x,t) = \frac{1}{2}c\,\mathrm{sech}^2((-x + ct) \bmod P - P/2).$$

In our numerical experiments, $c = 4$ and $P = 20$. For this problem, we have used the central difference operator to approximate ∂_x in the GEP and LIGEP schemes, since it gives good results and yields considerably shorter computational time than if the pseudospectral operator is used. The proposed methods all show very good stability conditions when applied to this problem, as expected by methods conserving some invariant. The initial shape of the wave is well kept for long integration times, even when quite large step sizes in space and time are used; Figure 3.3 gives a good illustration of this. As in the previous example, we again observe that little is lost in accuracy by choosing linearly implicit over fully implicit time integration. A close inspection of Figure 3.3 also indicates that the local energy-preserving schemes preserve the shape of the wave better than the global energy-preserving schemes, while on the other hand, the GEP and LIGEP schemes are better than the LEP and LILEP schemes at preserving the speed of the wave. This is confirmed in Table 3.2 by measuring

the shape error

$$\epsilon_{\text{shape}} := \min_{\tau} \| U^N - u(\cdot - \tau) \|_2^2$$

and phase error

$$\epsilon_{\text{phase}} := c\,|\operatorname{argmin}_{\tau} \| U^N - u(\cdot - \tau) \|_2^2 - ct|,$$

where U^N is the numerical solution at end time t.

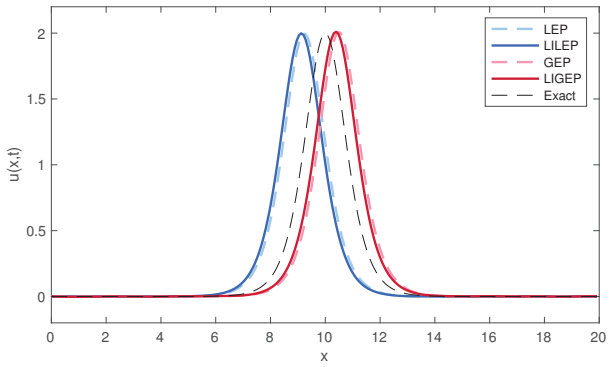


Figure 3.3: The soliton solution of the KdV equation at time $t = 100$, with $M = 250$ discretization points in space and a time step $\Delta t = 0.01$.

M		200			400			600	
	ϵ_{shape}	ϵ_{phase}	CT	ϵ_{shape}	ϵ_{phase}	CT	ϵ_{shape}	ϵ_{phase}	CT
LEP	4.67e-3	1.12	21.86	1.22e-3	3.81e-1	35.89	5.86e-4	2.43e-1	51.92
LILEP	4.10e-3	1.23	5.14	5.26e-4	4.88e-1	8.26	1.45e-4	3.50e-1	10.89
GEP	1.62e-2	8.61e-1	19.53	3.66e-3	1.16e-1	34.09	1.71e-3	2.32e-2	49.45
LIGEP	1.71e-2	7.50e-1	6.84	4.39e-3	5.19e-5	8.10	2.47e-3	1.31e-1	12.52

Table 3.2: Phase and shape errors and the computational time (CT) for different schemes applied to test problem 2 of the KdV equation, for varying number of discretization points M, with time step $\Delta t = 0.01$ and end time $t = 100$.

In Figure 3.4, we have plotted the computational time required to reach a certain accuracy in the global error for the different methods, both at time $t = 0.5$ and at time $t = 10$. We compare our methods to the fully implicit LEP and GEP schemes of [20], and also to two of the schemes studied in [1]: the multi-symplectic box scheme (MSB) and the implicit midpoint scheme (IMP). Most notably we see from both plots in Figure 3.4 that the linearly implicit schemes perform better than the fully implicit schemes. Also, we see that at time $t = 0.5$ the global error is lowest for the LILEP scheme, while at $t = 10$ it is lowest for the LIGEP scheme. This is in accordance with the schemes' phase and shape errors, which can be observed from Figure 3.3 and Table 3.2; with

increasing time, the phase error becomes more dominant, and thus the scheme with the smallest phase error becomes increasingly advantageous.

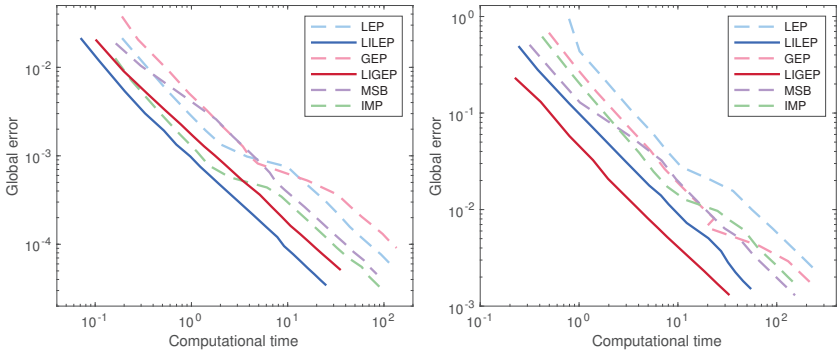


Figure 3.4: Computational time required to reach a given global error, with $\frac{\Delta x}{\Delta t}$ fixed, for test problem 2 of the KdV equation solved at time t. *Left:* $t = 0.5$, $\frac{\Delta x}{\Delta t} = 40$. *Right:* $t = 10$, $\frac{\Delta x}{\Delta t} = 8$.

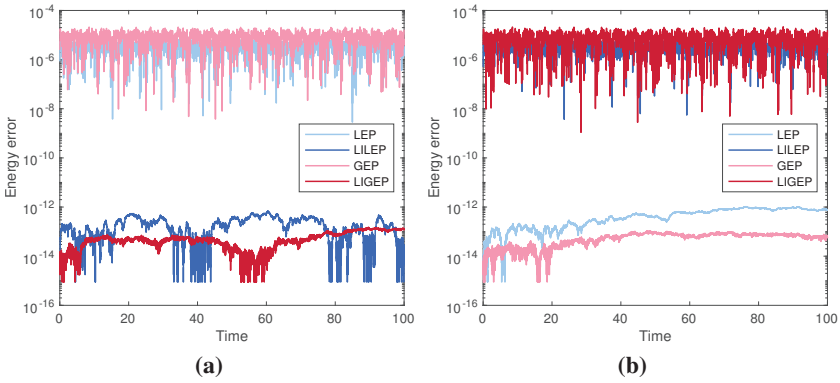
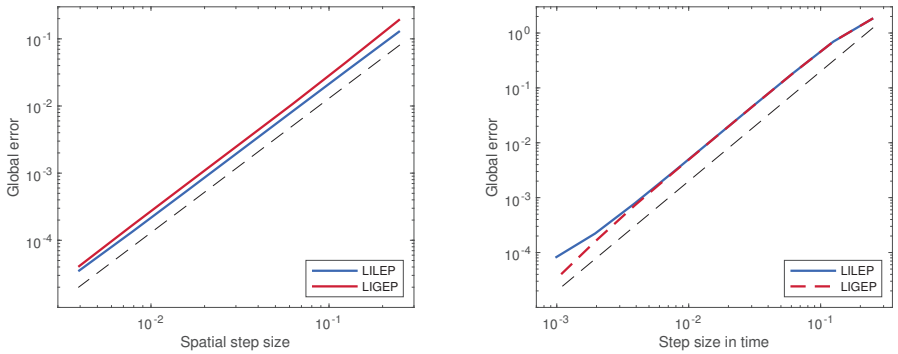


Figure 3.5: Error in discrete approximations to the global energy, by our methods and the fully implicit schemes of Gong et al. *Left:* The error in (3.5.2) for LEP/LILEP and the error in (3.5.6) for GEP/LIGEP, for test problem 2 solved with $M = 250$ discretization points in space and time step $\Delta t = 0.01$. *Right:* The error in (3.5.3) for LEP/LILEP and the error in (3.5.5) for GEP/LIGEP.

Figure 3.5 illustrates how the different schemes preserve a discrete approximation to the energy to machine precision. That is, the linearly implicit schemes LILEP and LIGEP preserve exactly the discrete energies (3.5.2) and (3.5.6), respectively, while keeping the discrete energies (3.5.3) and (3.5.5), respectively, within some bound which depends on the discretization parameters. Likewise,

the reverse is true for the fully implicit schemes. These observations fit well with our above results about the different discrete approximations to the energy: that for both the local energy preserving and the global energy preserving schemes, either discrete energy given can be seen as a modification of the other approximation. Finally, we have included plots in Figure 3.6 which confirm that our schemes are of second order in space and time.



(a) **(b)**

Figure 3.6: Order plots for the LILEP and LIGEP schemes, solving the second test problem for the KdV equation at time $t = 1$. The black, dashed line is a reference line with slope 2 in both plots. *Left:* Fixed temporal step $\Delta t = 2 \times 10^{-4}$. *Right:* Fixed spatial step $\Delta x = 4 \times 10^{-3}$.

3.5.2 Zakharov–Kuznetsov equation

Kahan's method is previously shown to have nice properties when applied to integrable ODE systems [8, 10], and to perform well compared to other linearly implicit methods when applied to the KdV and Camassa–Holm equations [15], which are completely integrable PDEs. We wish to test our methods also on non-integrable systems, as well as on higher-dimensional problems. Therefore we consider the Zakharov–Kuznetsov equation, which is a non-integrable PDE [22,34]. This two-dimensional generalisation of the KdV equation has a variety of applications, see e.g. [23] for a brief summary.

Applying the (2+1)-dimensional LILEP method (3.4.14) to the Zakharov–Kuznetsov equation (3.3.8) multi-symplectified as described in Example 3.2,

we find

$$\begin{aligned}
\delta_x\mu_t\mu_y\phi^n_{j,k} &= \mu_t\mu_x\mu_y u^n_{j,k},\\
\frac{1}{2}\delta_t\mu_x\mu_y\phi^n_{j,k}+\delta_x\mu_t\mu_y v^n_{j,k}+\delta_y\mu_t\mu_x w^n_{j,k} &= \mu_t\mu_x\mu_y p^n_{j,k}-\frac{1}{2}\mu_x\mu_y u^n_{j,k}\mu_x\mu_y u^{n+1}_{j,k},\\
\delta_x\mu_t\mu_y w^n_{j,k}-\delta_y\mu_t\mu_x v^n_{j,k} &= 0,\\
-\frac{1}{2}\delta_t\mu_x\mu_y u^n_{j,k}-\delta_x\mu_t\mu_y p^n_{j,k} &= 0,\\
-\delta_x\mu_t\mu_y u^n_{j,k}+\delta_y\mu_t\mu_x q^n_{j,k} &= -\mu_t\mu_x\mu_y v^n_{j,k},\\
-\delta_x\mu_t\mu_y q^n_{j,k}-\delta_y\mu_t\mu_x u^n_{j,k} &= -\mu_t\mu_x\mu_y w^n_{j,k}.
\end{aligned}$$

Upon eliminating all variables except u, we are left with

$$\delta_t\mu_t\mu_x^3\mu_y u^n_{j,k}+\frac{1}{2}\delta_x\mu_t\mu_x\mu_y(\mu_x\mu_y u^n_{j,k}\mu_x\mu_y u^{n+1}_{j,k})+\delta_x^3\mu_t^2\mu_y^2 u^n_{j,k}+\delta_x\delta_y^2\mu_t^2\mu_x^2 u^n_{j,k}=0.$$

The operator μ_t is again superfluous. Hence we get the scheme

$$\delta_t\mu_x^3\mu_y u^n_{j,k}+\frac{1}{2}\delta_x\mu_x\mu_y(\mu_x\mu_y u^n_{j,k}\mu_x\mu_y u^{n+1}_{j,k})+\delta_x^3\mu_t\mu_y^2 u^n_{j,k}+\delta_x\delta_y^2\mu_t\mu_x^2 u^n_{j,k}=0.$$

This scheme preserves

$$\begin{aligned}
\bar{\mathcal{E}}^n_L = \frac{1}{6}\Delta x\Delta y\sum_{j=0}^{M_x-1}\sum_{k=0}^{M_y-1}\Big(&2\delta_x\mu_y u^{n+1}_{j,k}\delta_x\mu_y u^n_{j,k}+(\delta_x\mu_y u^n_{j,k})^2\\
&+2\delta_y\mu_x u^{n+1}_{j,k}\delta_y\mu_x u^n_{j,k}+(\delta_y\mu_x u^n_{j,k})^2-(\mu_x\mu_y u^n_{j,k})^2(\mu_x\mu_y u^{n+1}_{j,k})\Big),
\end{aligned}$$

which is a two-step discrete approximation of the energy

$$\mathcal{E}=\int(\frac{1}{2}(\nabla u)^2-\frac{1}{6}u^3)\,d\Omega.$$

Similarly, applying the linearly implicit global energy-preserving method (3.4.25) to (3.3.10), we get the scheme

$$\delta_t u^n_{j,k}+\frac{1}{2}(D_x(u^n u^{n+1}))_{j,k}+\mu_t(D_x^3(u^n))_{j,k}+\mu_t(D_xD_y^2(u^n))_{j,k}=0,$$

which preserves the two-step discrete energy approximation

$$\begin{aligned}
\bar{\mathcal{E}}^n = \frac{1}{6}\Delta x\Delta y\sum_{j=0}^{M_x-1}\sum_{k=0}^{M_y-1}\Big(&2(D_xu^n)_{j,k}(D_xu^{n+1})_{j,k}+((D_xu^n)_{j,k})^2\\
&+2(D_yu^n)_{j,k}(D_yu^{n+1})_{j,k}+((D_yu^n)_{j,k})^2-(u^n_{j,k})^2u^{n+1}_{j,k}\Big).
\end{aligned}$$

Test problem

Taking a note from a numerical experiment performed in [4], we study the formation of cylindrical soliton pulses on the domain $[0, P] \times [0, P]$, $P = 30$, following the initial condition

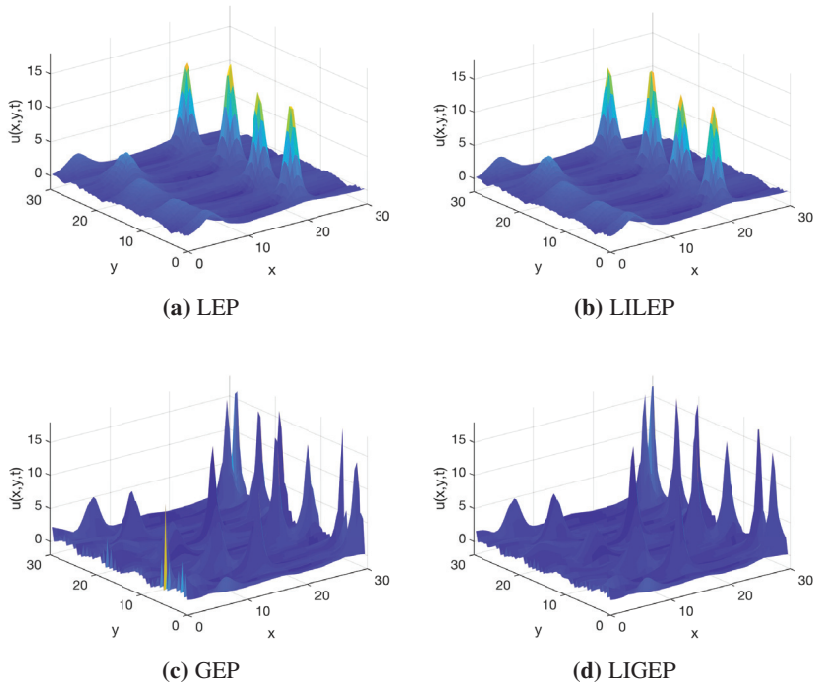
$$u_0(x, y) = 3c\,\mathrm{sech}^2\big(\frac{1}{2}\sqrt{c}(x - P/2)\big) + \xi(y),$$

where $\xi(y)$ is a random perturbation.

Upon trying the different schemes we can immediately conclude that the local energy-preserving schemes are superior for this problem when compared to the global energy-preserving schemes. The GEP and LIGEP schemes are too costly when the pseudospectral operator is used, and gives oscillatory behaviour in the y-direction when the central difference operator is used, unless the discretization in this direction is very fine. Although the global energy-preserving schemes with the central difference operator are slightly faster then the local energy-preserving schemes, as can be seen in Table 3.3, this is undermined by the cost of the extra discretization points needed to avoid oscillations in the former case. As was the case for the KdV problem, we see little difference between the linearly implicit schemes and their fully implicit counterparts. This can be seen in Figure 3.7, as can the oscillations in y-direction of the solution found by the GEP and LIGEP methods. The plots in Figure 3.7 can be compared to the plot in Figure 3.8, where the same problem is solved by the LILEP method using finer discretization in space and time. The initial random perturbation in y-direction over 75 points is then transferred over to 225 points using linear interpolation.

M	45	75	105	135	165	195	225	255
LEP	5.10	32.20	48.43	101.59	125.23	258.64	353.98	510.00
LILEP	2.04	8.87	14.57	31.02	37.25	78.98	108.02	157.91
GEP	3.62	19.54	41.87	73.59	122.31	186.74	258.19	352.36
LIGEP	1.38	6.00	13.45	23.79	39.31	60.27	83.32	113.13

Table 3.3: Running time, in seconds, for computing 100 steps in time by the various schemes and various number of discretization points $M = M_x = M_y$ in each spatial direction, solving our test problem for the Zakharov–Kuznetsov equation. As for the KdV equation test problems, a tolerance of 10^{-10} is used when solving the fully implicit schemes by Newton's method.



(a) LEP **(b)** LILEP

(c) GEP **(d)** LIGEP

Figure 3.7: The test problem of the Zakharov–Kuznetsov equation solved at time $t = 15$ by the different schemes, with $M = M_x = M_y = 75$ points in each spatial direction and $\Delta t = 0.1$.

3.6 Concluding remarks

In this paper, we propose two types of linearly implicit methods with conservation properties for cubic invariants of multi-symplectic PDEs. The linearly implicit local energy-preserving (LILEP) method preserves a discrete approximation to the local energy conservation law, and by extension, the global energy whenever periodic boundary conditions are considered. The linearly implicit global energy-preserving (LIGEP) method preserves the global energy without inheriting the local preservation from the continuous system.

We test our methods on two PDEs: the one-dimensional, integrable KdV equation and the two-dimensional, non-integrable Zakharov–Kuznetsov equation. The numerical experiments confirm that the proposed methods are of second order both in space and time and that they preserve the expected local and global energy conservation laws. We have observed excellent stability properties for the LILEP scheme in particular, and very high accuracy in the LIGEP scheme even for quite coarse discretization when a Fourier pseudospec-

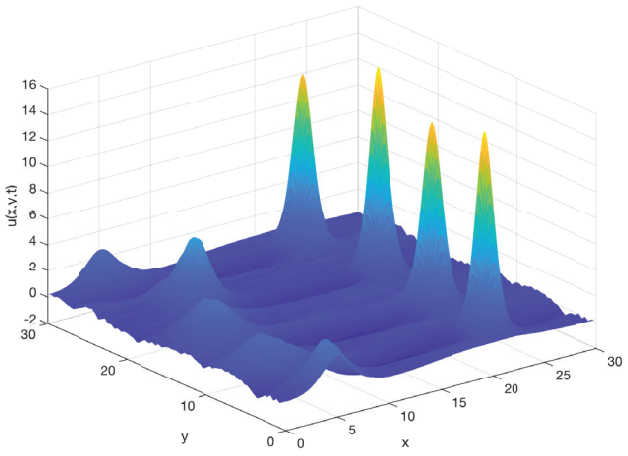


Figure 3.8: The test problem of the Zakharov–Kuznetsov equation solved at time $t = 15$ by the LILEP scheme, with $M = M_x = M_y = 225$ discretization points in each spatial direction and a temporal step size $\Delta t = 0.001$.

tral operator is used to approximate the spatial derivative. Compared to the fully implicit methods of Gong et al. in [20], which was an inspiration for this paper, our methods show comparable wave profiles, global errors and energy errors, at a significantly lower computational cost. For two-dimensional problems, where fully implicit schemes quickly become very expensive to compute, the combination of local energy-preservation and a linearly implicit method seems to provide for a very competitive method.

Although we have only considered the preservation of cubic invariants in this paper, our schemes can be extended to preserve higher order polynomials by the polarisation techniques for generalising Kahan's method suggested in [9]. This would result in $(p-2)$-step methods for preservation of a discrete p-order polynomial invariant.

Bibliography

[1] U. M. ASCHER AND R. I. MCLACHLAN, *On symplectic and multisymplectic schemes for the KdV equation*, J. Sci. Comput., 25 (2005), pp. 83–104.

[2] T. J. BRIDGES, *A geometric formulation of the conservation of wave action and its implications for signature and the classification of instabil-*

ities, Proc. Roy. Soc. London Ser. A, 453 (1997), pp. 1365–1395.

[3] T. J. BRIDGES, *Multi-symplectic structures and wave propagation*, vol. 121, 1997, pp. 147–190.

[4] T. J. BRIDGES AND S. REICH, *Multi-symplectic spectral discretizations for the Zakharov–Kuznetsov and shallow water equations*, Phys. D, 152/153 (2001), pp. 491–504. Advances in nonlinear mathematics and science.

[5] L. BRUGNANO, F. IAVERNARO, AND D. TRIGIANTE, *Hamiltonian boundary value methods (energy preserving discrete line integral methods)*, JNAIAM. J. Numer. Anal. Ind. Appl. Math., 5 (2010), pp. 17–37.

[6] W. CAI, H. LI, AND Y. WANG, *Partitioned averaged vector field methods*, J. Comput. Phys., 370 (2018), pp. 25–42.

[7] E. CELLEDONI, V. GRIMM, R. I. MCLACHLAN, D. I. MCLAREN, D. O'NEALE, B. OWREN, AND G. R. W. QUISPEL, *Preserving energy resp. dissipation in numerical PDEs using the "average vector field" method*, J. Comput. Phys., 231 (2012), pp. 6770–6789.

[8] E. CELLEDONI, R. I. MCLACHLAN, D. I. MCLAREN, B. OWREN, AND G. R. W. QUISPEL, *Integrability properties of Kahan's method*, J. Phys. A, 47 (2014), pp. 365202, 20.

[9] E. CELLEDONI, R. I. MCLACHLAN, D. I. MCLAREN, B. OWREN, AND G. R. W. QUISPEL, *Discretization of polynomial vector fields by polarization*, Proc. A., 471 (2015), pp. 20150390, 10.

[10] E. CELLEDONI, R. I. MCLACHLAN, B. OWREN, AND G. R. W. QUISPEL, *Geometric properties of Kahan's method*, J. Phys. A, 46 (2013), pp. 025201, 12.

[11] E. CELLEDONI, D. I. MCLAREN, B. OWREN, AND G. R. W. QUISPEL, *Geometric and integrability properties of Kahan's method: the preservation of certain quadratic integrals*, J. Phys. A, 52 (2019), pp. 065201, 9.

[12] Y. CHEN, S. SONG, AND H. ZHU, *The multi-symplectic Fourier pseudospectral method for solving two-dimensional Hamiltonian PDEs*, J. Comput. Appl. Math., 236 (2011), pp. 1354–1369.

[13] S. H. CHRISTIANSEN, H. Z. MUNTHE-KAAS, AND B. OWREN, *Topics in structure-preserving discretization*, Acta Numer., 20 (2011), pp. 1–119.

[14] M. DAHLBY AND B. OWREN, *A general framework for deriving integral preserving numerical methods for PDEs*, SIAM J. Sci. Comput., 33 (2011), pp. 2318–2340.

[15] S. EIDNES, L. LI, AND S. SATO, *Linearly implicit structure-preserving schemes for Hamiltonian systems*, arXiv preprint, arXiv:1901.03573, (2019).

[16] Z. FEI, V. M. PÉREZ-GARCÍA, AND L. VÁZQUEZ, *Numerical simulation of nonlinear Schrödinger systems: a new conservative scheme*, Appl. Math. Comput., 71 (1995), pp. 165–177.

[17] E. FRANCK, M. HÖLZL, A. LESSIG, AND E. SONNENDRÜCKER, *Energy conservation and numerical stability for the reduced MHD models of the non-linear JOREK code*, ESAIM Math. Model. Numer. Anal., 49 (2015), pp. 1331–1365.

[18] D. FURIHATA AND T. MATSUO, *Discrete variational derivative method: a structure-preserving numerical method for partial differential equations*, Chapman and Hall/CRC, 2010.

[19] D. FURIHATA AND T. MATSUO, *Discrete variational derivative method*, Chapman & Hall/CRC Numerical Analysis and Scientific Computing, CRC Press, Boca Raton, FL, 2011. A structure-preserving numerical method for partial differential equations.

[20] Y. GONG, J. CAI, AND Y. WANG, *Some new structure-preserving algorithms for general multi-symplectic formulations of Hamiltonian PDEs*, J. Comput. Phys., 279 (2014), pp. 80–102.

[21] E. HAIRER, C. LUBICH, AND G. WANNER, *Geometric numerical integration*, vol. 31 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 2006.

[22] H.-C. HU, *New exact solutions of Zakharov–Kuznetsov equation*, Commun. Theor. Phys. (Beijing), 49 (2008), pp. 559–561.

[23] H. IWASAKI, S. TOH, AND T. KAWAHARA, *Cylindrical quasi-solitons of the Zakharov–Kuznetsov equation*, Phys. D, 43 (1990), pp. 293–303.

[24] C. JIANG, W. CAI, AND Y. WANG, *A linear-implicit and local energy-preserving scheme for the sine-Gordon equation based on the invariant energy quadratization approach*, arXiv preprint, arXiv:1808.06854, (2018).

[25] C. JIANG, Y. GONG, W. CAI, AND Y. WANG, *A linearly implicit structure-preserving scheme for the Camassa–Holm equation based on multiple scalar auxiliary variables approach*, arXiv preprint, arXiv:1907.00167, (2019).

[26] W. KAHAN, *Unconventional numerical methods for trajectory calculations*, Unpublished lecture notes, (1993).

[27] R. A. LABUDDE AND D. GREENSPAN, *Energy and momentum conserving methods of arbitrary order for the numerical integration of equations of motion*, Numer. Math., 25 (1975), pp. 323–346.

[28] B. LEIMKUHLER AND S. REICH, *Simulating Hamiltonian dynamics*, vol. 14, Cambridge university press, 2004.

[29] Y. LI AND X. WU, *General local energy-preserving integrators for solving multi-symplectic Hamiltonian PDEs*, J. Comput. Phys., 301 (2015), pp. 141–166.

[30] J. E. MARSDEN, G. W. PATRICK, AND S. SHKOLLER, *Multisymplectic geometry, variational integrators, and nonlinear PDEs*, Comm. Math. Phys, 199 (1998), pp. 351–395.

[31] T. MATSUO AND D. FURIHATA, *Dissipative or conservative finite-difference schemes for complex-valued nonlinear partial differential equations*, J. Comput. Phys., 171 (2001), pp. 425–447.

[32] R. I. MCLACHLAN, G. QUISPEL, AND N. ROBIDOUX, *Geometric integration using discrete gradients*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 357 (1999), pp. 1021–1045.

[33] B. E. MOORE AND S. REICH, *Multi-symplectic integration methods for Hamiltonian PDEs*, Future Gener Comput Syst, 19 (2003), pp. 395–402.

[34] H. NISHIYAMA, T. NOI, AND S. OHARU, *Conservative finite difference schemes for the generalized Zakharov–Kuznetsov equations*, J. Comput. Appl. Math., 236 (2012), pp. 2998–3006.

[35] S. REICH, *Multi-symplectic Runge-Kutta collocation methods for Hamiltonian wave equations*, J. Comput. Phys., 157 (2000), pp. 473–499.

[36] Y. WANG, B. WANG, AND M. QIN, *Local structure-preserving algorithms for partial differential equations*, Sci. China Ser. A, 51 (2008), pp. 2115–2136.

[37] X. YANG, J. ZHAO, AND Q. WANG, *Numerical approximations for the molecular beam epitaxial growth model based on the invariant energy quadratization method*, J. Comput. Phys., 333 (2017), pp. 104–127.

[38] N. J. ZABUSKY AND M. D. KRUSKAL, *Interaction of "solitons" in a collisionless plasma and the recurrence of initial states*, Phys. Rev. Lett., 15 (1965), p. 240.

[39] V. ZAKHAROV AND E. KUZNETSOV, *Three-dimensional solitons*, Zh. Eksp. Teor. Fiz, 66 (1974), pp. 594–597.

[40] P. F. ZHAO AND M. Z. QIN, *Multisymplectic geometry and multisymplectic Preissmann scheme for the KdV equation*, J. Phys. A, 33 (2000), pp. 3613–3626.

Symplectic Lancozs and Arnoldi Method for Solving Linear Hamiltonian Systems of ODEs: Preservation of Energy and Other Invariants

Elena Celledoni and Lu Li

Published in: *Progress in Industrial Mathematics at ECMI 2016*

Symplectic Lanczos and Arnoldi Method for Solving Linear Hamiltonian Systems of ODEs: Preservation of Energy and Other Invariants

Abstract. Krylov subspace methods have become popular for the numerical approximation of matrix functions as for example for the numerical solution of large and sparse linear systems of ordinary differential equations. One well known technique is based on the method of Arnoldi which computes an orthonormal basis of the Krylov subspace. However, when applied to Hamiltonian linear systems of ODEs, this method fails to preserve the symplecticity of the solution under numerical discretization, or to preserve energy. In this work we apply the Symplectic Lanczos Method to construct a J-orthogonal basis of the Krylov subspace. This basis is then used to construct a numerical approximation which is energy preserving. The symplectic Lanczos method is widely used to approximate eigenvalues of large and sparse Hamiltonian matrices, but the approach for solving linear Hamiltonian systems is not well known in the literature. We also show that under appropriate additional assumptions on the structure of the linear Hamiltonian system, the Arnoldi method can preserve certain invariants of the system. We finally investigate numerically the energy and global error behaviour for the methods.

4.1 Arnoldi Projection Method

Consider a linear Hamiltonian initial value problem of the form

$$\begin{aligned} \dot{y} &= Ay \\ y(0) &= y_0 \quad t > 0, \end{aligned} \tag{4.1.1}$$

where $y(t) \in \mathbb{R}^{2m}$, $A \in \mathbb{R}^{2m\times 2m}$ and $JA = H$ is symmetric, and $y_0 \in \mathbb{R}^{2m}$. We denote by J the $2m \times 2m$ matrix

$$J = \begin{pmatrix} 0 & I_m \\ -I_m & 0 \end{pmatrix}, \tag{4.1.2}$$

with I_m the $m \times m$ identity matrix. The energy of system (4.1.1) is

$$\mathcal{H}(y) = \frac{1}{2} y^T JAy \equiv \frac{1}{2} {y_0}^T H y_0, \tag{4.1.3}$$

and it is preserved along solution trajectories, i.e. $\frac{d\mathcal{H}(y(t))}{dt} = 0$.

Consider the Krylov subspace of dimension n, generated by the matrix A and the vector y_0:

$$\mathcal{K}_n(A, y_0) := \text{span}\{y_0, Ay_0, \cdots, A^{n-1}y_0\}. \tag{4.1.4}$$

The basic idea of Krylov projection methods is to build a numerical approximation for the solution of (4.1.1) evolving in the Krylov subspace. This is done by solving (projected) linear systems of ordinary differential equations (ODEs) of much lower dimension than the original system.

One well known Krylov projection method is the one based on the Arnoldi algorithm [1,3,5]. This algorithm generates an orthonormal basis for $\mathcal{K}_n(A, y_0)$, which is stored in a $2m \times n$ matrix V_n, and an upper Hessenberg $n \times n$ matrix H_n, such that V_n and H_n satisfy

$$\begin{aligned} AV_n = V_n H_n + w_{n+1} e_n^T, \quad w_{n+1} = h_{n+1,n} v_{n+1}, \\ {V_n}^T V_n = I_n, \end{aligned} \tag{4.1.5}$$

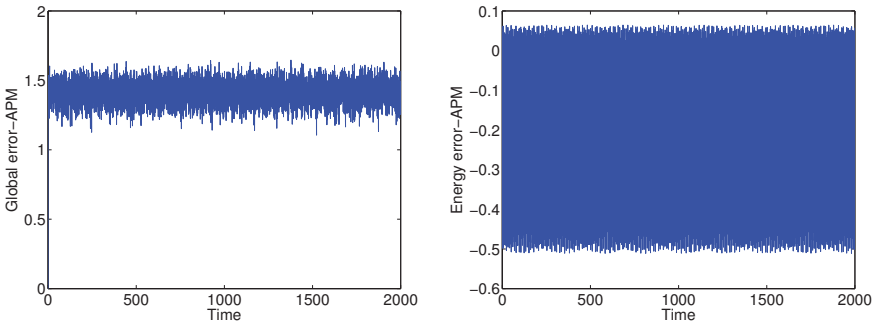
where v_{n+1} is the last column of V_{n+1} and $h_{n+1,n}$ is the $(n+1, n)$ entry of H_{n+1}. The Arnoldi projection method (APM) is then defined by searching for an approximation $y_A(t) = V_n z(t)$ of $y(t)$, where z is the solution of the following smaller system

$$\begin{aligned} \dot{z} &= H_n z, \\ z(0) &= z_0. \end{aligned} \tag{4.1.6}$$

When applied to Hamiltonian systems, the APM fails in general to preserve symplecticity or energy. However, we will see that in special cases, the APM can show a very good behaviour with respect to energy-preservation.

In Fig. 4.1 we report numerical tests performed with the APM on the problem (4.1.1) when A is Hamiltonian and simultaneously skew-symmetric. In the case when A is Hamiltonian and simultaneously skew-symmetric, the APM shows almost preservation of energy and boundedness of the global error over long time integration. This does not occur for general Hamiltonian matrices, as can be confirmed by numerical experiments with JA symmetric block diagonal or symmetric block tridiagonal. We will explain this behaviour in the next section.

For all the experiments including all the figures in the following chapters, we always consider $A \in \mathbb{R}^{200\times 200}$, and we will fix the dimension of the Krylov subspace to be 4; A and y_0 are randomly generated.



(a) Block skew-symmetric A **(b)** Block skew-symmetric A

Figure 4.1: Fig. 4.1a: we plot the global error $|| y(t) - y_A(t) ||_2$ versus time, where $y(t)$ is the reference solution. The reference solution is always computed using Cayley transform. **Fig.** 4.1b: we plot the energy error, where energy error is $\mathcal{H}(y_A(0)) - \mathcal{H}(y_A(t))$, and $y_A(0) = y_0$.

4.1.1 APM Case When $JA = AJ$

In this section we consider a Hamiltonian matrix A such that A and J commute.

Proposition 4.1. *Suppose A is a Hamiltonian matrix. Then J and A commute if and only if the matrix A is skew-symmetric.*

So, in this case, the numerical experiments of Fig. 4.1 show that the global error and the energy error remain bounded over time. One can show that in this case (4.1.1) has m preserved first integrals in involution, and APM preserves r modified first integrals.

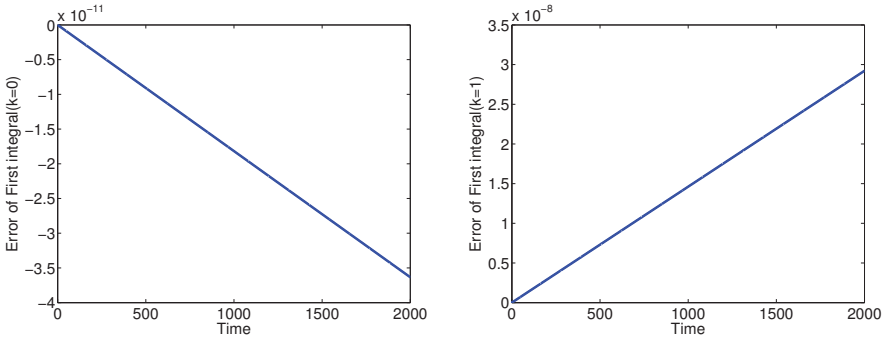
Proposition 4.2. *If $JA = AJ$, the Hamiltonian system (4.1.1) has m first integrals in involution, $\mathcal{H}_k(y) := \frac{1}{2} y^T A^{2k} y$ for $k = 0, 1, \ldots, m-1$, and in involution with the Hamiltonian $\mathcal{H}$.*

Proposition 4.3. *Applying the Arnoldi projection method to the Hamiltonian system (4.1.1), under the assumption $JA = AJ$, the numerical approximation preserves the following first integrals*

$$\mathcal{H}_k^A(y) := \frac{1}{2} y^T V_n (H_n)^{2k} V_n^T y, \quad k = 0, 1, \ldots r,$$

with $r = n/2 - 1$ if n is even and $r = (n-1)/2 - 1$ if n is odd.

In Fig. 4.2 we have performed experiments with $n = 4$, under the assumption that $JA = AJ$, APM preserves $r = 2$ first integrals, these are $\mathcal{H}_0^A(y_A) = \frac{1}{2} {y_A}^T y_A$, and $\mathcal{H}_1^A(y_A) = -\frac{1}{2} {y_A}^T V_n {H_n}^2 {V_n}^T y_A$.



(a) Block skew-symmetric A **(b)** Block skew-symmetric A

Figure 4.2: Preservation of the two first integrals by APM for a skew-symmetric and Hamiltonian matrix A. **Fig.** 4.2a: numerical error $\mathcal{H}_0^A(y_A(0)) - \mathcal{H}_0^A(y_A(t))$, and $y_A(0) = y_0$. **Fig.** 4.2b: numerical error $\mathcal{H}_1^A(y_A(0)) - \mathcal{H}_1^A(y_A(t))$.

The case $JA = AJ$ is not the only case when the energy error for APM appears to be bounded. We will introduce and explain another example in the following subsection. In what follows, we rewrite $JA = H$ in a block form

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{pmatrix}, \tag{4.1.7}$$

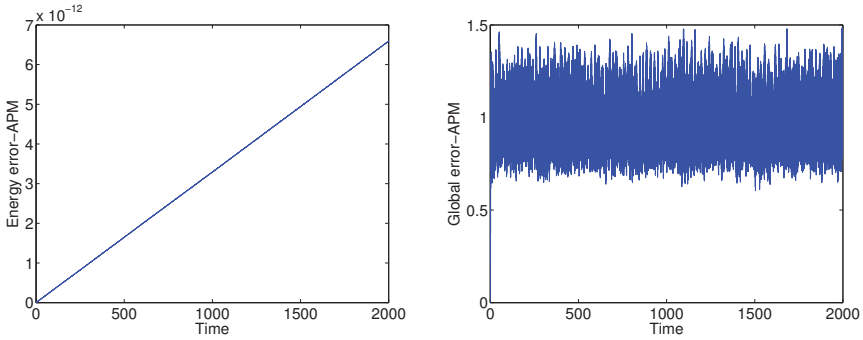
and then the original Hamiltonian system (4.1.1) has the form

$$\begin{aligned} \dot{p} &= -H_{12}^T p - H_{22} q \\ \dot{q} &= H_{11} p + H_{12} q, \end{aligned} \tag{4.1.8}$$

where $(p^T, q^T)^T = y$.

4.1.2 APM Case When $H_{1,2} = 0$, $H_{1,1} = I$ and $y_0 = ({p_0}^T, 0)^T$

If we consider Hamiltonian system (4.1.8) with $H_{1,2} = 0$, $H_{1,1} = I$ and an initial vector of the form $y_0 = ({p_0}^T, 0)^T$, we can show that the APM preserves the energy.



(a) Energy error versus time **(b)** Global error versus time

Figure 4.3: For both experiments we consider A, such that $H_{1,2} = 0$, $H_{1,1} = I$, and an initial vector $y_0 = ({p_0}^T, 0)^T$. (4.3a) Energy error versus time. (4.3b) Global error versus time.

In Fig. 4.3, one can see the good performance of the APM when $H_{1,2} = 0$, $H_{1,1} = I$ and $y_0 = ({p_0}^T, 0)^T$. The energy of the original system is preserved and the global error is bounded. This can be explained by the following proposition.

Proposition 4.4. *In the case when $H_{1,2} = 0$, $H_{1,1} = I$ and $y_0 = ({p_0}^T, 0)^T$ (and subject to a suitable permutation of the equations), the projected system (4.1.6) for the APM is a Hamiltonian system. The energy of the original system (4.1.1) is preserved by the numerical solution of the APM.*

Remark 4.1. *The APM applied to the special case when $H_{1,2} = 0$, $H_{1,1} = I$ and $y_0 = ({p_0}^T, 0)^T$ corresponds to performing a structure preserving model reduction in the spirit of [4]. Let V_n be the basis of the Krylov subspace $\mathcal{K}_n(-H_{22}, p_0)$. The approximations to the position q and the momentum p are restricted to evolve on this Krylov subspace and to take the form $V_n\hat{p}$ and $V_n\hat{q}$. Deriving the Euler-Lagrange equations and the corresponding Hamiltonian equations under this assumption, leads to a projected Hamiltonian system which coincides with (4.1.6).*

4.2 Symplectic Lanczos Projection Method

Instead of using an orthonormal basis, we consider applying the Symplectic Lanczos algorithm to construct a J-orthogonal basis for the Krylov subspace [2, 6].

Denote by J_{2n} the matrix (4.1.2) with $m = n$. Given the Hamiltonian matrix $A \in \mathbb{R}^{2m,2m}$ and the starting vector $y_0 \in \mathbb{R}^{2m}$, the symplectic Lanczos method

generates a sequence of $2m \times 2n$ matrices

$$S_{2n} = [v_1, ..., v_n, w_1, ...w_n], \tag{4.2.1}$$

which satisfy

$$AS_{2n} = S_{2n}H_{2n} + r_{n+1}e_{2n}^T, \tag{4.2.2}$$

where H_{2n} is a tridiagonal Hamiltonian $2m \times 2m$ matrix, S_{2n} is a symplectic matrix, i.e.,

$$S_{2n}{}^T J S_{2n} = J_{2n}, \tag{4.2.3}$$

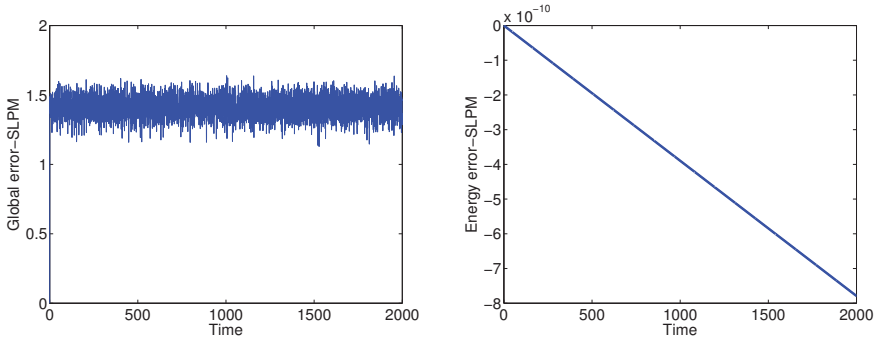
and $r_{n+1} = \zeta_{n+1} v_{n+1}$ is J-orthogonal to the columns of S_{2n}, see [6] for details.

The Symplectic Lanczos projection method (which we here denote SLPM) constructs the approximation $y_S = S_{2n}\hat{z}(t)$ of $y(t)$, where $\hat{z}$ is the exact flow for the following smaller system

$$\begin{aligned} \dot{\hat{z}} &= H_{2n}\hat{z}, \\ \hat{z}(0) &= \hat{z}_0. \end{aligned} \tag{4.2.4}$$

Proposition 4.5. *The SLPM method preserves the energy when applied to the system (4.1.1) with A a Hamiltonian matrix.*

In Fig. 4.4, we report the performance of the SLPM on symmetric block diagonal JA, similar results have been obtained with a range of different Hamiltonian matrices. As expected, the energy is preserved and the global error remains bounded.



(a) Block diagonal JA **(b)** Block diagonal JA

Figure 4.4: Fig. 4.4a: we plot the global error $|| y(t) - y_S(t) ||_2$ versus time. **Fig.** 4.4b: we plot the energy error, where energy error is $\mathcal{H}(y_A(0)) - \mathcal{H}(y_S(t))$, and $y_S(0) = y_0$.

4.3 Conclusion

In this paper we have reported some experiments with different Krylov subspace methods for linear Hamiltonian systems: a method based on the symplectic Lanczos method, which always preserves the energy; and a projection method based on the classical Arnoldi algorithm. We have seen that the Arnoldi projection method can preserve some modified integrals when applied to special systems (i.e., when A is Hamiltonian and skew-symmetric), and the APM preserves the energy when applied to some other special Hamiltonian matrices. We have recently obtained other Krylov subspace methods based on Gram Schmidt processes similar to the Arnoldi method, which preserve energy by solving exactly smaller Hamiltonian systems obtained by projection. The properties of all these Krylov methods and an appropriate comparison in terms of performance, stability with respect to propagation of rounding errors, and computational complexity will be discussed in a forthcoming paper.

Bibliography

[1] W. E. ARNOLDI, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] P. BENNER, H. FASSBENDER, AND M. STOLL, *A Hamiltonian Krylov–Schur-type method based on the symplectic Lanczos process*, Linear Algebra Appl., 435 (2011), pp. 578–600.

[3] E. GALLOPOULOS AND Y. SAAD, *Efficient solution of parabolic equations by Krylov approximation methods*, SIAM J. Sci. and Stat. Comput., 13 (1992), pp. 1236–1264.

[4] S. LALL, P. KRYSL, AND J. E. MARSDEN, *Structure-preserving model reduction for mechanical systems*, Phys. D, 184 (2003), pp. 304–318.

[5] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, SIAM, 2003.

[6] D. S. WATKINS, *On Hamiltonian and symplectic Lanczos processes*, Linear Algebra Appl., 385 (2004), pp. 23–45.

Krylov projection methods for linear Hamiltonian systems

Elena Celledoni and Lu Li

Published in: *Numerical Algorithms*

Krylov projection methods for linear Hamiltonian systems

Abstract. We study geometric properties of Krylov projection methods for large and sparse linear Hamiltonian systems. We consider in particular energy-preservation. We discuss the connection to structure preserving model reduction. We illustrate the performance of the methods by applying them to Hamiltonian PDEs.

5.1 Introduction

Large and sparse linear Hamiltonian systems arise in many fields of science and engineering, examples are models in network dynamics [27] and the semi-discretization of Hamiltonian partial differential equations (PDEs), like the wave equation [15, 22] and Maxwell's equations [21, 25]. In the context of Hamiltonian PDEs, the energy conservation law often plays a crucial role in the proof of existence and uniqueness of solutions [26]. Energy-preservation under numerical discretization can be advantageous as it testifies correct qualitative behaviour of the numerical solution, and it is also useful to prove convergence of numerical schemes [24]. There is an extensive literature on energy-preserving methods for ordinary differential equations (ODEs) [10, 11, 18, 23], but these methods need to be implemented efficiently to be competitive for large and sparse systems arising in numerical PDEs. Krylov projection methods are attractive for discrete PDE problems because they are iterative, accurate and they allow for restart and preconditioning strategies. But their structure preserving properties are not completely understood and should be further studied. This paper is a contribution in this direction.

It is well known that integration methods cannot be simultaneously symplectic and energy-preserving for general Hamiltonian systems [30]. However, the situation changes when we restrict to linear systems. An example is the midpoint rule which is symplectic and is also energy-preserving on linear problems; this is because the energy is quadratic for linear problems and the midpoint rule preserves all quadratic invariants. The midpoint method is implicit and requires the solution of one linear system of algebraic equations at each time step. The structure preserving properties are then retained only to the precision of the linear iterative solver. In this paper, we investigate preservation of geometric properties in Krylov projection methods for the exponential function. These are popular methods for the solution of discrete linear time-dependent PDEs [9,13], but because of the Krylov projection, structure is only preserved to the accu-

racy of the method. On the other hand, we show that some of these Krylov projection methods can be energy-preserving to a higher level of precision, and can preserve several first integrals simultaneously. We finally discuss the connections to structure-preserving model reduction and variational principles. In particular, we consider modified Hamilton's principle as the natural variational formulation for projection methods based on block J-orthogonal basis. Previous work in the context of structure preserving Krylov projection methods can be found in [2,20] and for Hamiltonian eigenvalue problems for example in [8].

The structure of this paper is as follows. We discuss symplecticity in section 6.1. Section 5.3 is devoted to the preservation of first integrals. Section 5.4 is devoted to projection methods based on block J-orthogonal bases and their connection to structure preserving model reduction. In Section 5.5, the geometric properties of the considered methods are illustrated by numerical examples.

5.2 Krylov projection and symplecticity

Consider a linear Hamiltonian initial value problem of the form

$$\dot{y} = f(y) = JHy, \quad y(t_0) = y_0, \qquad J = J_{2m} = \begin{bmatrix} 0 & I_m \\ -I_m & 0 \end{bmatrix}, \quad (5.2.1)$$

where $y(t) \in \mathbb{R}^{2m}$, $H \in \mathbb{R}^{2m\times 2m}$ is symmetric, $H^T = H$, $y_0 \in \mathbb{R}^{2m}$, and I_m is the $m \times m$ identity matrix. In what follows we denote by A the product $A = JH$. The matrix J is skew-symmetric, $J^T = -J$, and it defines a symplectic inner product[1] on $\mathbb{R}^{2m}$, $\omega(x, y) := x^T Jy$. Considering the energy function $\mathcal{H}(y) := \frac{1}{2} y^T Hy$, we have the gradient of $\mathcal{H}$ is $\nabla\mathcal{H}(y) = Hy$. The vector field of equation (5.2.1) is a Hamiltonian vector field, i.e. $\omega(f(y), v) = \nabla\mathcal{H}(y)^T v$, $\forall v \in \mathbb{R}^{2m}$. From this it follows that the flow map,

$$\varphi_t : \mathbb{R}^{2m} \to \mathbb{R}^{2m}, \qquad y_0 \mapsto y(t),$$

is a symplectic map [17], i.e. it satisfies

$$\Psi_{y_0}(t)^T J \Psi_{y_0}(t) = J, \qquad \text{where} \quad \Psi_{y_0}(t) := \frac{\partial \varphi_t(y_0)}{\partial y_0}.$$

A non-constant function $\mathcal{I}(y)$ is a first integral of the ODE $\dot{y} = f(y)$, if $\mathcal{I}(y)$ satisfies $\frac{d\mathcal{I}(y)}{dt}|_{y=y(t)} = \nabla\mathcal{I}(y)\dot{y} = \nabla\mathcal{I}(y) f(y) = 0$ for all y. So $\mathcal{I}(y)$ is constant

[1] A symplectic inner product on a vector space is a nondegenerate skew-symmetric bilinear form [3].

along the solution trajectory: $\mathcal{I}(y(t)) - \mathcal{I}(y(t_0)) = \int_{t_0}^{t} \nabla \mathcal{I}(y)\dot{y}\,dt = 0$. The energy function $\mathcal{H}(y)$ is a first integral of (5.2.1). An approximation one-step method for (5.2.1) is said to be energy-preserving if $\mathcal{H}$ is constant along the numerical solution, and symplectic if the numerical one-step method (numerical flow map)

$$\phi_h : \mathbb{R}^{2m} \to \mathbb{R}^{2m}, \qquad y_0 \mapsto \tilde{y} \approx y(t_0 + h)$$

is such that

$$\frac{\partial \phi_h(y_0)}{\partial y_0}^T J \frac{\partial \phi_h(y_0)}{\partial y_0} = J,$$

[17].

The idea of Krylov projection methods is to build numerical approximations for (5.2.1) in the Krylov subspace:

$$\mathcal{K}_r(A, y_0) := \text{span}\{y_0, Ay_0, \cdots, A^{r-1}y_0\},$$

which is a subspace of $\mathbb{R}^{2m}$ of dimension $r << 2m$. Let us consider even dimension $r = 2n$. A basis of $\mathcal{K}_{2n}(A, y_0)$ is constructed. The most well-known Krylov projection method is the one based on the Arnoldi algorithm [4] generating an orthonormal basis for $\mathcal{K}_{2n}(A, y_0)$. The method gives rise to a $2m \times 2n$ matrix V_{2n} with orthonormal columns, and to an upper Hessenberg $2n \times 2n$ matrix T_{2n} such that $I_{2n} = {V_{2n}}^T V_{2n}$, and $T_{2n} = V_{2n}^T A V_{2n}$. The approximation of $y(t)$ is

$$y_A(t) := V_{2n} z(t), \quad \text{where} \quad \dot{z} = T_{2n} z, \quad z(0) = z_0 = V_{2n}^T y_0. \tag{5.2.2}$$

We will denote this method by Arnoldi projection method (APM). Consider J_{2n} and the corresponding symplectic inner product in $\mathbb{R}^{2n}$, $\omega(u, v) := u^T J_{2n} v$. If $n < m$, unless we make further assumptions on H, the projected system (5.2.2) is not a Hamiltonian system in $\mathbb{R}^{2n}$, this is because $J_{2n}^{-1} T_{2n} = J_{2n}^{-1} V_{2n}^T J H V_{2n}$ is in general not symmetric and $J_{2n}^{-1} T_{2n} z$ is in general not the gradient of some energy function.

Instead of using an orthonormal basis, one can construct a J-orthogonal basis for $\mathcal{K}_{2n}(A, y_0)$ using the symplectic Lanczos algorithm [6]. The matrix S_{2n} whose columns are the vectors of this J-orthogonal basis satisfies

$$S_{2n}^T J S_{2n} = J_{2n}.$$

We will denote the corresponding Krylov projection method by Symplectic Lanczos projection method (SLPM). The projected system for SLPM is analog to (5.2.2), with V_{2n} replaced by S_{2n}, T_{2n} by $J_{2n} S_{2n}^T H S_{2n}$ and an appropriate z_0 (see Section 5.3.3). This projected system is a Hamiltonian system. But for $n < m$, the approximation $y_S(t) := S_{2n} z(t)$ is not symplectic. In fact, y_S is the

solution of the system

$$\dot{y}_S = (S_{2n} J_{2n} S_{2n}^T)\, H\, y_S, \quad y_S(t_0) = y_0, \tag{5.2.3}$$

and (5.2.3) is a Poisson system with Poisson structure given by the matrix $S_{2n}J_{2n}S_{2n}^T$ which is skew-symmetric and depends on the initial condition[2]. For $n = m$, $J_{2m} = J$, and $y_S = y$. However, the case $n < m$ is the most relevant for the use of the method in practice. In spite of not preserving the symplectic inner product ω, SLPM clearly shares important structural properties with the exact solution of (5.2.1) and is energy-preserving, see Section 5.3.3.

The symplectic Lanczos algorithm is not the only way to obtain a basis which is symplectic for the Krylov subspace. We will consider block J-orthogonal bases in Section 5.4 and show that they can be viewed as techniques of structure preserving model reduction, in the spirit of [19]. We propose one Krylov algorithm based on these ideas.

5.3 Preservation of first integrals and energy

We first present a result about the first integrals for a general linear Hamiltonian system. Recall that two first integrals F and G of an ODE are said to be in involution if their Poisson bracket $\{F, G\} := (\nabla F)^T J \nabla G$ vanishes [17].

Proposition 5.1. *For $A = JH$ where J is skew-symmetric and invertible, and H is symmetric and invertible, the system $\dot{y} = Ay$, $y(t_0) = y_0$ has the following first integrals in involution, $\mathcal{H}_k(y) = \frac{1}{2} y^T H A^{2k} y$, for $k = 0, 1, \ldots$. The Hamiltonian of the system is $\mathcal{H} = \mathcal{H}_0$.*

Proof. We consider the derivative of $\mathcal{H}_k$ along solution trajectories $y(t)$

$$\begin{aligned}
\frac{d}{dt}\mathcal{H}_k(y(t)) &= \frac{1}{2}\left[\dot{y}^T H (JH)^{2k} y + y^T H (JH)^{2k} \dot{y}\right] \\
&= \frac{1}{2}\left[-y^T H J H (JH)^{2k} y + y^T H (JH)^{2k} J H y\right] \\
&= \frac{1}{2}\left[-y^T H (JH)^{2k+1} y + y^T H (JH)^{2k+1} y\right] \\
&= 0,
\end{aligned}$$

so $\mathcal{H}_k(y)$, $k = 0, \ldots$, are preserved: $\mathcal{H}_k(y(t)) = \mathcal{H}_k(y_0)$. The integrals are in

[2] A Poisson system in $\mathbb{R}^d$ is a system of the type $\dot{y} = \Omega \nabla \mathcal{H}(y)$, where Ω is a skew-symmetric matrix, not necessarily invertible and can depend on y. Ω must satisfy the Jacobi identity, [17]. In our case, Ω depends on y_0.

involution because their Poisson bracket is zero,

$$\begin{aligned}\{\mathcal{H}_k, \mathcal{H}_p\} &= (\nabla\mathcal{H}_k)^T J \nabla\mathcal{H}_p = (A^{2k}y)^T HJH(A^{2p}y) \\ &= y^T((JH)^{2k})^T HJH(JH)^{2p}y = y^T H(JH)^{2(k+p)+1}y = 0,\end{aligned}$$

where we have used the skew-symmetry of $H(JH)^{2(k+p)+1}$.

□

In what follows, we will discuss the preservation of the first integrals of Proposition 5.1 when applying Krylov projection methods.

5.3.1 Preservation of first integrals for the APM

It can be observed from numerical simulations that the APM fails in general to preserve energy when applied to Hamiltonian systems, Figure 5.1 (left), Section 5.5, but structure-preserving properties can be ensured for such method via a simple change of inner product. Assume that H is symmetric and positive definite so that $\langle\cdot,\cdot\rangle_H := \langle\cdot, H\cdot\rangle$ defines an inner product. We modify the Arnoldi algorithm by replacing the usual inner product $\langle\cdot,\cdot\rangle$ by $\langle\cdot,\cdot\rangle_H$. We then show that the numerical solution given by this method preserves to machine accuracy certain first integrals. The modified Arnoldi algorithm (see Algorithm 5.1) generates a H-orthonormal basis, which is stored in the $2m \times n$ matrix V_n, satisfying $V_n^T H V_n = I_n$. This algorithm generates a skew-symmetric tridiagonal matrix T_n such that

$$\begin{aligned} AV_n &= V_n T_n + w_{n+1}e_n^T, \quad & w_{n+1} &= h_{n+1,n}v_{n+1}, \\ {V_n}^T H V_n &= I_n, & {V_n}^T H w_{n+1} &= 0. \end{aligned}$$

In what follows, we consider the Krylov projection method

$$y_H := V_n z, \quad \text{where } z \text{ satisfies} \quad \dot{z} = T_n z, \quad z(t_0) = {V_n}^T H y_0.$$

Proposition 5.2. *The numerical approximation y_H for the solution y of* (5.2.1) *preserves the following first integrals:*

$$\bar{\mathcal{H}}_k(y_H) = \frac{1}{2} y_H^T H V_n (T_n)^{2k} {V_n}^T H y_H$$

for all $k = 0, 1, \ldots$.

Proof. We observe that $T_n = {V_n}^T HJHV_n$ is skew-symmetric. So the ODE system for z has first integrals: $\mathcal{I}_k(z) = \frac{1}{2} z^T (T_n)^{2k} z$, for all $k = 0, 1, \ldots$.

Therefore, we have $\bar{\mathcal{H}}_k(y_H) = \frac{1}{2} y_H^T H V_n (T_n)^{2k} V_n{}^T H y_H = \frac{1}{2} z^T (T_n)^{2k} z$ and $\frac{d}{dt}\bar{\mathcal{H}}_k(y_H(t)) = \frac{d}{dt}\mathcal{I}_k(z(t)) = 0$, so the first integrals are preserved. □

Remark 5.1. *If n is even, the above Krylov projection method induces a projected problem which is conjugate to a Hamiltonian system, i.e., it can be written in the form* (5.2.1) *via change of variables. Since H_n is skew-symmetric, H_n can be factorized as $H_n = U_n J_n D_n U_n^{-1}$ where D_n is diagonal. Then, H_n can be transformed to a Hamiltonian matrix by a similarity transformation using U_n.*

5.3.2 Hamiltonian system with $JA = AJ$

We now consider J given by (5.2.1). Assume that A and J commute, then A is skew-symmetric, and the Hamiltonian system (5.2.1) has two Hamiltonian structures, one associated to A with Hamiltonian $\frac{1}{2} y^T y$, the other to J with Hamiltonian $\frac{1}{2} y^T H y$. The APM with Euclidean inner product $\langle \cdot, \cdot \rangle$ preserves modified first integrals. To proceed, we first give the following result.

Proposition 5.3. *Suppose A is a Hamiltonian matrix. Then J and A commute if and only if the matrix A is skew-symmetric.*

Proof. Suppose A is a Hamiltonian matrix and $A = JH$, where J and H are defined as in equation (5.2.1). Then the fact that J and A commute implies that $JJH = JHJ$, i.e. $-H = JHJ$ and by multiplying J^{-1} from right side, we get $-(JH)^T = JH$, namely $A^T = -A$. On the other hand, the fact that A is skew-symmetric implies that $(JH)^T = -JH$ and using this we get $JA = JJH = -J(JH)^T = JHJ = AJ$. □

The first integrals of the system (5.2.1) are given by the following proposition.

Proposition 5.4. *If $JA = AJ$, the Hamiltonian system* (5.2.1) *has the following first integrals in involution, $\mathcal{H}_k(y) = \frac{1}{2} y^T A^{2k} y$ for $k = 0, 1, \ldots$, and the first integrals are in involution with the Hamiltonian $\mathcal{H}(y) = \frac{1}{2} y^T H y$.*

Proof. From Proposition 5.3 we know that A is skew-symmetric. Then Proposition 5.1 holds with J replaced by A, and H replaced by the identity matrix. The integrals are in involution with the Hamiltonian $\mathcal{H}(y) = \frac{1}{2} y^T H y$ in fact

$$\{\mathcal{H}_k, \mathcal{H}\} = y^T A^{2k} J H y = y^T A^{2k} A y = 0, \quad k = 0, \ldots.$$

□

Remark 5.2. *By a direct application of Proposition 5.2, the APM to the Hamiltonian system* (5.2.1)*, under the assumption $JA = AJ$, gives a numerical approximation $y_A := V_n z$ which preserves the following modified first integrals*

$$\tilde{\mathcal{H}}_k(y_A) := \frac{1}{2} y_A^T V_n (H_n)^{2k} V_n^T y_A, \quad k = 0, 1, \ldots.$$

We next prove that the Hamiltonian of (5.2.1) is bounded by y_A under the assumption that J and A commute.

Proposition 5.5. *Assume the APM is applied to* (5.2.1)*. Under the assumption $JA = AJ$, the energy $\mathcal{H}(y) = \frac{1}{2} y^T J^{-1} A y$, is bounded along the numerical solution.*

Proof. This result follows directly from Remark 5.2 with $k = 0$, i.e.,

$$\frac{1}{2} {y_A}^T J^{-1} A y_A \leq \frac{1}{2} {y_A}^T y_A \| J^{-1} A \|_2 = \frac{1}{2} {y_0}^T y_0 \| J^{-1} A \|_2.$$

□

Proposition 5.5 explains the good behaviour of the APM in [12].

5.3.3 Symplectic Lanczos projection method

We now consider the symplectic Lanczos projection method (SLPM). Krylov subspace methods based on the symplectic Lanczos algorithm are widely used for the computation of eigenvalues of large and sparse Hamiltonian matrices [5, 14, 29]. For their use in the approximation of linear Hamiltonian systems see [1], [13].

Given $A \in \mathbb{R}^{2m,2m}$ and the starting vector $y_0 \in \mathbb{R}^{2m}$, the symplectic Lanczos method generates a sequence of matrices

$$S_{2n} = [v_1, ..., v_n, w_1, ... w_n] \quad \text{satisfying} \quad AS_{2n} = S_{2n} T_{2n} + r_{n+1} e_{2n}^T,$$

where T_{2n} is a tridiagonal Hamiltonian matrix, and $r_{n+1} = \zeta_{n+1} v_{n+1}$ is J-orthogonal with respect to the columns of S_{2n}. Since S_{2n} has J-orthogonal columns, i.e., ${S_{2n}}^T J S_{2n} = J_{2n}$, we know that

$$T_{2n} = J_{2n}^{-1} S_{2n}^T J A S_{2n} = J_{2n} S_{2n}^T H S_{2n},$$

and the projected system is a Hamiltonian system, where $z_0 = J_{2n}^{-1} S_{2n}^T J y_0$. More-

over, we have

$$\mathcal{H}_S(z) = \frac{1}{2} z^T J_{2n}^{-1} T_{2n} z \equiv \frac{1}{2} z_0^T J_{2n}^{-1} T_{2n} z_0. \tag{5.3.1}$$

Proposition 5.6. *The SLPM is an energy-preserving method for* (5.2.1).

Proof. The result follows by computing the Hamiltonian of (5.2.1) along numerical trajectories $y_S = S_{2n} z$, $\mathcal{H}(y_S) = \frac{1}{2} {y_S}^T J^{-1} A y_S$, and then using (5.3.3) and (5.3.1). □

5.4 Projection methods based on block J-orthogonal basis

We now consider a general strategy for Krylov projection methods to obtain J-orthogonal bases, this will lead automatically to energy preserving methods for (5.2.1). In what follows we will use the notation $(q^T, p^T)^T = y$ and write H in block form, and rewrite (5.2.1) accordingly:

$$\begin{aligned} \dot{q} &= H_{12}^T q + H_{22} p, \\ \dot{p} &= -H_{11} q - H_{12} p, \end{aligned} \qquad H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix}.$$

Assume that we can construct two matrices with linearly independent columns $V_n \in \mathbb{R}^{m \times n}$ and $W_n \in \mathbb{R}^{m \times n}$ such that ${V_n}^T W_n = I_n$. Then the matrix

$$S_{2n} := \begin{bmatrix} V_n & 0 \\ 0 & W_n \end{bmatrix} \tag{5.4.1}$$

has J-orthogonal columns. We will approximate y by the following projection method: $y \approx y_B$ defined by

$$y_B = S_{2n} z, \quad \text{where } z \text{ satisfies} \quad \dot{z} = J_{2n} S_{2n}^T J^{-1} A S_{2n} z, \quad z(t_0) = z_0, \tag{5.4.2}$$

and for the SLPM $z_0 = J_{2n}^{-1} S_{2n}^T J y_0$.

Proposition 5.7. *If $y_0 = S_{2n} z(t_0)$, then the energy of the original Hamiltonian system* (5.2.1) *will be preserved by the numerical solution* (5.4.1)-(5.4.2).

Proof. Notice that $\mathcal{H}(S_{2n} z) = \frac{1}{2} z^T S_{2n}^T J^{-1} A S_{2n} z$ is constant with respect to t because z is the solution of a Hamiltonian system with energy of the form $\mathcal{E}(z) = \frac{1}{2} z^T (S_{2n}^T J^{-1} A S_{2n}) z$. The result then follows directly from the fact that $\mathcal{E}(z) \equiv \mathcal{E}(z_0) = \mathcal{H}(y_0)$. □

We here propose one strategy to construct S_{2n} as in (5.4.1) with $W_n^T V_n = I_n$ and $V_n = W_n$. Let K_n be the Krylov matrix $2m \times n$, and consider the first m rows of K_n and the last m separately:

$$K_n := [y_0, Ay_0, \ldots, A^{n-1} y_0], \qquad K_n = \begin{bmatrix} K_n^q \\ K_n^p \end{bmatrix}.$$

We then find an orthonormal basis V_n for $\text{span}\{K_n^q, K_n^p\} \subset \mathbb{R}^m$ by either a QR-factorisation (algorithm 5.2 in the Appendix [3]) or a Gram-Schmidt process.

5.4.1 Structure preserving model reduction using Krylov subspaces

In this section we consider Hamilton's phase space variational principle (also called modified Hamilton's principle) [16, Ch. 8-5] which is the fundament of the projection methods based on block J-orthogonal basis. Since our system (5.2.1) is given in the form of an Hamiltonian system, it is natural to use the phase space variational principle, which is formulated in terms of the variables p and q and the Hamiltonian function $H(q, p)$, rather than the classical Hamilton's principle which is formulated in terms of q and $\dot{q}$ and the Lagrangian function $L(q, \dot{q})$. Following [19], we restrict the phase space variational principle to low dimensional subspaces of $\mathbb{R}^m$ and derive the projected Hamiltonian system taking variations on the low dimensional subspaces.

Assume $[q^T, p^T]^T := y$ and q and p are m-dimensional vectors belonging to $\mathbb{R}^m$ and its dual respectively, and that the Hamiltonian $\mathcal{H} : \mathbb{R}^m \times (\mathbb{R}^m)^* \to \mathbb{R}$ is $\mathcal{H}(q, p) := \mathcal{H}(y)$.[4] Considering the action functional $\mathcal{S} : \mathbb{R}^m \times (\mathbb{R}^m)^* \to \mathbb{R}$

$$\mathcal{S}(q, p) := \int_{t_0}^{t_{end}} \left(p^T \dot{q} - \mathcal{H}(q, p) \right) dt, \tag{5.4.3}$$

Hamilton's phase space variational principle states that

$$\delta \mathcal{S} = 0$$

for fixed $q_0 = q(t_0)$ and $q_{end} = q(t_{end})$, and it is equivalent to Hamilton's equations (5.2.1), [16, Ch. 8-5] . By projecting q and p separately on appropriate subspaces $\text{span}\{V_n\} \subset \mathbb{R}^m$ and $\text{span}\{W_n\} \subset (\mathbb{R}^m)^*$, i.e., $q \approx V_n \hat{q}$ and $p \approx W_n \hat{p}$, one restricts the variational principle to $\text{span}\{V_n\} \times \text{span}\{W_n\}$:

[3] Notice that to obtain a stable algorithm it is an advantage to replace the Krylov matrix with an orthonormal matrix obtained by the Arnoldi algorithm.

[4] The duality pairing between $\mathbb{R}^m$ and $(\mathbb{R}^m)^*$ is here simply $\langle p, q \rangle := p^T q$.

$\hat{\mathcal{S}}(\hat{q},\hat{p}) := \mathcal{S}(V_n\,\hat{q}, W_n\,\hat{p})$. Taking variations

$$0 = \delta\hat{\mathcal{S}}(\hat{q},\hat{p}) = \delta\int_{t_0}^{t_{end}} (W_n\hat{p})^T V_n\dot{\hat{q}} - \mathcal{H}(V_n\hat{q}, W_n\hat{p})\,dt$$

for fixed endpoints $\hat{q}_0 = \hat{q}(t_0)$ and $\hat{q}_{end} = \hat{q}(t_{end})$, we obtain the Hamiltonian equations associated to this reduced variational principle

$$\begin{aligned}\dot{\hat{p}} &= -{V_n}^T H_{12} W_n\hat{p} - {V_n}^T H_{11} V_n\hat{q},\\ \dot{\hat{q}} &= {W_n}^T H_{22} W_n\hat{p} + {W_n}^T H_{12}^T V_n\hat{q},\end{aligned}$$

which coincide with the system for z in (5.4.2). This explains the connection of the projection methods based on block J-orthogonal basis, (5.4.1) and (5.4.2), with the techniques of structure preserving model reduction derived in [19] and here modified for the phase space variational principle.

Assuming additional structure on H, we will show in the next section that the usual APM applied to the resulting system falls in the same class of projection methods based on block J-orthogonal basis and is a structure preserving model reduction method in the spirit of [19]. Model reduction methods for general second order systems obtained projecting the differential systems onto Krylov subspaces using an Arnoldi or a Lanczos process have been previously studied [7].

5.4.2 Special case $H_{1,2} = O$, $H_{2,2} = I$.

This special case is directly related to the setting in [19]. Denoting $y = (q^T, p^T)^T$, we consider the corresponding Hamiltonian system

$$\dot{y} = Ay \quad \text{with} \quad A = \begin{bmatrix} 0 & I \\ -H_{11} & 0 \end{bmatrix}. \tag{5.4.4}$$

and we notice that $p = \dot{q}$ in this case. The action functional (5.4.3) from the previous section is the integral of the Lagrangian density function

$$L(q(t),\dot{q}(t)) = \frac{1}{2}\dot{q}(t)^T\dot{q}(t) - \frac{1}{2}q(t)^T H_{11}\,q(t), \tag{5.4.5}$$

and in this case because $\dot{q} = p$ the phase space variational principle coincides with Hamilton's principle. Let V_n be the basis of the Krylov subspace $\mathcal{K}_n(-H_{11}, p_0)$ obtained via the Arnoldi algorithm. The reduced Lagrangian becomes

$$L(\hat{q}(t),\dot{\hat{q}}(t)) = \frac{1}{2}\dot{\hat{q}}(t)^T\dot{\hat{q}}(t) - \frac{1}{2}\hat{q}(t)^T {V_n}^T H_{11} V_n\hat{q}(t), \tag{5.4.6}$$

and the corresponding Hamiltonian equations are

$$\begin{aligned} \dot{\hat{q}} &= \hat{p}, \\ \dot{\hat{p}} &= -{V_n}^T H_{11} V_n \hat{q}(t). \end{aligned} \tag{5.4.7}$$

By solving (5.4.7), we obtain $(\hat{q}^T, \hat{p}^T)^T$ and then can construct the model reduction approximation $((V_n\hat{q})^T, (V_n\hat{p})^T)^T \approx (q^T, p^T)^T$.

Proposition 5.8. *When applied to* (5.4.4) *with* $y_0 = (0, p_0^T)^T$*, the model reduction procedure outlined in* (5.4.5)-(5.4.7) *coincides with the APM.*

Proof. Let $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^2$ be the two vectors of the canonical basis in $\mathbb{R}^2$. Denote by $\otimes$ the Kronecker tensor product. We have

$$\mathcal{K}_{2n}(A, y_0) = \text{span}\{\mathbf{e}_1 \otimes p_0, \mathbf{e}_2 \otimes p_0, \mathbf{e}_1 \otimes (-H_{11}p_0), \mathbf{e}_2 \otimes (-H_{11})p_0, \ldots\}.$$

Denote by $\mathbb{U}_{2n} \in \mathbb{R}^{2m\times 2n}$ the orthogonal matrix generated by the usual Arnoldi algorithm with matrix A, vector $y_0 = (0, p_0^T)^T$ and Euclidean inner product. Then $\mathbb{U}_{2n}$ is given by

$$\mathbb{U}_{2n} = \begin{bmatrix} 0 & v_1 & 0 & v_2 & 0 & \ldots & 0 & v_n \\ v_1 & 0 & v_2 & 0 & v_3 & \ldots & v_n & 0 \end{bmatrix},$$

and satisfies

$${\mathbb{U}_{2n}}^T A \mathbb{U}_{2n} = \Pi_{2n} \begin{bmatrix} 0 & I_n \\ -V_n^T H_{11} V_n & 0 \end{bmatrix} {\Pi_{2n}}^T \quad \text{and} \quad \mathbb{U}_{2n}\Pi_{2n} = \begin{bmatrix} V_n & O \\ O & V_n \end{bmatrix},$$

where $v_1, v_2, \ldots v_n$ are the columns of V_n and Π_{2n} is a $2n \times 2n$ permutation matrix. After a permutation of the variables $w = {\Pi_{2n}}^T z$, the projected system by APM $\dot{z} = {\mathbb{U}_{2n}}^T A \mathbb{U}_{2n} z$, $z(t_0) = {\mathbb{U}_{2n}}^T y_0$ can be rewritten in the form (5.4.1)-(5.4.2). □

5.5 Numerical Examples

In this section, several numerical examples are presented to illustrate the behavior of the following methods:

- APM: Arnoldi projection method using Euclidean inner product, Section 5.3;
- APMH: Arnoldi projection method using the inner product $\langle\cdot,\cdot\rangle_H$, Section 5.3;

- SLPM: symplectic Lanczos projection method, Section 5.3.3;
- BJPM: block J-orthogonal projection method with QR factorization, Section 5.4.1.

These methods are applied to solve randomly generated linear Hamiltonian systems, and linear systems arising from the discretization of Hamiltonian PDEs. If not stated otherwise, the dimension of the original space is set to be $2m = 400$ and the dimension of the Krylov subspace is chosen to be $2n = 4$, for all Krylov methods compared. The reference solution is computed using the Cayley transformation (midpoint rule) with step-size 0.004. The solution of the projected system (5.2.2) is obtained with the same method and the same step-size used for the reference solution. All the errors in energy and in first integrals are relative errors.

To obtain a desired global error accuracy on $[0, T]$ for large T, we either use a sufficiently large dimension of the Krylov subspace or use a time-stepping (restart) procedure. More precisely, this entails subdividing $[0, T]$ into subintervals $[t_r, t_{r+1}]$ and performing the projection on each subinterval recomputing the basis of the Krylov subspace with starting vector $y_r \approx y(t_r)$. In the experiments, we use subintervals of size $t_{r+1} - t_r = 0.2$. The restart procedure is of practical interest because it allows to use a Krylov subspace of low dimension. In exact arithmetic the first integrals would be preserved exactly, however, due to the propagation of roundoff errors, a small linear drift in the preserved quantities is observed. The numerical experiments show that the drift in the energy error can be lessened by applying the restart technique. However, the restart compromises the preservation of the first integrals of Propositions 6.1 and 5.4 for APM and APMH simply because the basis V_n is recomputed on each subinterval.

5.5.1 Randomly generated Hamiltonian matrices

Case $JA = AJ$: APM

In the experiment considered in Figure 5.1 (left), $H = J^{-1}A$ is block diagonal, symmetric and positive definite but with no particular extra structure. There is a clear drift in the energy for the APM, and the energy is preserved for the APMH and SLPM. Similar experiments show that the global error of APMH and SLPM is bounded, while the global error of the APM is not (these errors are not reported here). If we apply the APM to an example where $JA = AJ$, the first integrals are preserved and the energy error and global error are bounded, see [12].

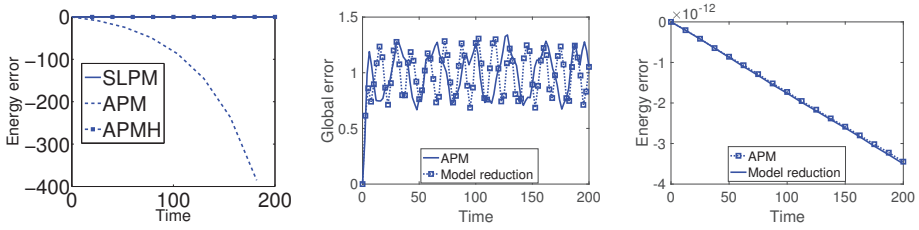


Figure 5.1: Left: Methods without the restart applied to a block diagonal matrix A. Energy drift for APM and energy conservation for SLPM and APMH. **Middle**: Global error of APM and Model reduction versus time. **Right**: Relative energy error of APM and Model reduction versus time.

Case $H_{1,2} = O$, $H_{2,2} = I$: Model reduction

In this numerical test, we consider a Hamiltonian matrix A of the special form (5.4.4) with an initial vector of the form $y_0 = (0, p_0^T)^T$ and we apply the APM to this system. For comparison, we use the model reduction procedure described in Section 5.4.2: we generate the orthogonal matrix V_n using the Arnoldi algorithm with matrix $-H_{11}$ and vector p_0. The methods behave as predicted, (see Figure 5.1, the middle and right figures). The experiment confirms that the APM in this case behaves as the model reduction method and preserves the energy. A small linear drift is observed at the level of roundoff and we will consider this error propagation in the next subsection.

Full matrices: Comparison of APMH, SLPM, BJPM

In this subsection, we consider a randomly generated, full Hamiltonian matrix $A = JH$. In Figure 5.2, we report numerical results for the methods APMH, SLPM and BJPM without restart. The left panel of Figure 5.2, reports the relative energy error for the methods. The right panel of Figure 5.2, illustrates the convergence of the methods: the global error at $T = 2$ decreases when the dimension of the Krylov subspace increases.

APMH is the method that better preserves the energy, but a linear error growth in time at the level of roundoff can be observed for all the methods and also in the error of the first integrals for APMH. To examine this propagation of roundoff errors, we compare the relative energy error and the error in the Cayley transformation as a function of time, see middle panel of Figure 5.2. For $t_k = \Delta t\, k$, we denote the Cayley transformation approximating $\exp(t_k T_n)$ by $\mathrm{Cay}(t_k T_n) := \left((I - \frac{\Delta t}{2} T_n)^{-1}(I + \frac{\Delta t}{2} T_n)\right)^k$. The error in the Cayley transformation is measured by verifying the orthogonality of the

matrix $\mathrm{Cay}(t_k T_n)$. After one step ($t_1 = \Delta t$), this error is close to machine accuracy, $\|(\mathrm{Cay}(\Delta t T_n))^T \mathrm{Cay}(\Delta t T_n)) - I\|_2 = 1.1224e-16$, but we see that $\|(\mathrm{Cay}(t_k T_n))^T \mathrm{Cay}(t_k T_n)) - I\|_2$ grows with t_k and comparably to the relative energy error. Likely, this error is the main cause of the roundoff error propagation in the energy. In this experiment, we have chosen $\Delta t := 2^{-s}$, where s is the minimum positive integer such that $2^{-s}\|T_n\|_1 \leq \frac{1}{2}$, see for example [28].

In Figure 5.3 (left), we see that the roundoff error drift is mitigated by applying the restart technique. In the right figure, we observe that for the methods with restart, the global error behaves well and stops increasing after a certain time.

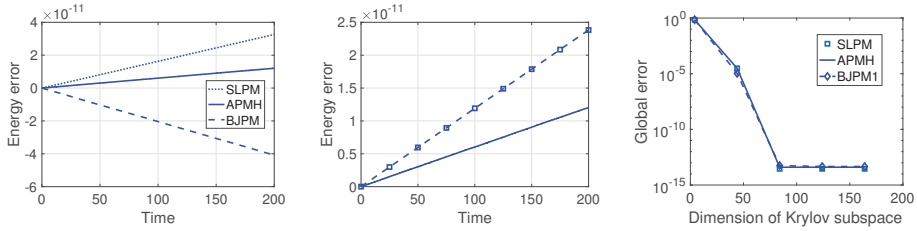


Figure 5.2: Krylov projection methods applied to full matrices. **Left**: energy error for SLPM, APMH and BJPM, methods without restart. In this experiment $\Delta t := 2^{-s}$, where $s > 0$ is such that $2^{-s}\|T_n\|_1 \leq \frac{1}{2}$. **Middle**: relative energy error of APMH (solid line), reference line $\|(\mathrm{Cay}(t_k T_n))^T \mathrm{Cay}(t_k T_n)) - I\|_2$ (dotted square). Error in orthogonality and skew-symmetry: $\|V_n^T H V_n - I\|_2 = 2.8728e-16$, $\|T_n^T + T_n\|_2 = 0$ and $\|(\mathrm{Cay}(\Delta t T_n))^T \mathrm{Cay}(\Delta t T_n)) - I\|_2 = 1.1224e-16$. Same step-size as left panel. **Right**: global error versus dimension of the Krylov subspace.

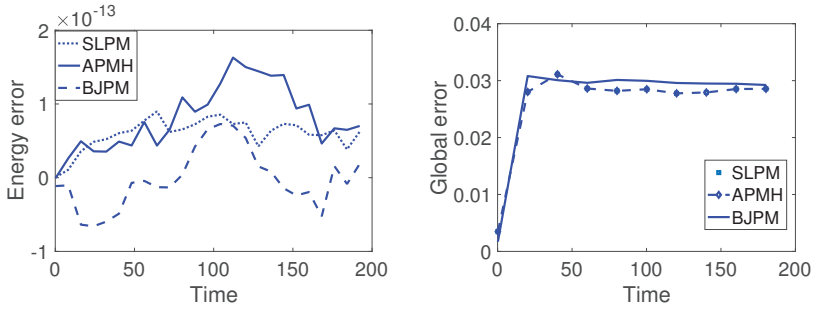


Figure 5.3: Left: energy error. **Right**: global error, methods with restart.

5.5.2 Hamiltonian PDEs

In this section we apply the methods to Hamiltonian PDEs, including the wave equations and the Maxwell's Equations.

Wave equation

We consider the 2D wave equations

$$\dot{\phi} = \psi, \qquad \dot{\psi} = \triangle\phi,$$

on $[0,1] \times [0,1]$ with homogeneous Dirichlet boundary conditions $\phi(t,0,y) = \phi(t,1,y) = \phi(t,x,0) = \phi(t,x,1) = 0$ and a randomly generated initial vector. Semi-discretizing on an equispaced grid $x_i = i\Delta x$ and $y_j = j\Delta y$, $\Delta x = \Delta y$, $i,j = 0,\ldots,N$ and assuming $u(x_i,y_j) \approx U_{i,j}$, we obtain a system

$$\dot{U} = AU, \quad U(0) = U_0, \qquad A = \begin{bmatrix} 0 & I \\ G & 0 \end{bmatrix}$$

with G the discrete 2D Laplacian obtained by using central differences. This is a Hamiltonian system with energy $\mathcal{H} = \frac{1}{2}U^T JAU \equiv \frac{1}{2}U(0)^T JAU(0)$. We perform experiments with all the Krylov projection methods discussed in this paper. The left figure in Figure 5.4 shows that all the methods are energy-preserving.

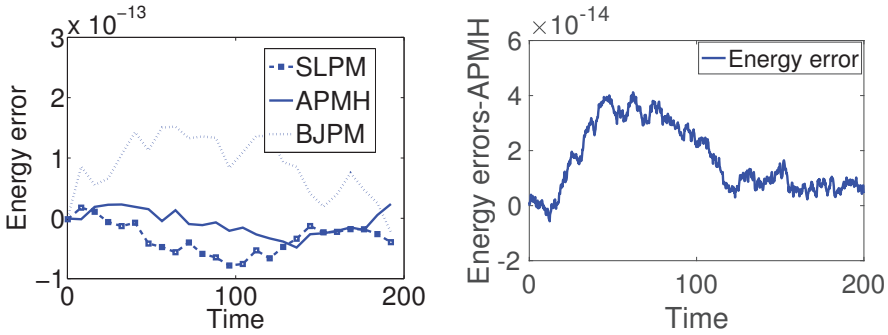


Figure 5.4: Left: energy error for Wave equation in $2d$; methods with restart are considered and the dimension of the problem is 392, namely $N = 15$. **Right**: energy error for Maxwell's equations in 1d; method with restart is considered.

1D Maxwell's equations

We consider 1D Maxwell's equations

$$\begin{aligned} \partial_t E &= \partial_x B, \\ \partial_t B &= \partial_x E \end{aligned} \tag{5.5.1}$$

for $x \in [0,1]$ and $t > 0$ with boundary conditions $E(0,t) = E(1,t) = 0$, $B_x(0,t) = B_x(1,t) = 0$ and initial conditions $E(x,0) = \sin(\pi x)$ and $B(x,0) = \cos(\pi x)$. After semi-discretization with $E(x_i,t) \approx E_i(t)$ and $B(x_i,t) \approx B_i(t)$, $i = 0,\dots,N$, we get a system of ODEs

$$\dot{U} = \bar{S}DU, \quad U(0) = U_0, \tag{5.5.2}$$

where $U = [E_1,...,E_{N-1},B_0,...,B_N]^T$ and

$$\bar{S} = \frac{1}{2h}\begin{bmatrix} 0_{N-1,N+1} & G \\ -G^T & 0_{N+1,N-1} \end{bmatrix}, \qquad G = \begin{bmatrix} -2 & 0 & 1 & & & & \\ & -1 & 0 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 0 & 1 & \\ & & & & -1 & 0 & 2 \end{bmatrix}$$

and $D = \mathrm{diag}(I_{N-1}, \frac{1}{2}, I_{N-1}, \frac{1}{2})$. Equation (5.5.2) fits the framework of section 5.3, with $\bar{S}$ skew-symmetric and D symmetric and positive definite; therefore APMH can be applied to this problem. The numerical approximation of U obtained applying the APMH preserves the first integrals $\mathcal{H}_k(\bar{U})$ of Proposition 6.1. The tables about preservations of first integrals are not reported here. In the right panel of Figure 5.4, we consider the Maxwell equation (5.5.1) with fixed and given initial value and also the restart technique is used. We observe that the energy is preserved well.

5.5.3 Numerical results for 3D Maxwell's equations

We consider 3D Maxwell's equations in CGS units for the electromagnetic field in a vacuum

$$\begin{aligned} \partial_t E &= -c\nabla \times B, \\ \partial_t B &= c\nabla \times E. \end{aligned} \tag{5.5.3}$$

The boundary conditions are zero and the initial conditions are randomly generated for both fields. We consider $c = 1$. We get the following Hamiltonian system after semi-discretization:

$$\dot{U} = AU, \quad A = \begin{bmatrix} 0 & -G_1 \\ G_1 & 0 \end{bmatrix}, \qquad U(0) = U_0, \tag{5.5.4}$$

where $U = [E^x_{1,1,1}, ..., E^z_{N-1,N-1,N-1}, B^x_{1,1,1}, ..., B^z_{N-1,N-1,N-1}]^T$ and G_1, symmetric and of the size $3(N-1)^3$, is the discretization of the curl operator $\nabla\times$.

Remark 5.3. *The matrix A is skew-symmetric in equation* (5.5.4)*. Therefore the APMH with $J = A, H = I$ applied to the system* (5.5.4)*, equals the APM and preserves the first integrals $\mathcal{H}_k(\bar{U})$ of Proposition 6.1.*

Remark 5.4. *Equation* (5.5.4) *can be rewritten as a Hamiltonian equation $\dot{U} = JHU$, with $H = J^{-1}A$ a symmetric matrix. Therefore we can also apply SLPM and BJPM to system* (5.5.3) *and the energy $\mathcal{H}(U) = \frac{1}{2}U^T J^{-1} AU$ is preserved. However, APMH cannot be used here because H is not a positive definite matrix, and the inner product $\langle\cdot,\cdot\rangle_H$ is degenerate. This can lead to instabilities and both global error and energy error might blow up during the iteration.*

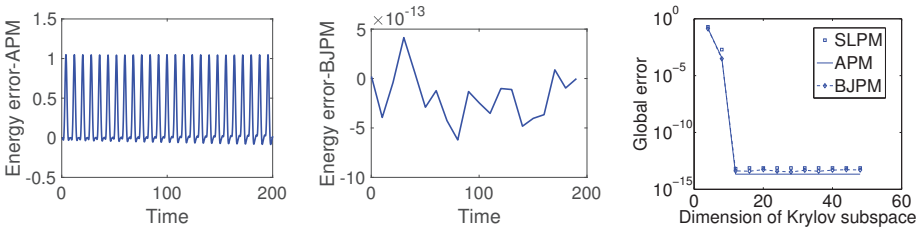


Figure 5.5: The dimension of the problem is 384, namely $N = 5$. **Left**: the energy error considered here is following from the energy in Remark 5.3. **Middle**: the method with the restart technique is used, and the energy error is based on Remark 5.4. **Right**: we consider L^2 norm of the global error at $t = T = 2$ as a function of the dimension of the Krylov subspace.

The left figure in Figure 5.5 shows that the energy error of APM is bounded as stated in Remark 5.3. The middle one in Figure 5.5 shows that the energy $\mathcal{H}(U) = \frac{1}{2}U^T J^{-1} AU$ is preserved for BJPM as stated in Remark 5.4. In the right panel of Figure 5.5, we report convergence plots for the methods. As the dimension of the Krylov subspace increases, the global error decreases very fast for all the methods. All the methods converge well also for larger end time, such as $T = 200$. Also in this example, we observed a small linear growth in the error of the first integrals, due to the propagation of round-off errors (figures are not presented here).

Bibliography

[1] S. Agoujil, A. Bentbib, and A. Kanber, *A structure preserving approximation method for Hamiltonian exponential matrices*, Appl. Numer. Math., 62 (2012), pp. 1126–1138.

[2] A. ARCHID AND A. H. BENTBIB, *Approximation of the matrix exponential operator by a structure-preserving block Arnoldi-type method*, Appl. Numer. Math., 75 (2014), pp. 37–47.

[3] V. I. ARNOL'D, B. DUBROVIN, A. KIRILLOV, AND I. KRICHEVER, *Dynamical Systems IV: Symplectic Geometry and Its Applications*, vol. 4, Springer Science & Business Media, 2001.

[4] W. E. ARNOLDI, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[5] P. BENNER AND H. FAβBENDER, *An implicitly restarted symplectic Lanczos method for the Hamiltonian eigenvalue problem*, Linear Algebra Appl., 263 (1997), pp. 75–111.

[6] P. BENNER, H. FASSBENDER, AND M. STOLL, *A Hamiltonian Krylov–Schur-type method based on the symplectic Lanczos process*, Linear Algebra Appl., 435 (2011), pp. 578–600.

[7] P. BENNER, V. MEHRMANN, AND D. C. SORENSEN, *Dimension reduction of large-scale systems*, vol. 35, Springer, 2005.

[8] P. BENNER, V. MEHRMANN, AND H. XU, *A numerically stable structure-preserving method for computing the eigenvalues of real Hamiltonian or symplectic pencils*, Numer. Math., 78 (1998), pp. 329–358.

[9] M. A. BOTCHEV AND J. G. VERWER, *Numerical integration of damped Maxwell equations*, SIAM J. Sci. Comput., 31 (2009), pp. 1322–1346.

[10] L. BRUGNANO, F. IAVERNARO, AND D. TRIGIANTE, *Hamiltonian boundary value methods (energy preserving discrete line integral methods)*, JNAIAM. J. Numer. Anal. Ind. Appl. Math., 5 (2010), pp. 17–37.

[11] E. CELLEDONI, V. GRIMM, R. I. MCLACHLAN, D. I. MCLAREN, D. O'NEALE, B. OWREN, AND G. R. W. QUISPEL, *Preserving energy resp. dissipation in numerical PDEs using the "average vector field" method*, J. Comput. Phys., 231 (2012), pp. 6770–6789.

[12] E. CELLEDONI AND L. LI, *Symplectic Lanczos Arnoldi method for solving linear Hamiltonian systems: preservation of energy and other invariants*, in Progress in industrial mathematics at ECMI 2016, vol. 26 of Math. Ind., Springer, Cham, 2017, pp. 553–559.

[13] T. EIROLA AND A. KOSKELA, *Krylov integrators for Hamiltonian systems*, BIT, (2018), pp. 1–20.

[14] H. FASSBENDER, *A detailed derivation of the parameterized SR algorithm and the symplectic Lanczos method for Hamiltonian matrices*, Preprint, (2006).

[15] K. FENG AND M.-Z. QIN, *The symplectic methods for the computation of Hamiltonian equations*, in Numerical Methods for Partial Differential Equations, Springer, 1987, pp. 1–37.

[16] H. GOLDSTEIN, *Classical Mechanics*, Pearson Education India, 2011.

[17] E. HAIRER, C. LUBICH, AND G. WANNER, *Geometric numerical integration*, vol. 31 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 2006.

[18] R. A. LABUDDE AND D. GREENSPAN, *Energy and momentum conserving methods of arbitrary order for the numerical integration of equations of motion*, Numer. Math., 25 (1975), pp. 323–346.

[19] S. LALL, P. KRYSL, AND J. E. MARSDEN, *Structure-preserving model reduction for mechanical systems*, Phys. D, 184 (2003), pp. 304–318.

[20] L. LOPEZ AND V. SIMONCINI, *Preserving geometric properties of the exponential matrix by block Krylov subspace methods*, BIT, 46 (2006), pp. 813–830.

[21] J. E. MARSDEN AND A. WEINSTEIN, *The Hamiltonian structure of the Maxwell-Vlasov equations*, Phys. D, 4 (1982), pp. 394–406.

[22] R. MCLACHLAN, *Symplectic integration of Hamiltonian wave equations*, Numer. Math., 66 (1993), pp. 465–492.

[23] R. I. MCLACHLAN, G. QUISPEL, AND N. ROBIDOUX, *Geometric integration using discrete gradients*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 357 (1999), pp. 1021–1045.

[24] R.-D. RICHTMYER AND K.-W. MORTON, *Difference Methods for Initial-value Problems*, Interscience Publishers John Wiley & Sons, Inc., Academia Publishing House of the Czechoslovak Acad, 1967.

[25] Y. SUN AND P. TSE, *Symplectic and multisymplectic methods for Maxwell's equations*, J. Comput. Phys., 230 (2010), pp. 2076–2094.

[26] M. E. TAYLOR, *Partial differential equations I. Basic theory*, vol. 115 of Applied Mathematical Sciences, Springer, New York, second ed., 2011.

[27] A. VAN DER SCHAFT AND D. JELTSEMA, *Port-Hamiltonian systems theory: An introductory overview*, Found. and Trends in Syst. and Control, 1 (2014), pp. 173–378.

[28] R. C. WARD, *Numerical computation of the matrix exponential with accuracy estimate*, SIAM J. on Num. Anal., 14 (1977), pp. 600–610.

[29] D. S. WATKINS, *On Hamiltonian and symplectic Lanczos processes*, Linear Algebra Appl., 385 (2004), pp. 23–45.

[30] G. ZHONG AND J. E. MARSDEN, *Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators*, Phys. Lett. A, 133 (1988), pp. 134–139.

Appendix

Algorithm 5.1 Arnoldi's algorithm with modified inner product

Input: a matrix $J \in \mathbb{R}^{m\times m}$, $H \in \mathbb{R}^{m\times m}$, a vector $b \in \mathbb{R}^m$, a number $n \in \mathbb{N}$ and a tolerance $\iota \in \mathbb{R}$.
$A = JH$
$v_1 = b/\langle b, b\rangle_H^{\frac{1}{2}}$
for $j = 1 : n$ **do**
  compute $w_j = Av_j$
  for $k = 1 : 2$ **do**
    for $i = 1 : j$ **do**
      $h_{i,j} = \langle v_i, w_j\rangle_H$
      $w_j = w_j - h_{i,j}v_i$
    end for
  end for
  $h_{j+1,j} = \langle w_j, w_j\rangle_H^{\frac{1}{2}}$
  if $h_{j+1,j} < \iota$ **then**
    Stop
  end if
  $v_{j+1} = w_j/h_{j+1,j}$
end for
Output: $T_n, V_n, v_{n+1}, h_{n+1,n}$.

Algorithm 5.2 Algorithm to generate V_n (by QR factorization)

Matrix $A \in \mathbb{R}^{2m\times 2m}$, vector $b \in \mathbb{R}^{2m}$, number $n \in \mathbb{N}$.
$v = b$
$K_n = v$
for $i = 1 : n-1$ **do**
  $v = Av$
  $K_n = [K_n, v]$
end for
$K_n^q = K_n(1 : m, :)$
$K_n^p = K_n(m+1 : 2m, :)$
$[Q, R] = qr([K_n^q, K_n^p])$
$V_n = Q(:, 1 : k), \quad k = \text{rank}([K_n^q, K_n^p]) \le 2n$
Output V_n.

Rounding error analysis for the energy error of APMH

Lu Li

Rounding error analysis for the energy error of APMH

In this chapter, we study the propagation of roundoff errors in the energy error of the modified Arnoldi projection method in Paper 4 in this thesis. In the first section, we recall the method applied to a linear Hamiltonian system. In the second section, we consider an analysis for the propagation of rounding errors in the computation of numerical energy.

6.1 Energy-preserving methods for linear Hamiltonian systems based on Arnoldi algorithm

Consider a linear Hamiltonian initial value problem of the form

$$\dot{y} = f(y) = JHy, \quad y(t_0) = y_0 \in \mathbb{R}^m, \qquad J = J_m = \begin{bmatrix} 0 & I_{m/2} \\ -I_{m/2} & 0 \end{bmatrix}, \quad (6.1.1)$$

where m is an even number, and $H \in \mathbb{R}^{m\times m}$ is a symmetric and positive definite matrix. We denote by APMH the projection method based on the Arnoldi algorithm with respect to the inner product $\langle\cdot,\cdot\rangle_H$ [3]. With this inner product, the Arnoldi algorithm generates an H-orthonormal basis. Suppose the first n vectors of this basis are stored in the $m \times n$ matrix V_n. The algorithm generates an $n \times n$ skew-symmetric matrix T_n such that

$$AV_n = V_n T_n + w_{n+1} e_n^T, \quad w_{n+1} = h_{n+1,n} v_{n+1},$$
$$V_n{}^T H V_n = I_n, \qquad V_n{}^T H w_{n+1} = 0.$$

The APMH method is to solve y for (6.1.1) by considering the Krylov projection method

$$y_H := V_n z, \quad \text{where} \quad z \quad \text{satisfies} \quad \dot{z} = T_n z, \quad z(t_0) = V_n{}^T H y_0. \quad (6.1.2)$$

Proposition 6.1. *[3] The numerical approximation y_H for the solution y of* (6.1.1) *preserves the energy $\mathcal{H}(y_H) = \frac{1}{2} y_H^T H y_H$, and also the following first integrals:*

$$\mathcal{H}_k(y_H) = \frac{1}{2} y_H^T H V_n (T_n)^{2k} V_n{}^T H y_H,$$

for all $k = 0, 1, \ldots$.

In particular, we focus on the energy preservation here. Note that one can not get the exact solution for z in the projected system (6.1.2), and a numerical

time-stepping method must be used. To ensure the energy preservation of APMH, an energy-preserving numerical method is preferred for solving z, and we choose the midpoint rule in our experiments in [3]. Even if the APMH is an energy-preserving method for linear Hamiltonian systems, a drift in the energy at the level of roundoff can be observed in numerical experiments, see [3]. In this note we consider an analysis of the propagation of roundoff errors in the preservation of energy for this Krylov projection method.

6.2 Propagation of rounding errors in the energy

Let us denote the conversion of a real number into its floating point representation by $\mathrm{fl}(\cdot)$. Consider the time interval $[0, T]$ and denote the time step size by $\Delta t = T/N$, where N is the number of the iterations for solving $y(T)$. From the analysis in [4], a better choice of the time step size was to be $\Delta t := 2^{-s}$, where s satisfies $2^{-s}\|T_n\|_1 \leq 1/2$. We chose s to be the minimum positive negative number such that $2^{-s}\|T_n\| \leq 1/2$ for the experiments in [3]. In the next, we will introduce two lemmas which provide the basis for our analysis.

Lemma 6.1. *[2] Denote the machine precision by ε, for two vectors $x, y \in \mathbb{R}^m$, we have*

$$|fl(x^T y) - x^T y| \leq m\varepsilon |x^T||y| + \mathcal{O}(\varepsilon^2).$$

Lemma 6.2. *[2] Consider a $n \times m$ matrix A and a $m \times r$ matrix B, and suppose $m\varepsilon \leq 0.01$, we have*

$$|fl(AB) - AB| \leq m\varepsilon |A||B| + \mathcal{O}(\varepsilon^2).$$

In section 6.1 we have seen that the APMH method can be divided in three main parts:

- Performing the modified Arnoldi algorithm to compute V_n and T_n;
- Finding the numerical solution of the projected system (6.1.2) with the midpoint rule, which amounts to applying the matrix $C = (I - \frac{\Delta t}{2} T_n)^{-1}(I + \frac{\Delta t}{2} T_n)$ to a vector repeatedly several times $z_k = C z_{k-1}$;
- Calculating $y_k = V_n z_k$, for $k = 1, 2, \ldots, N$.

Propagation of rounding errors occurs in all the three parts. The rounding errors in the first part has already been determined by the matrix factorization. So in this note we will focus on the roundoff analysis in the last two parts. Let us

suppose $\mathrm{fl}(V_n) = V_n + \varepsilon E_1$, and $\mathrm{fl}(T_n) = T_n + \varepsilon E_2$ by the modified Arnoldi algorithm [3]. According to the roundoff analysis in [1] and by similar calculation, we can deduce that $|E_1|$ and $|E_2|$ depends on A and y_0. Let us also assume that $\mathrm{fl}(C) = C + \varepsilon E_3$. By our calculation $|E_3|$ depends on Δt, T_n and E_2. In this note, we focus more on the propagation of rounding errors in applying the midpoint rule to solve z_k and y_k, for $k = 1, 2, \ldots$, and we will not present the details about the bound of $|E_1|$, $|E_2|$ and $|E_3|$. In the following two propositions, we will consider the propagation of rounding errors at each time step $t_k = k\Delta t$ based on the relations $\mathrm{fl}(V_n) = V_n + \varepsilon E_1$, $\mathrm{fl}(T_n) = T_n + \varepsilon E_2$ and $\mathrm{fl}(C) = C + \varepsilon E_3$ in the computation of z_k for $k = 0, 1, \ldots, N$.

Proposition 6.2. *Let $er_{z_0} := fl(z_0) - z_0$, then we have*

$$er_{z_0} = \varepsilon E_1^T H y_0 + V_n^T b_{01} + b_{02} + \mathcal{O}(\varepsilon^2), \tag{6.2.1}$$

with $|b_{01}| \le m\varepsilon |H||y_0| + \mathcal{O}(\varepsilon^2)$, and $|b_{02}| \le m\varepsilon |V_n^T||H||y_0| + \mathcal{O}(\varepsilon^2)$.

Proof. In exact arithmetics we would compute $z_0 = V_n^T H y_0$, however, taking the rounding errors into account, we get

$$\mathrm{fl}(z_0) = \mathrm{fl}\big(\mathrm{fl}(V_n)\mathrm{fl}(Hy_0)\big) = (V_n + \varepsilon E_1)^T (Hy_0 + b_{01}) + b_{02},$$

where

$$\begin{aligned} &|b_{01}| \le m\varepsilon |H||y_0| + \mathcal{O}(\varepsilon^2), \\ &|b_{02}| \le m\varepsilon |V_n^T + \varepsilon E_1^T||Hy_0 + b_{01}| + \mathcal{O}(\varepsilon^2) \le m\varepsilon |V_n^T||H||y_0| + \mathcal{O}(\varepsilon^2). \end{aligned}$$

Therefore we obtain

$$\begin{aligned} er_{z_0} &= \varepsilon E_1^T H y_0 + V_n^T b_{01} + b_{02} + \mathcal{O}(\varepsilon^2), \\ |er_{z_0}| &\le \varepsilon |E_1^T||H||y_0| + 2m\varepsilon |V_n^T||H||y_0| + \mathcal{O}(\varepsilon^2). \end{aligned}$$

□

Proposition 6.3. *Let $er_{z_k} := fl(z_k) - z_k$, and $b_{k+1} := fl(z_{k+1}) - fl(C)fl(z_k)$, then we have*

$$er_{z_k} = C^k er_{z_0} + \varepsilon \sum_{i=0}^{k-1} (C^{k-1-i} E_3 C^i) z_0 + \sum_{i=1}^{k} C^{k-i} b_i + \mathcal{O}(\varepsilon^2), \quad k = 1, 2, \ldots, N, \tag{6.2.2}$$

with $|b_i| \le n\varepsilon |C|^i |z_0| + \mathcal{O}(\varepsilon^2)$, $i = 1, 2, \ldots, k$.

Proof. This proof is done by induction on k. For $k = 1$, the exact computation

gives $z_1 = Cz_0$, however, taking the rounding errors into account, we have

$$\mathrm{fl}(z_1) = (C + \varepsilon E_3)(z_0 + er_{z_0}) + b_1,$$

where $|b_1| \le n\varepsilon(|C| + \varepsilon|E_3|)(|z_0| + |er_{z_0}|) + \mathcal{O}(\varepsilon^2) = n\varepsilon|C||z_0| + \mathcal{O}(\varepsilon^2)$. This implies that

$$er_{z_1} = Cer_{z_0} + \varepsilon E_3 z_0 + b_1 + \mathcal{O}(\varepsilon^2).$$

Suppose (6.2.2) is satisfied for k. From

$$\mathrm{fl}(z_{k+1}) = \mathrm{fl}(Cz_k) = (C + \varepsilon E_3)(z_k + er_{z_k}) + b_{k+1},$$

where $|b_{k+1}| \le n\varepsilon(|C| + \varepsilon|E_3|)(|z_k| + |er_{z_k}|) + \mathcal{O}(\varepsilon^2) = n\varepsilon|C|^{k+1}|z_0| + \mathcal{O}(\varepsilon^2)$, we can obtain

$$\begin{aligned} er_{z_{k+1}} &= Cer_{z_k} + \varepsilon E_3 z_k + b_{k+1} + \mathcal{O}(\varepsilon^2) \\ &= C\big(C^k er_{z_0} + \varepsilon \sum_{i=0}^{k-1} (C^{k-1-i} E_3 C^i) z_0 + \sum_{i=1}^{k} C^{k-i} b_i\big) + \varepsilon E_3 z_k + b_{k+1} + \mathcal{O}(\varepsilon^2), \\ &= C^{k+1} er_{z_0} + \varepsilon \sum_{i=0}^{k} (C^{k-i} E_3 C^i) z_0 + \sum_{i=1}^{k+1} C^{k+1-i} b_i + \mathcal{O}(\varepsilon^2). \end{aligned}$$

□

In the next proposition, we will consider the rounding error in y_k based on the above analysis for z_k.

Proposition 6.4. *Let $er_{y_k} := fl(y_k) - y_k$, and $c_k := fl(y_k) - fl(V_n)fl(z_k)$, then we have*

$$er_{y_k} = V_n er_{z_k} + \varepsilon E_1 z_k + c_k + \mathcal{O}(\varepsilon^2), \quad \text{for} \quad k = 1, 2, \ldots, N. \tag{6.2.3}$$

Proof. Recall that $y_k = V_n z_k$ and with the rounding errors considered, we have

$$\mathrm{fl}(y_k) = (V_n + \varepsilon E_1)(z_k + er_{z_k}) + c_k,$$

where $|c_k| \le n\varepsilon(|V_n| + \varepsilon|E_1|)(|z_k| + |er_{z_k}|) + \mathcal{O}(\varepsilon^2) = n\varepsilon|V_n||z_k| + \mathcal{O}(\varepsilon^2)$. Therefore, we obtain

$$er_{y_k} = V_n er_{z_k} + \varepsilon E_1 z_k + c_k + \mathcal{O}(\varepsilon^2).$$

□

Based on the estimation of rounding error in y_k, we give the rounding error in the energy in the following proposition.

Proposition 6.5. *Let* $er_{H_k} := fl(\mathcal{H}(y_k)) - \mathcal{H}(y_k)$, *with* $\mathcal{H}(y_k) = \frac{1}{2} y_k^T H y_k$ *defined as in Proposition 6.1, and* $d_k = fl(Hy_k) - Hfl(y_k)$, $f_k = fl(y_k^T H y_k) - fl(y_k)^T fl(Hy_k)$. *We then have*

$$er_{H_k} = 2y_k^T H er_{y_k} + y_k^T d_k + f_k + \mathcal{O}(\varepsilon^2), \quad for \quad k = 1, 2, \ldots, N. \tag{6.2.4}$$

Proof. From $\mathrm{fl}(y_k) = y_k + er_{y_k}$, we have

$$\mathrm{fl}(Hy_k) = Hy_k + Her_{y_k} + d_k,$$

where $|d_k| \leq m\varepsilon |H|(|er_{y_k}| + |y_k|) + \mathcal{O}(\varepsilon^2) = m\varepsilon |H||y_k| + \mathcal{O}(\varepsilon^2)$. Therefore we get

$$\begin{aligned} \mathrm{fl}(y_k^T H y_k) &= (y_k + er_{y_k})^T (Hy_k + Her_{y_k} + d_k) + f_k, \\ er_{H_k} &= 2y_k^T H er_{y_k} + y_k^T d_k + f_k + \mathcal{O}(\varepsilon^2). \end{aligned}$$

where

$$\begin{aligned} |f_k| &\leq m\varepsilon (|er_{y_k}| + |y_k|)^T (|H||er_{y_k}| + |H||y_k| + |d_k|) + \mathcal{O}(\varepsilon^2) \\ &= m\varepsilon |y_k|^T |H||y_k| + \mathcal{O}(\varepsilon^2). \end{aligned}$$

□

In order to give the rounding error in energy error, we will consider the rounding error in the original energy in the following proposition.

Proposition 6.6. *Let* $er_{H_0} := fl(y_0^T H y_0) - y_0^T H y_0$, $g_0 := fl(Hy_0) - Hy_0$ *and* $p_0 := fl(y_0^T H y_0) - y_0^T fl(Hy_0)$, *we have*

$$er_{H_0} = y_0^T g_0 + p_0, \tag{6.2.5}$$

with $|g_0| \leq m\varepsilon |H||y_0| + \mathcal{O}(\varepsilon^2)$ *and* $|p_0| \leq m\varepsilon |y_0^T||H||y_0| + \mathcal{O}(\varepsilon^2)$

Proof. From $\mathrm{fl}(Hy_0) = Hy_0 + g_0$, where $|g_0| \leq m\varepsilon |H||y_0| + \mathcal{O}(\varepsilon^2)$, we have

$$\mathrm{fl}(y_0^T H y_0) = y_0^T (Hy_0 + g_0) + p_0,$$

where $|p_0| \leq m\varepsilon |y_0^T||Hy_0 + g_0| + \mathcal{O}(\varepsilon^2) = m\varepsilon |y_0^T||H||y_0| + \mathcal{O}(\varepsilon^2)$. Therefore we can obtain

$$er_{H_0} = y_0^T g_0 + p_0.$$

□

Based on the results in proposition 6.5 and 6.6, we finally give the rounding error in the energy error in the following theorem.

Theorem 6.1. *Let* $er_k := fl(\frac{y_k^T H y_k - y_0^T H y_0}{y_0^T H y_0})$ *and*

$$\beta_1 = \frac{2\varepsilon \sum_{i=0}^{k-1} z_0^T \left((C^T)^{1+i} E_3 C^i\right) z_0}{z_0^T z_0}$$
$$\beta_2 = \frac{2\sum_{i=1}^{k} z_0^T (C^T)^i b_i}{z_0^T z_0}$$
$$\beta_3 = \frac{2z_0^T er_{z_0} + 2y_k^T H(\varepsilon E_1 z_k + c_k) + y_k^T d_k + f_k - (y_0^T g_0 + p_0)}{z_0^T z_0}$$

then we have

$$er_k = \beta_1 + \beta_2 + \beta_3 + \mathcal{O}(\varepsilon^2), \quad for \quad k = 1, 2, \ldots, N. \tag{6.2.6}$$

Proof. From (6.2.4), (6.2.5) and (6.2.3), (6.2.2), and using $C^T C = I$, $y_k^T H y_k = y_0^T H y_0$, we can obtain

$$\begin{aligned}
&\mathrm{fl}(y_k^T H y_k - y_0^T H y_0) \\
&= er_{H_k} - er_{H_0} + \mathcal{O}(\varepsilon^2) \\
&= 2y_k^T H e r_{y_k} + y_k^T d_k + f_k - (y_0^T g_0 + p_0) + \mathcal{O}(\varepsilon^2) \\
&= 2y_k^T H(V_n er_{z_k} + \varepsilon E_1 z_k + c_k) + y_k^T d_k + f_k - (y_0^T g_0 + p_0) + \mathcal{O}(\varepsilon^2) \\
&= 2z_0^T (C^T)^k er_{z_k} + 2y_k^T H(\varepsilon E_1 z_k + c_k) + y_k^T d_k + f_k - (y_0^T g_0 + p_0) + \mathcal{O}(\varepsilon^2) \\
&= 2\varepsilon \sum_{i=0}^{k-1} z_0^T \left((C^T)^{1+i} E_3 C^i\right) z_0 + 2\sum_{i=1}^{k} z_0^T (C^T)^i b_i + 2z_0^T er_{z_0} \\
&\quad + 2y_k^T H(\varepsilon E_1 z_k + c_k) + y_k^T d_k + f_k - (y_0^T g_0 + p_0) + \mathcal{O}(\varepsilon^2).
\end{aligned}$$

Note that $y_0^T H y_0 = z_0^T z_0$, and

$$\mathrm{fl}(\frac{y_k^T H y_k - y_0^T H y_0}{y_0^T H y_0}) = \frac{\mathrm{fl}(y_k^T H y_k - y_0^T H y_0)}{y_0^T H y_0} + \mathcal{O}(\varepsilon^2),$$

then we can obtain (6.2.6).

□

In the next proposition, we show that the term β_3 in theorem 6.1 can be bounded by a constant independent of the simulating time.

Proposition 6.7.

$$|\beta_3| \le \big(2\sigma_{max}(|E_1^T||H||V_n|) + 4m\sigma_{max}(|V_n^T||H||V_n|) + 2n + 2m\big)\varepsilon,$$

where $\sigma_{max}(A)$ is the largest eigenvalue of A.

Proof. Conisder c_k in Proposition 6.4, and d_k, f_k in Proposition 6.5, we can suppose that there exist $\tilde{c}_k$, $\tilde{d}_k$ and $\tilde{f}_k$ such that $c_k = n\varepsilon V_n\tilde{c}_k$ with $|\tilde{c}_k| \le |z_k|$, $d_k = m\varepsilon HV_n\tilde{d}_k$ with $|\tilde{d}_k| \le |z_k|$ and $f_k = m\varepsilon y_k^T HV_n\tilde{f}_k$ with $|\tilde{f}_k| \le |z_k|$. Thus, we obtain

$$\begin{aligned} &2y_k^T Hc_k + y_k^T d_k + f_k = \varepsilon z_k^T V_n^T HV_n(2n\tilde{c}_k + m\tilde{d}_k + m\tilde{f}_k), \\ &|2y_k^T Hc_k + y_k^T d_k + f_k| \le (2n+2m)\varepsilon|z_k|^T|z_k|. \end{aligned} \tag{6.2.7}$$

By using (6.2.1) and (6.2.5), we have

$$|2z_0^T er_{z_0} - (y_0^T g_0 + p_0)| \le 2\varepsilon|z_0|^T|E_1^T||H||V_n||z_0| + 4m\varepsilon|z_0|^T||V_n^T||H||V_n||z_0| \tag{6.2.8}$$

Note that $z_k^T z_k = z_0^T z_0$, and then from (6.2.7) and (6.2.8), we can obtain

$$\begin{aligned} |\beta_3| &= |2z_0^T er_{z_0} + 2y_k^T H(\varepsilon E_1 z_k + c_k) + y_k^T d_k + f_k - (y_0^T g_0 + p_0)|/|z_0^T z_0| \\ &\le \big(2\sigma_{\max}(|E_1^T||H||V_n|) + 4m\sigma_{\max}(|V_n^T||H||V_n|) + 2n + 2m\big)\varepsilon. \end{aligned}$$

□

We conclude this section with a remark which connects the analysis in theorem 6.1 and the numerical experiment in Figure 5.2 (the middle figure) in chapter 5.

Remark 6.1. *Note from* (6.2.6) *that the rounding errors in the relative energy error mainly consists of three terms, β_1, β_2 and β_3. From proposition 6.7, the bound of β_3 does not increase with the number of iteration N. However, we see clearly that β_1 and β_2 increases with N, or we say with T if $\Delta t = T/N$ is fixed. Therefore β_1 and β_2 are the dominant terms in the rounding error of the energy error. They are mainly due to the propagation of the rounding errors in the calculation of $C^N z_0$. When we calculate $C^N z_0$, we need to multiply with matrix C for N times. Consider that there is an error in the calculation of C, i.e. $\varepsilon E_3 = fl(C) - C$. The rounding errors in the calculation for $C^N z_0$ comes from two aspects: one is due to the propagation of εE_3 which results in the term*[1] *$\frac{1}{2}\beta_1$ in* (6.2.6)*; the other is due to the errors by matrix multiplications in $C^N z_0$*

[1] The rounding error in the energy $(z_0 C^N)^T V_n^T HV_n C^N z_0$ $C^N z_0$ will be doubled compared to the rounding error in $C^N z_0$. Therefore we have the 1/2 in front of β_1 and β_2.

which leads to $\frac{1}{2}\beta_2$ *in* (6.2.6). *Therefore, a reasonable bound for* $|\beta_1 + \beta_2|$ *is* $N \cdot \|fl(C)^T C - I_n\|_2$, *which is close to* $\|(I - \frac{T}{2}T_n)^{-1}(I + \frac{T}{2}T_n)\|_2$. *The numerical experiment, namely the middle figure of Figure 5.2 in chapter 5 in this thesis, coincides with this analysis.*

Bibliography

[1] A. BJÖRCK, *Solving linear least squares problems by Gram-Schmidt orthogonalization*, Nordisk Tidskr. Informations-Behandling, 7 (1967), pp. 1–21.

[2] G. H. GOLUB AND C. F. VAN LOAN, *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, third ed., 1996.

[3] L. LI AND E. CELLEDONI, *Krylov projection methods for linear hamiltonian systems*, Numer. Algorithms, (2019), pp. 1–18.

[4] R. C. WARD, *Numerical computation of the matrix exponential with accuracy estimate*, SIAM J. Numer. Anal., 14 (1977), pp. 600–610.