Journal of Computational Physics 472 (2023) 111697

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Explicit exactly energy-conserving methods for Hamiltonian systems

Stefan Bilbao [a,*], Michele Ducceschi [b], Fabiana Zama [c]

[a] Acoustics and Audio Group, University of Edinburgh, 12 Nicolson Square, Edinburgh, EH8 9DF, United Kingdom
[b] Department of Industrial Engineering, University of Bologna, Bologna, Italy
[c] Department of Mathematics, University of Bologna, Bologna, Italy

ARTICLE INFO

Article history:
Received 29 June 2022
Received in revised form 12 October 2022
Accepted 12 October 2022
Available online 18 October 2022

Keywords:
Hamiltonian systems
Geometric numerical integration
Energy-conserving methods
Explicit methods
Finite difference methods

ABSTRACT

For Hamiltonian systems, simulation algorithms that exactly conserve numerical energy or pseudo-energy have seen extensive investigation. Most available methods either require the iterative solution of nonlinear algebraic equations at each time step, or are explicit, but where the exact conservation property depends on the exact evaluation of an integral in continuous time. Under further restrictions, namely that the potential energy contribution to the Hamiltonian is non-negative, newer techniques based on invariant energy quadratisation allow for exact numerical energy conservation and yield linearly implicit updates, requiring only the solution of a linear system at each time step. In this article, it is shown that, for a general class of Hamiltonian systems, and under the non-negativity condition on potential energy, it is possible to arrive at a fully explicit method that exactly conserves numerical energy. Furthermore, such methods are unconditionally stable, and are of comparable computational cost to the very simplest integration methods (such as Störmer-Verlet). A variant of this scheme leading to a conditionally-stable method is also presented, and follows from a splitting of the potential energy. Various numerical results are presented, in the case of the classic test problem of Fermi, Pasta and Ulam and for nonlinear systems of partial differential equations, including those describing high amplitude vibration of strings and plates.

1. Introduction

The design of exact energy conserving numerical methods for nonlinear Hamiltonian systems goes back at least as far as the work of LaBudde and Greenspan, for the single particle subject to a central potential [1], Marciniak [2] for the N-body problem, Strauss and Vazquez [3] for the nonlinear Klein-Gordon equation, and Greenspan for the nonlinear harmonic oscillator [4], and was greatly generalized by Simo et al. [5] in their work on energy- and momentum-conserving algorithms. The concept of energy conservation in time stepping schemes goes back much further in the case of linear systems—see the discussion and references in [5], particularly with regard to the energy conservation of the Crank-Nicholson method for linear systems. An exact numerical conservation of a "pseudo-energy" (henceforth simply "energy" or "numerical energy" in this article) of a numerical scheme for a conservative system may be contrasted with schemes that conserve energy approximately—see, e.g., early work by Hughes et al. [6]. The same point of view is taken with other geometrical numerical

* Corresponding author.
E-mail address: sbilbao@ed.ac.uk (S. Bilbao).

https://doi.org/10.1016/j.jcp.2022.111697

integration techniques, such as, e.g., symplectic methods, where energy is conserved approximately, following the result from Zhong and Marsden [7] that, under some restrictions, exact energy conservation and symplecticity for Hamiltonian systems cannot be obtained simultaneously [8]. Various studies examine the degree to which energy is conserved for symplectic and non-symplectic methods [9,10]. In this paper, we focus on exact conservation of numerical energy rather than approximate conservation, and we will not consider momentum conservation. Part of the reason for the focus here on exact numerical energy conservation is the possibility of determining conditions for numerical stability as has been pointed out earlier—see [5] and Remark 8 in [11].

Most exact numerical energy-conserving algorithms presented to date are implicit [3,2,5,12], requiring the solution of a nonlinear system of algebraic equations at each time step, normally through an iterative method (such as, e.g., Newton-Raphson). This is obviously a computational bottleneck, and a fully explicit formulation is preferable. Note, however, that for linear systems, explicit and exact energy-conserving methods are available—see, e.g., [13]. Exact explicit and conservative integrators have also been designed on a case-by-case basis—normally new ad hoc stability considerations appear in a problem-dependent way regarding the choice of the time step. See [14] for a treatment of the three-body problem, and [15] for the Kepler problem.

In a recent article by Marazzato et al. [11], an explicit Hamiltonian integrator is presented that preserves a numerical energy exactly, for general nonlinear systems. Two features are worth noting: a) the conserved quantity is defined through a continuous time integration, and thus must itself be approximated, leading effectively to approximate (non-exact) energy conservation in a discrete time implementation—with fine enough approximation, numerical energy conservation to machine precision may be achieved; and b) when the energy of the model system is non-negative, as it commonly is in practice, the numerical energy does not inherit this property, and thus conditions for numerical stability do not immediately follow. Other exactly conservative methods, also reliant on the exact evaluation of a continuous integration have been proposed previously [16,17].

In the context of gradient flows for diffusive systems, recent progress has been made in developing energy-stable methods through a variable transformation representing the energy, generally a nonlinear function of the state, as a quadratic form—such methods are referred to as invariant energy quadratisation (IEQ) approaches [18,19]. The main result is that it is possible to arrive at updating equations for a time-stepping method that are linearly implicit—so that the update depends only on the solution of a linear system, rather than the solution of a system of nonlinear algebraic equations. This allows for the sidestepping of the many difficulties associated with iterative solvers, and reduced computational cost. Alongside the more recently introduced scalar auxiliary variable (SAV) approaches [20], such techniques have been applied to a wide variety of problems [21]. More recently, IEQ/SAV approaches have been applied to Hamiltonian systems, as in the case of diagonally-implicit Runge Kutta methods [22]. For an interesting overview of the relationship between quadratisation techniques and linearly implicit schemes, see the recent article by Sato et al. [23].

A linearly implicit method will require the solution of a linear system at each time step, and will most likely constitute the computational bottleneck in a simulation as a whole. As will be shown here, using IEQ and SAV approaches, it is possible to arrive at energy-stable (indeed exactly lossless in machine arithmetic) numerical designs that are fully explicit. The explicit character of the update derives from the availability of a closed form inverse for the linearly implicit system that arises, and furthermore the ability to solve an $N \times N$ linear system in $O(N)$ operations, through the Sherman-Morrison inversion theorem [24]. This property is fully general, and not dependent on the particular form of the Hamiltonian, except through the additional non-negativity requirement on the potential energy. As a result, computer execution time for exactly energy conserving methods for Hamiltonian systems is very nearly on par with that of the simplest explicit numerical schemes, such as, e.g., Störmer-Verlet integration. This design addresses the two points a) and b) above with reference to the scheme presented in [11].

In Section 2, a restricted class of Hamiltonian systems is introduced. Under the constraint that the potential energy is non-negative, quadratisation techniques can be used to arrive at a simplified system of three equations in momentum, position, and a single scalar auxiliary variable derived from the potential energy. A further generalised form is also discussed, where a quadratic term is extracted or split from the expression for the potential energy before quadratisation, leading to a distinct starting point for numerical designs. Time stepping methods are introduced in Section 3, including first the rudimentary Störmer-Verlet method, the energy-conserving method presented in [11], and then an exact energy-conserving and unconditionally stable method obtained using IES/SAV approaches, and allowing an explicit update. A variation of this last scheme to the case of a split potential energy is also exactly energy-conserving, and leads to a distinct conditionally-stable method. In Section 4, three examples are presented: the Fermi-Pasta-Ulam ODE system, and two PDE systems then reduced, by semi-discretisation, to Hamiltonian ODE systems: the nonlinear vibration of a string, and the high-amplitude vibration of a thin plate. Various numerical results are presented, illustrating convergence rates, exact numerical energy conservation to machine precision, and relative computation times. Some concluding remarks appear in Section 5.

Preliminary results have been presented recently at the 2022 European Nonlinear Dynamics Conference [25].

2. Hamiltonian systems

Consider a system of particles, with N generalised coordinates $\mathbf{q} = [q_1, \ldots, q_N]^\intercal$ and momenta $\mathbf{p} = [p_1, \ldots, p_N]^\intercal$. Here, $^\intercal$ indicates a transposition operation, so $\mathbf{p}$ and $\mathbf{q}$ are $N \times 1$ column vectors. Both are functions of time $t \geq 0$, so $\mathbf{p} = \mathbf{p}(t)$ and $\mathbf{q} = \mathbf{q}(t)$. Suppose also that an associated Hamiltonian $H(\mathbf{p}, \mathbf{q})$ is defined by

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

for some constant symmetric positive definite $N \times N$ matrix $\mathbf{M}$ referred to as the mass matrix. $V(\mathbf{q})$ is the potential energy for the system of particles. H and V are scalar functions of time t, through their dependence on $\mathbf{p}$ and $\mathbf{q}$.

The system dynamics follow from Hamilton's equations:

$$\dot{\mathbf{q}} = \nabla_{\mathbf{p}} H \qquad \dot{\mathbf{p}} = -\nabla_{\mathbf{q}} H\,, \tag{2}$$

where dots indicate differentiation with respect to time t, and $\nabla_{\mathbf{p}}$ and $\nabla_{\mathbf{q}}$ are gradients with respect to $\mathbf{p}$ and $\mathbf{q}$, respectively. For the particular form of Hamiltonian given in (1), Hamilton's equations become

$$\dot{\mathbf{q}} = \mathbf{M}^{-1}\mathbf{p} \qquad \dot{\mathbf{p}} = -\nabla_{\mathbf{q}} V\,. \tag{3}$$

Through time differentiation of the first of (3), and substitution of the second, a second order system of ordinary differential equations in $\mathbf{q}$ results:

$$\mathbf{M}\ddot{\mathbf{q}} = -\nabla_{\mathbf{q}} V\,. \tag{4}$$

Such a second order form serves as the starting point for many numerical methods, including the classic Störmer-Verlet method. See Section 3.1. In the first order representation (3), initial conditions are required for $\mathbf{q}$ and $\mathbf{p}$, so $\mathbf{q}(0) = \mathbf{q}^{(0)}$ and $\mathbf{p}(0) = \mathbf{p}^{(0)}$, for given N vectors $\mathbf{q}^{(0)}$ and $\mathbf{p}^{(0)}$. For the second order form (4), one may set $\dot{\mathbf{q}}(0) = \mathbf{M}^{-1}\mathbf{p}^{(0)}$.

The defining feature of a Hamiltonian system is that, given initial conditions $\mathbf{p}^{(0)}$ and $\mathbf{q}^{(0)}$, the energy $H(t)$ is constant for all $t \geq 0$:

$$\dot{H} = 0 \quad \rightarrow \qquad H(t) = H(0) = \text{constant} \quad \forall t \geq 0\,, \tag{5}$$

where $H(0) = H(\mathbf{p}^{(0)}, \mathbf{q}^{(0)})$, evaluated from (1) using initial conditions $\mathbf{q}^{(0)}$ and $\mathbf{p}^{(0)}$.

2.1. Comments

The system and Hamiltonian in (1) are not the most general possible. A few remarks are offered here:

- The mass matrix $\mathbf{M}$ is assumed constant—a common assumption [5]. In many cases of interest, it is furthermore diagonal, but we will consider the more general form here, and indicate cases in which a diagonal form of $\mathbf{M}$ will have an impact on computational performance.
- The Hamiltonian in (1) is separable, so that $H(\mathbf{p},\mathbf{q}) = T(\mathbf{p}) + V(\mathbf{q})$, for kinetic energy T and potential energy V. In some cases of interest [8], the mass matrix $\mathbf{M}$ may be of the form $\mathbf{M} = \mathbf{M}(\mathbf{q})$, disturbing this splitting, but these cases will not be considered here.
- Following from the points above, the Hamiltonian is quadratic in $\mathbf{p}$, meaning that the associated dynamical system is linear in $\mathbf{p}$.
- The individual displacements q_i and momenta p_i, $i = 1, \ldots, N$, are assumed scalar here, and represent either displacements/momenta in a signal coordinate direction, or individual components of more general vector displacements/momenta. Both cases will be seen in the numerical examples in Section 4.
- Loss is easily introduced into system (3), rendering the system no longer Hamiltonian but dissipative. For example, considering linear loss, one may write:

$$\dot{\mathbf{q}} = \mathbf{M}^{-1}\mathbf{p} \qquad \dot{\mathbf{p}} = -\nabla_{\mathbf{q}} V - \mathbf{R}\mathbf{p}\,, \tag{6}$$

  for a positive semi-definite $N \times N$ matrix $\mathbf{R}$. In this case, the conservation law (5) may be generalized as

$$\dot{H} = -\mathbf{p}^{\mathsf{T}}\mathbf{R}\mathbf{p} \leq 0 \quad \rightarrow \qquad H(t) \leq H(0) \quad \forall t \geq 0\,, \tag{7}$$

  where H is as defined in (1).

2.2. Non-negativity of the potential energy

As a further constraint, assume that

$$V(\mathbf{q}) \geq 0 \qquad \forall \mathbf{q} \in \mathbb{R}^N\,. \tag{8}$$

This constraint is a natural one in many applications, but not all—for example, the gravitational potential used in the calculations of planetary orbits is not of this form [26]. This further implies, from (1), that

$$H \geq 0 \qquad \forall \mathbf{p}, \mathbf{q} \in \mathbb{R}^N\,. \tag{9}$$

Furthermore, as $\mathbf{M} > 0$, from (5), one has the following bound on $\mathbf{p}(t)$ in terms of the initial conditions:

$$\|\mathbf{p}(t)\| \leq \sqrt{2\lambda_{\max}(\mathbf{M})H(0)} \qquad \forall t \geq 0\,, \tag{10}$$

where $\lambda_{\max}(\mathbf{M})$ is the maximal eigenvalue of $\mathbf{M}$, and $\|\cdot\|$ indicates a Euclidean norm. If $V(\mathbf{q})$ is radially unbounded [27], so that $V(\mathbf{q}) \to +\infty$ as $\|\mathbf{q}\| \to +\infty$, then a further bound follows for $\|\mathbf{q}\|$.

The non-negativity property of V is essential to invariant energy quadratisation methods—indeed, it can be generalized to the case of V bounded from below, so that $V(\mathbf{q}) \geq c$, for any constant c [20], but the non-negativity condition (8) above will be used here for simplicity.

2.3. Potential energy quadratisation

Under the non-negativity condition on V, from (8), one may define

$$V = \tfrac{1}{2}\psi^2\,. \tag{11}$$

Hamilton's equations become

$$\dot{\mathbf{q}} = \mathbf{M}^{-1}\mathbf{p} \qquad \dot{\mathbf{p}} = -\psi\nabla_{\mathbf{q}}\psi\,. \tag{12}$$

Furthermore, using the chain rule,

$$\dot{\psi} = (\nabla_{\mathbf{q}}\psi)^{\mathsf{T}}\dot{\mathbf{q}}\,. \tag{13}$$

Finally, introducing

$$\mathbf{g} \triangleq \nabla_{\mathbf{q}}\psi = \frac{1}{\sqrt{2V}}\nabla_{\mathbf{q}}V\,, \tag{14}$$

a system of three equations results:

$$\dot{\mathbf{q}} = \mathbf{M}^{-1}\mathbf{p} \qquad \dot{\mathbf{p}} = -\psi\mathbf{g} \qquad \dot{\psi} = \mathbf{g}^{\mathsf{T}}\dot{\mathbf{q}}\,. \tag{15}$$

An auxiliary initial condition $\psi(0)$ may be set as $\psi(0) = \sqrt{2V(\mathbf{q}^{(0)})}$ in terms of the given initial condition $\mathbf{q}(0) = \mathbf{q}^{(0)}$.

The key feature of system (15) is that, if $\mathbf{g}$ is assumed known at at any given time instant, the resulting equations are linear in $\mathbf{p}$, $\mathbf{q}$ and ψ. Though in the present case, $\mathbf{g}$ does indeed have a dependence on $\mathbf{q}$, in the numerical setting, $\mathbf{g}$ can be evaluated directly using previously computed values of the solution, and thus the update becomes a matter of solving a linear system. This is the essence of the "linearly implicit" property of schemes arising from energy quadratisation. As will be shown subsequently here, under the appropriate numerical design, the linear system is of a particularly simple form with a known easily-computed inverse leading, effectively, to an explicit update. Such a property is independent of the particular form of the potential energy $V(\mathbf{q})$, provided the non-negativity constraint (8) is satisfied.

2.4. Potential energy splitting

In some cases, the expression for the potential energy naturally takes the form

$$V(\mathbf{q}) = \frac{1}{2}\mathbf{q}^{\mathsf{T}}\mathbf{K}\mathbf{q} + V'(\mathbf{q})\,, \tag{16}$$

where $\mathbf{K}$ is a symmetric positive semi-definite $N \times N$ matrix, and where $V' \geq 0$. The first term could represent the stored energy of the system due to linear mechanisms, and V' that due to additional nonlinear effects. Now, using

$$V' = \tfrac{1}{2}\psi^2\,, \tag{17}$$

one arrives at a system of equations generalizing (15):

$$\dot{\mathbf{q}} = \mathbf{M}^{-1}\mathbf{p} \qquad \dot{\mathbf{p}} = -\mathbf{K}\mathbf{q} - \psi\mathbf{g} \qquad \dot{\psi} = \mathbf{g}^{\mathsf{T}}\dot{\mathbf{q}}\,. \tag{18}$$

Though equivalent to (15), in a numerical setting, such a splitting allows for a larger family of numerical designs, treating the linear and nonlinear parts of the problem separately. See Section 3.4.

2.5. Potential energy gauge and regularisation

A well-known technique in IEQ/SAV approaches is the introduction of a constant shift of the global energy (see e.g. [20,28])—Hamilton's equations (3) are unchanged (i.e. gauge invariant), but some numerical schemes, such as IEQ/SAV-based approaches are affected [29]. This amounts to the replacement

$$V \rightarrow V + \epsilon \tag{19}$$

for a suitably chosen shift constant $\epsilon \geq 0$. Under the non-negativity constraint (8), $V + \epsilon$ is thus bounded away from zero, regularizing the calculation of $\mathbf{g}$ as defined in (14). The regularization approach applies equally to the case of a split potential, under the replacement $V' \rightarrow V' + \epsilon$.

3. Numerical methods

In this section, we assume time discretization using a constant time step k, such that solutions are computed at times $t^n = nk$, $n = 0, 1, \ldots$. A discrete-time vector $\mathbf{u}^n$ represents an approximation to a continuous time vector $\mathbf{u}(t)$ at times $t = t^n$.

3.1. Störmer-Verlet integration

The most basic approach to the integration of the Hamiltonian system as defined in Section 2 is through direct approximation of the second order form in (4), using centred differences, as:

$$\mathbf{q}^{n+1} = 2\mathbf{q}^n - \mathbf{q}^{n-1} - k^2 \mathbf{M}^{-1} \nabla_{\mathbf{q}} V|_{\mathbf{q}=\mathbf{q}^n} . \tag{20}$$

This discretisation is referred to as Störmer-Verlet, and was known to Newton—see [30] and the references therein. It is a fully explicit two-step scheme—in order to advance the solution to time step $n+1$, a direct evaluation of the gradient of the potential energy V at time step n is required. The values of the state $\mathbf{q}^n$ at $n = 0$ and $n = 1$ may be initialised, using the initial conditions $\mathbf{q}^{(0)}$ and $\mathbf{p}^{(0)}$, as

$$\mathbf{q}^0 = \mathbf{q}^{(0)} \qquad \mathbf{q}^1 = \mathbf{q}^{(0)} + k\mathbf{M}^{-1}\mathbf{p}^{(0)} - \frac{k^2}{2}\mathbf{M}^{-1}\nabla_{\mathbf{q}} V|_{\mathbf{q}=\mathbf{q}^{(0)}} \tag{21}$$

to second order in k, using a Taylor series approximation.

Störmer-Verlet is second-order accurate in the time step k, time-reversible, and symplectic, but not energy-conserving except in an approximate sense [30]. It can exhibit numerical instability depending on both the choice of time step and the size of the initial conditions. Such instabilities will be mentioned in Section 4.

3.2. Explicit approximately energy-conserving method

Consider the following scheme approximating (3), proposed by Marazzato et al. ([11], Eq. (6)):

$$\mathbf{q}^{n+1} = \mathbf{q}^n + k\mathbf{M}^{-1}\mathbf{p}^{n+\frac{1}{2}} \qquad \mathbf{p}^{n+\frac{3}{2}} = \mathbf{p}^{n-\frac{1}{2}} - 2\int_{nk}^{(n+1)k} \nabla_{\mathbf{q}} V(\tilde{\mathbf{q}}^n(t))dt . \tag{22}$$

This is a time interleaved scheme, with $\mathbf{p}^{n+\frac{1}{2}}$ and $\mathbf{q}^n$ defined at alternating multiples of $k/2$. The first of (22) is a standard interleaved approximation to the first of (3). The second of (22) relies on a continuous integration of $\nabla_{\mathbf{q}} V$ over the time interval $t \in [nk, (n+1)k]$, and over the known free-flight trajectory $\tilde{\mathbf{q}}^n(t)$ defined by

$$\tilde{\mathbf{q}}^n(t) = \mathbf{q}^n + (t - nk)\mathbf{M}^{-1}\mathbf{p}^{n+\frac{1}{2}} . \tag{23}$$

The scheme (22) may be rewritten as a three-step method in $\mathbf{q}^n$ alone as

$$\mathbf{q}^{n+2} = \mathbf{q}^{n+1} + \mathbf{q}^n - \mathbf{q}^{n-1} - 2k\mathbf{M}^{-1}\int_{nk}^{(n+1)k} \nabla_{\mathbf{q}} V|_{\mathbf{q}}(\bar{q}^n(t))dt \tag{24}$$

and may be initialised, to second order in k, as

$$\mathbf{q}^0 = \mathbf{q}^{(0)} \qquad \mathbf{q}^1 = \mathbf{q}^{(0)} + k\mathbf{M}^{-1}\mathbf{p}^{(0)} - \frac{k^2}{2}\mathbf{M}^{-1}\nabla V|_{\mathbf{q}=\mathbf{q}^{(0)}} \qquad \mathbf{q}^2 = \mathbf{q}^{(0)} + 2k\mathbf{M}^{-1}\mathbf{p}^{(0)} - 2k^2\mathbf{M}^{-1}\nabla V|_{\mathbf{q}=\mathbf{q}^{(0)}} \tag{25}$$

Other conservative schemes proposed also rely on such a continuous integration [16]. Except for particular functional forms of the potential energy $V(\mathbf{q})$, this integral is not available in closed form and must be approximated. However, once the integral is evaluated, the scheme (22) is fully explicit.

The scheme (22) has an associated numerical energy that is exactly conserved [11]:

$$\mathbf{H}^n = \tfrac{1}{2}\mathbf{p}^{n+\frac{1}{2}}\mathbf{M}^{-1}\mathbf{p}^{n-\frac{1}{2}} + V(\mathbf{q}^n) = \text{constant}\,. \tag{26}$$

There are two important points to mention here:

- Scheme (22) is exactly conservative, but depends on a continuous integration over a free-flight trajectory as given in (23) in order to achieve this property. Thus the exact conservation property is approached in the limit of increasing accuracy in the approximation of the continuous integration. It is also important to point out here that a fine-grained approximation of the continuous integration will require multiple evaluations of the gradient $\nabla_{\mathbf{q}} V$ which, depending on the functional form of V, represents an additional cost that grows with the desired accuracy of the approximation.
- When $V \geq 0$, the numerical energy (26) does not inherit the non-negativity property of the model system, and thus cannot be used in order to bound solution growth, as pointed out in Remark 8 of [11].

Under the additional non-negativity constraint (8) on V, it is possible to demonstrate a method that conserves a pseudo-energy that does not depend on a continuous integration, and that furthermore inherits the non-negativity property of the Hamiltonian of the model system, allowing for useful bounds on solution size. Furthermore, through the exploitation of matrix structure, it will be shown that such a method is fully explicit. Additionally, only one function evaluation $\nabla_{\mathbf{q}} V$ is required per time step.

3.3. Explicit exactly energy-conserving method

Returning now to the form (15) of Hamilton's equations obtained under quadratisation of the potential energy, and the introduction of the new variable ψ as in (11), consider the following scheme, written in terms of the interleaved time series $\mathbf{p}^{n+\frac{1}{2}}$ and $\mathbf{q}^n$, and the scalar time series $\psi^{n+\frac{1}{2}}$:

$$\mathbf{q}^{n+1} = \mathbf{q}^n + k\mathbf{M}^{-1}\mathbf{p}^{n+\frac{1}{2}} \tag{27a}$$

$$\mathbf{p}^{n+\frac{1}{2}} = \mathbf{p}^{n-\frac{1}{2}} - \tfrac{k}{2}\mathbf{g}^n\left(\psi^{n+\frac{1}{2}} + \psi^{n-\frac{1}{2}}\right) \tag{27b}$$

$$\psi^{n+\frac{1}{2}} = \psi^{n-\frac{1}{2}} + \tfrac{1}{2}\left(\mathbf{g}^n\right)^{\mathsf{T}}\left(\mathbf{q}^{n+1} - \mathbf{q}^{n-1}\right)\,. \tag{27c}$$

Here, $\mathbf{g}^n$ is defined as

$$\mathbf{g}^n = \nabla_{\mathbf{q}}\psi|_{\mathbf{q}=\mathbf{q}^n} = \frac{1}{\sqrt{2V(\mathbf{q}^n)}}\nabla_{\mathbf{q}} V|_{\mathbf{q}=\mathbf{q}^n}\,. \tag{28}$$

All difference approximations are centred, and the scheme as a whole is thus reversible and second-order accurate.

The system (27) may be manipulated into an explicit update form in the following way. Beginning from (27a), one may write:

$$\begin{aligned}\mathbf{q}^{n+1} &\overset{(27a)}{=} 2\mathbf{q}^n - \mathbf{q}^{n-1} + k\mathbf{M}^{-1}\left(\mathbf{p}^{n+\frac{1}{2}} - \mathbf{p}^{n-\frac{1}{2}}\right)\\ &\overset{(27b)}{=} 2\mathbf{q}^n - \mathbf{q}^{n-1} - \tfrac{k^2}{2}\mathbf{M}^{-1}\mathbf{g}^n\left(\psi^{n+\frac{1}{2}} + \psi^{n-\frac{1}{2}}\right)\\ &\overset{(27c)}{=} 2\mathbf{q}^n - \mathbf{q}^{n-1} - k^2\mathbf{M}^{-1}\mathbf{g}^n\psi^{n-\frac{1}{2}} - \tfrac{k^2}{4}\mathbf{M}^{-1}\mathbf{g}^n(\mathbf{g}^n)^{\mathsf{T}}\left(\mathbf{q}^{n+1} - \mathbf{q}^{n-1}\right)\,.\end{aligned} \tag{29}$$

Finally, the update has the form

$$\mathbf{A}^n\mathbf{q}^{n+1} = \mathbf{b}^n\,, \tag{30}$$

where

$$\mathbf{A}^n = \mathbf{I} + \boldsymbol{\alpha}^n(\boldsymbol{\beta}^n)^{\mathsf{T}} \qquad \mathbf{b}^n = 2\mathbf{q}^n - 2k\boldsymbol{\alpha}^n\psi^{n-\frac{1}{2}} - \left(\mathbf{I} - \boldsymbol{\alpha}^n(\boldsymbol{\beta}^n)^{\mathsf{T}}\right)\mathbf{q}^{n-1}\,. \tag{31}$$

Here, $\mathbf{I}$ is the $N \times N$ identity matrix, and the vectors $\boldsymbol{\alpha}^n$ and $\boldsymbol{\beta}^n$ are defined in terms of $\mathbf{g}^n$ by $\boldsymbol{\alpha}^n = \frac{k}{2}\mathbf{M}^{-1}\mathbf{g}^n$ and $\boldsymbol{\beta}^n = \frac{k}{2}\mathbf{g}^n$. Thus, given $\mathbf{q}^n$, $\mathbf{q}^{n-1}$ and $\psi^{n-\frac{1}{2}}$, both $\mathbf{A}^n$ and $\mathbf{b}^n$ may be explicitly constructed. Notice that the matrix $\mathbf{A}^n$ is positive definite by construction, due to the positive definiteness of $\mathbf{M}$, and thus the update (30) always has a unique solution. Once the

update in (30) has been performed, $\psi^{n+\frac{1}{2}}$ may be computed from (27c) explicitly, using $\mathbf{q}^{n-1}$, $\mathbf{q}^{n+1}$ and $\psi^{n-\frac{1}{2}}$. Scheme (30) requires initial values for $\mathbf{q}^0$ and $\mathbf{q}^1$, which may be set in the same way as for Störmer-Verlet, as in (21), and also $\psi^{\frac{1}{2}}$ which may be set to second order in k in terms of the initial displacement $\mathbf{q}^{(0)}$ and momentum $\mathbf{p}^{(0)}$ as

$$\psi^{\frac{1}{2}} = \sqrt{2V(\mathbf{q}^{(0)})} + \frac{k}{2\sqrt{2V(\mathbf{q}^{(0)})}} \left(\nabla_{\mathbf{q}} V|_{\mathbf{q}=\mathbf{q}^{(0)}}\right)^{\mathsf{T}} \mathbf{M}^{-1}\mathbf{p}^{(0)} \\ + \frac{k^2}{8}\left(\mathbf{p}^{(0)}\right)^{\mathsf{T}} \mathbf{M}^{-1}\left(\nabla_{\mathbf{q}} \frac{1}{\sqrt{2V}} \nabla_{\mathbf{q}} V\right)|_{\mathbf{q}=\mathbf{q}^{(0)}} \mathbf{M}^{-1}\mathbf{p}^{(0)} - \left(\frac{k^2}{8\sqrt{2V}} \left(\nabla_{\mathbf{q}} V\right)^{\mathsf{T}} \mathbf{M}^{-1} \nabla_{\mathbf{q}} V\right)|_{\mathbf{q}=\mathbf{q}^{(0)}}. \tag{32}$$

Though the solution of (30) apparently requires an $N \times N$ linear system solution, in fact, the inverse is available in closed form, and allows for an explicit solution in $O(N)$ operations. $\mathbf{A}^n$ is a rank-1 perturbation of the identity, and thus Sherman Morrison inversion [24] yields:

$$\left(\mathbf{A}^n\right)^{-1} = \mathbf{I} - \frac{\boldsymbol{\alpha}^n (\boldsymbol{\beta}^n)^{\mathsf{T}}}{1 + (\boldsymbol{\beta}^n)^{\mathsf{T}} \boldsymbol{\alpha}^n}. \tag{33}$$

Computational cost is thus on par with the other methods presented here. Notice in particular, though, that the solution of a linear system involving $\mathbf{M}^{-1}$ is required in order to compute $\boldsymbol{\alpha}^n$, performed once per time step. This requirement of a linear system solution is common to all the methods in this section; the computational cost of performing this operation, which may indeed be the bottleneck, is not considered here. In many cases, though, $\mathbf{M}$ is diagonal, or even a simple multiple of the identity, and thus its inversion is trivial.

Exact conservation of a numerical energy follows directly from scheme (27). Left-multiplying (27b) by $\frac{1}{2}\left(\mathbf{p}^{n+\frac{1}{2}} + \mathbf{p}^{n-\frac{1}{2}}\right)^{\mathsf{T}} \mathbf{M}^{-1}$ gives:

$$\begin{aligned} &\tfrac{1}{2}\left(\mathbf{p}^{n+\frac{1}{2}} + \mathbf{p}^{n-\frac{1}{2}}\right)^{\mathsf{T}} \mathbf{M}^{-1}\left(\mathbf{p}^{n+\frac{1}{2}} - \mathbf{p}^{n-\frac{1}{2}}\right) + \tfrac{k}{4}\left(\mathbf{p}^{n+\frac{1}{2}} + \mathbf{p}^{n-\frac{1}{2}}\right)^{\mathsf{T}} \mathbf{M}^{-1}\mathbf{g}^n\left(\psi^{n+\frac{1}{2}} + \psi^{n-\frac{1}{2}}\right) \overset{(27b)}{=} 0 \\ &\tfrac{1}{2}\left(\mathbf{p}^{n+\frac{1}{2}} + \mathbf{p}^{n-\frac{1}{2}}\right)^{\mathsf{T}} \mathbf{M}^{-1}\left(\mathbf{p}^{n+\frac{1}{2}} - \mathbf{p}^{n-\frac{1}{2}}\right) + \tfrac{1}{4}\left(\mathbf{q}^{n+1} - \mathbf{q}^{n-1}\right)^{\mathsf{T}} \mathbf{g}^n\left(\psi^{n+\frac{1}{2}} + \psi^{n-\frac{1}{2}}\right) \overset{(27a)}{=} 0 \\ &\tfrac{1}{2}\left(\mathbf{p}^{n+\frac{1}{2}} + \mathbf{p}^{n-\frac{1}{2}}\right)^{\mathsf{T}} \mathbf{M}^{-1}\left(\mathbf{p}^{n+\frac{1}{2}} - \mathbf{p}^{n-\frac{1}{2}}\right) + \tfrac{1}{2}\left(\psi^{n+\frac{1}{2}} - \psi^{n-\frac{1}{2}}\right)\left(\psi^{n+\frac{1}{2}} + \psi^{n-\frac{1}{2}}\right) \overset{(27c)}{=} 0. \end{aligned} \tag{34}$$

This may be identified as

$$H^{n+\frac{1}{2}} = H^{n-\frac{1}{2}} = \text{constant}, \tag{35}$$

where the numerical energy $H^{n+\frac{1}{2}}$ is

$$H^{n+\frac{1}{2}} = \tfrac{1}{2}\left(\mathbf{p}^{n+\frac{1}{2}}\right)^{\mathsf{T}} \mathbf{M}^{-1}\mathbf{p}^{n+\frac{1}{2}} + \tfrac{1}{2}\left(\psi^{n+\frac{1}{2}}\right)^2 \geq 0. \tag{36}$$

It is worth comparing this expression, for which non-negativity is ensured, with the expression (26) for the scheme in Section 3.2. In this case, the kinetic energy term is exact, whereas the potential energy term is approximate but non-negative. In (26), the potential energy term can be recovered exactly in the limit of increasing quadrature accuracy, but the kinetic energy term is approximate and also unsigned. In the scheme (27) presented here, the momentum $\mathbf{p}^{n+\frac{1}{2}}$ may be bounded in terms of the energy H, for all time steps n, as

$$\|\mathbf{p}^{n+\frac{1}{2}}\| \leq \sqrt{2\lambda_{\max}(\mathbf{M})H}, \tag{37}$$

which is identical to the bound (10) for the model system. The scheme (27) is thus unconditionally numerically stable. A known issue here is that the numerical energy may not correspond exactly to the energy for the physical system. This issue has been approached recently by various authors [31,32]; but here, we are interested in the non-negativity property of the numerical conserved energy, and its utility as a means of ensuring numerically stable behaviour.

3.4. Split potential form

Consider now the split form of the potential energy V, as described in Section 2.4, where a positive semi-definite quadratic form has been separated from V to leave a residual energy contribution $V' \geq 0$, from which an auxiliary variable

ψ may be defined as in (17). Consider the following scheme, now modified with respect to (27) through the addition of a linear term in (38b) below:

$$\mathbf{q}^{n+1} = \mathbf{q}^n + k\mathbf{M}^{-1}\mathbf{p}^{n+\frac{1}{2}} \tag{38a}$$

$$\mathbf{p}^{n+\frac{1}{2}} = \mathbf{p}^{n-\frac{1}{2}} - k\mathbf{K}\mathbf{q}^n - \tfrac{k}{2}\mathbf{g}^n\left(\psi^{n+\frac{1}{2}} + \psi^{n-\frac{1}{2}}\right) \tag{38b}$$

$$\psi^{n+\frac{1}{2}} = \psi^{n-\frac{1}{2}} + \tfrac{1}{2}\left(\mathbf{g}^n\right)^{\mathsf{T}}\left(\mathbf{q}^{n+1} - \mathbf{q}^{n-1}\right), \tag{38c}$$

where now,

$$\mathbf{g}^n = \nabla_{\mathbf{q}}\psi|_{\mathbf{q}=\mathbf{q}^n} = \frac{1}{\sqrt{2V'(\mathbf{q}^n)}}\nabla_{\mathbf{q}}V'|_{\mathbf{q}=\mathbf{q}^n}. \tag{39}$$

An explicit update follows as in the case of the non-split form in (30), with $\mathbf{A}^n$ as given in (31), but with $\mathbf{b}^n$ now defined as

$$\mathbf{b}^n = \left(2\mathbf{I} - k^2\mathbf{M}^{-1}\mathbf{K}\right)\mathbf{q}^n - 2k\boldsymbol{\alpha}^n\psi^{n-\frac{1}{2}} - \left(\mathbf{I} - \boldsymbol{\alpha}^n(\boldsymbol{\beta}^n)^{\mathsf{T}}\right)\mathbf{q}^{n-1}. \tag{40}$$

An expression for a conserved numerical energy follows as before, now taking the form:

$$H^{n+\frac{1}{2}} = \tfrac{1}{2}\left(\mathbf{p}^{n+\frac{1}{2}}\right)^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{p}^{n+\frac{1}{2}} + \tfrac{1}{2}(\mathbf{q}^{n+1})^{\mathsf{T}}\mathbf{K}\mathbf{q}^n + \tfrac{1}{2}\left(\psi^{n+\frac{1}{2}}\right)^2 = \text{constant}. \tag{41}$$

Because the second term is of indefinite sign, the numerical energy is not necessarily non-negative. Because it is a quadratic form, however, it may be bounded, using (38a), as:

$$\tfrac{1}{2}(\mathbf{q}^{n+1})^{\mathsf{T}}\mathbf{K}\mathbf{q}^n \geq -\tfrac{1}{8}(\mathbf{q}^{n+1} - \mathbf{q}^n)^{\mathsf{T}}\mathbf{K}(\mathbf{q}^{n+1} - \mathbf{q}^n) = -\tfrac{k^2}{8}\left(\mathbf{p}^{n+\frac{1}{2}}\right)^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{K}\mathbf{M}^{-1}\mathbf{p}^{n+\frac{1}{2}}. \tag{42}$$

It then follows that

$$H^{n+\frac{1}{2}} \geq \tfrac{1}{2}\left(\mathbf{p}^{n+\frac{1}{2}}\right)^{\mathsf{T}}\left(\mathbf{M}^{-1} - \frac{k^2}{4}\mathbf{M}^{-1}\mathbf{K}\mathbf{M}^{-1}\right)\mathbf{p}^{n+\frac{1}{2}}. \tag{43}$$

Given that $\mathbf{M} > 0$, $\mathbf{M}^{-1}$ may be factored into unique upper and lower triangular factors as $\mathbf{M}^{-1} = \mathbf{M}^{-\frac{T}{2}}\mathbf{M}^{-\frac{1}{2}}$, a condition for non-negativity of H is

$$k \leq \frac{2}{\lambda_{\max}\left(\mathbf{M}^{-\frac{1}{2}}\mathbf{K}\mathbf{M}^{-\frac{T}{2}}\right)}. \tag{44}$$

The scheme is now conditionally stable, with the stability condition (44) corresponding to that for a linear system when $V' = 0$.

3.5. Remarks

- **Linear Conditions:** One of the interesting features of the non-split form (27) is that, even if the potential V is quadratic in $\mathbf{q}$, implying a linear system, the scheme is not linear. If, however, the split form (38) is used, then under linear conditions, $V' = 0$, and scheme (38) reduces exactly to Störmer-Verlet (20). Note that under linear conditions, Störmer-Verlet does indeed possess an exactly conserved numerical energy, as given in (41) with $\psi^{n+\frac{1}{2}} = 0$.
- **Generalized Splitting:** The non-split form (27) and the split form (38) may be combined to yield a large family of conservative schemes in an obvious way. If

$$V(\mathbf{q}) = \tfrac{1}{2}\mathbf{q}^{\mathsf{T}}\mathbf{K}\mathbf{q} + V_{\text{nonlinear}}(\mathbf{q}), \tag{45}$$

where $V_{\text{nonlinear}} \geq 0$ and consists of higher order terms in $\mathbf{q}$, then any splitting of the form

$$V(\mathbf{q}) = \tfrac{1}{2}\mathbf{q}^{\mathsf{T}}\mathbf{K}_0\mathbf{q} + V'(\mathbf{q}) \qquad V'(\mathbf{q}) = V_{\text{nonlinear}} + \tfrac{1}{2}\mathbf{q}^{\mathsf{T}}\mathbf{K}'\mathbf{q}, \tag{46}$$

with $\mathbf{K} = \mathbf{K}_0 + \mathbf{K}'$, $\mathbf{K}_0 \geq 0$, $\mathbf{K}' \geq 0$ will yield a conservative form. The resulting numerical scheme, with $V' = \frac{1}{2}\psi^2$ will inherit conservation of energy, which will be non-negative under a condition analogous to (44), but depending on $\mathbf{K}_0$.

- **Generalized Update for $\mathbf{q}^n$:** Consider again the non-split scheme (27), which may be rewritten as

$$\mathbf{q}^{n+1} = \mathbf{q}^n + k\mathbf{M}^{-1}\mathbf{p}^{n+\frac{1}{2}} \tag{47a}$$

$$\mathbf{p}^{n+\frac{1}{2}} = \mathbf{p}^{n-\frac{1}{2}} - \tfrac{k}{2}\mathbf{g}^n\left(\psi^{n+\frac{1}{2}} + \psi^{n-\frac{1}{2}}\right) \tag{47b}$$

$$\psi^{n+\frac{1}{2}} = \psi^{n-\frac{1}{2}} + \tfrac{k}{2}\left(\mathbf{g}^n\right)^{\mathsf{T}}\mathbf{M}^{-1}\left(\mathbf{p}^{n+\frac{1}{2}} + \mathbf{p}^{n-\frac{1}{2}}\right). \tag{47c}$$

  In this form, exact energy conservation follows from (47b) and (47c) only—it is independent of the values of $\mathbf{g}^n$, which are derived solely from $\mathbf{q}^n$. Further opportunities for generalization are thus available—(47a) could be replaced by any consistent update for $\mathbf{q}^{n+1}$, and the exact numerical energy conservation property remains undisturbed.
- **Variable Time Steps:** Though the case of variable time steps will not be discussed here in detail (and was indeed investigated in [11]), an exact energy-conserving scheme follows immediately from the form given in (47) above. Consider time instants t^n, and $t^{n+1/2}$, for integer n, where $t^n < t^{n+1/2} < t^{n+1}$. From these, one may define two sequences of time steps: $k^n = t^{n+1/2} - t^{n-1/2}$ and $k^{n+1/2} = t^{n+1} - t^n$. Supposing that $\mathbf{q}^n$ is a time series defined for $t = t^n$, and similarly $\mathbf{p}^{n+\frac{1}{2}}$ and $\psi^{n+\frac{1}{2}}$ are defined for $t = t^{n+\frac{1}{2}}$, then a scheme follows as:

$$\mathbf{q}^{n+1} = \mathbf{q}^n + k^{n+\frac{1}{2}}\mathbf{M}^{-1}\mathbf{p}^{n+\frac{1}{2}} \tag{48a}$$

$$\mathbf{p}^{n+\frac{1}{2}} = \mathbf{p}^{n-\frac{1}{2}} - \tfrac{k^n}{2}\mathbf{g}^n\left(\psi^{n+\frac{1}{2}} + \psi^{n-\frac{1}{2}}\right) \tag{48b}$$

$$\psi^{n+\frac{1}{2}} = \psi^{n-\frac{1}{2}} + \tfrac{k^n}{2}\left(\mathbf{g}^n\right)^{\mathsf{T}}\mathbf{M}^{-1}\left(\mathbf{p}^{n+\frac{1}{2}} + \mathbf{p}^{n-\frac{1}{2}}\right). \tag{48c}$$

  It is direct to show that, regardless of the choices of t^n and $t^{n+\frac{1}{2}}$, the scheme (48) conserves the energy (36) exactly. Thus such a generalisation is unconditionally stable, under the same reasoning as in the case of constant time steps.
- **Regularisation:** Regularisation through a potential energy shift, as described in Section 2.5, impacts on the calculation of $\mathbf{g}$ in (28) and (39), through a replacement $V \to V + \epsilon$ or $V' \to V' + \epsilon$, respectively.
- **Loss:** Following from the final comment in Section 2.1, the scheme presented here extends easily to include loss, as per the extended system definition in (6). For the non-split form (27), (27b) can be adjusted as

$$\mathbf{p}^{n+\frac{1}{2}} = \mathbf{p}^{n-\frac{1}{2}} - \tfrac{k}{2}\mathbf{g}^n\left(\psi^{n+\frac{1}{2}} + \psi^{n-\frac{1}{2}}\right) - \tfrac{k}{2}\mathbf{M}\mathbf{R}\left(\mathbf{p}^{n+\frac{1}{2}} + \mathbf{p}^{n-\frac{1}{2}}\right) \tag{49}$$

  The numerical energy conservation law (41) now becomes, using the same expression for total energy $H^{n+\frac{1}{2}}$,

$$H^{n+\frac{1}{2}} - H^{n-\frac{1}{2}} = -\tfrac{k}{4}\left(\mathbf{p}^{n+\frac{1}{2}} + \mathbf{p}^{n-\frac{1}{2}}\right)^{\mathsf{T}}\mathbf{R}\left(\mathbf{p}^{n+\frac{1}{2}} + \mathbf{p}^{n-\frac{1}{2}}\right) \leq 0 \qquad \to \qquad H^{n+\frac{1}{2}} \leq H^{n-\frac{1}{2}}. \tag{50}$$

  From an implementation standpoint, the ability to make use of a fast inversion technique (Sherman-Morrison) is unaffected by the addition of loss.

4. Examples

In this section, various Hamiltonian systems are simulated using schemes (27) and (38), beginning with the Fermi-Pasta-Ulam ODE problem in Section 4.1, and then progressing to more complex ODE systems derived as semi-discretisations to PDE systems. These include the coupled transverse-longitudinal vibration of a string at high amplitudes, in Section 4.2, and the high amplitude vibration of a thin plate in Section 4.3.

In all cases, a useful measure of the exact energy conservation property is the relative energy deviation error,

$$\text{Relative error} = \frac{H^{n+\frac{1}{2}} - H^{\frac{1}{2}}}{H^{\frac{1}{2}}}. \tag{51}$$

In double precision floating point arithmetic, it is normally on the order of machine precision, or approximately 10^{-16}. Depending on the state size of the system in question, however, larger deviations are possible.

4.1. Fermi-Pasta-Ulam problem

As a simple example, consider the classic system of a linear arrangement masses connected by linear and nonlinear springs, as proposed originally by Fermi, Pasta and Ulam [33], and later adapted as a test problem by various authors [34,11], the form of which is followed here.

Consider a system of $N = 2M$ masses in a linear arrangement, which longitudinal displacements q_i and momenta p_i, $i = 1, \ldots, 2M$. The system Hamiltonian is defined by

$$H = \frac{1}{2}\sum_{i=1}^{2M} p_i^2 + V \qquad \text{where} \qquad V = \frac{\omega^2}{4}\sum_{i=1}^{M}(q_{2i} - q_{2i-1})^2 + \sum_{i=0}^{M}(q_{2i+1} - q_{2i})^4 \,, \tag{52}$$

where, in the expression above, $q_0 = q_{2M+1} = 0$. Thus each mass is connected, in an alternating arrangement, to a linear spring, and a cubic nonlinear spring.

V is clearly non-negative here, and thus, after consolidation of the displacements q_i and p_i into vectors $\mathbf{p}$ and $\mathbf{q}$ of size $N \times 1$, the scheme as presented in (27) follows immediately, with $\mathbf{M} = \mathbf{I}_N$, the $N \times N$ identity matrix, and with V as given in (52) above. This scheme is explicit, exactly conservative, and unconditionally stable.

A split form (16) of the Hamiltonian also follows, using

$$\mathbf{K} = \frac{\omega^2}{2}\mathbf{I}_M \otimes \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad V' = \sum_{i=0}^{M}(q_{2i+1} - q_{2i})^4 \geq 0 \,, \tag{53}$$

where here, $\otimes$ indicates a Kronecker product. An exactly energy conserving method follows, as in (38), and is stable under the condition (44), which reduces in this case to

$$k \leq \frac{2}{\omega} \,. \tag{54}$$

4.1.1. Numerical results

We use the same settings as in [11], and choose $\omega = 50$, and $M = 3$. Simulations are run here with a time step of $k = 10^{-3}$ s; the reference solution is computed using Störmer-Verlet, with a time step of $k = 2^{-20} \approxeq 10^{-6}$ s. For initial conditions, we set $\mathbf{p}^{(0)} = \mathbf{0}$, and $\mathbf{q}^{(0)} = [0, 0, 0, \alpha, 0, 0]^\mathsf{T}$, for different values of α, with numerical initialisation for the various schemes given in the relevant sections. See Fig. 1, illustrating a comparison between time histories using the exactly conservative schemes (27) and (38) with Störmer-Verlet (20), and the Hamiltonian scheme (22), using simple midpoint quadrature rule (replicating the results in [11]). In all cases, the onset of errors is slightly faster for (27) and (22) than for the other schemes, indicating a higher error (see also Fig. 3).

For the schemes (27) and (38), numerical energy, as defined by (36) and (41) respectively, is conserved to machine precision. See Fig. 2, illustrating the relative deviation in energy, as defined in (51), for scheme (27) for the Fermi-Pasta-Ulam system, under the conditions as described above, and for $\alpha = 100$. It is easily seen that the relative error is on the order of machine roundoff error in double-precision floating point, or approximately 10^{-16}. Furthermore, it is possible to directly observe the quantisation of the relative energy to the lowest bits in the machine number representation.

As a basic test of convergence, consider the L^2 error defined, over a simulation duration $n = 0, \ldots, N_f$, by

$$\text{Error} = \sqrt{\sum_{n=0}^{N_f} k \|\mathbf{q}^n - \mathbf{q}_{\text{ref}}(t = nk)\|^2} \,, \tag{55}$$

where $\mathbf{q}_{\text{ref}}$ are computed using Störmer-Verlet with a time step of $k = 2^{-20} \approxeq 10^{-6}$ s. The error appears in Fig. 3, for the exact energy conserving method in (27), the split potential method in (38), Störmer-Verlet and the Hamiltonian scheme (22), over a range of time steps. In this case, the total simulation duration is 1 s, and three different values of the initial condition amplitude α are chosen: $\alpha = 10, 50, 100$. No regularisation was employed in this case (see Section 2.5). In general, the errors for Störmer-Verlet and the split scheme (38) track each other quite closely, with the non-split method (27), and scheme (22) performing somewhat worse, and increasingly so at higher amplitudes of the initial condition. Second order accuracy is easily observed in all cases. Stability for Störmer-Verlet and scheme (22) is dependent on the size of the initial condition; the range of time steps over which Störmer-Verlet and (22) are unstable is indicated as a grey region in the figure. Note that scheme (22) employs midpoint quadrature—it could well be that the accuracy and stability range improve using better approximations to the continuous integration.

4.2. Nonlinear string vibration

As a second example, consider the motion of a taut string. For sufficiently large displacements, the linear wave equation must be augmented by appropriate nonlinear terms to account for the amplitude-dependent physical phenomena observed

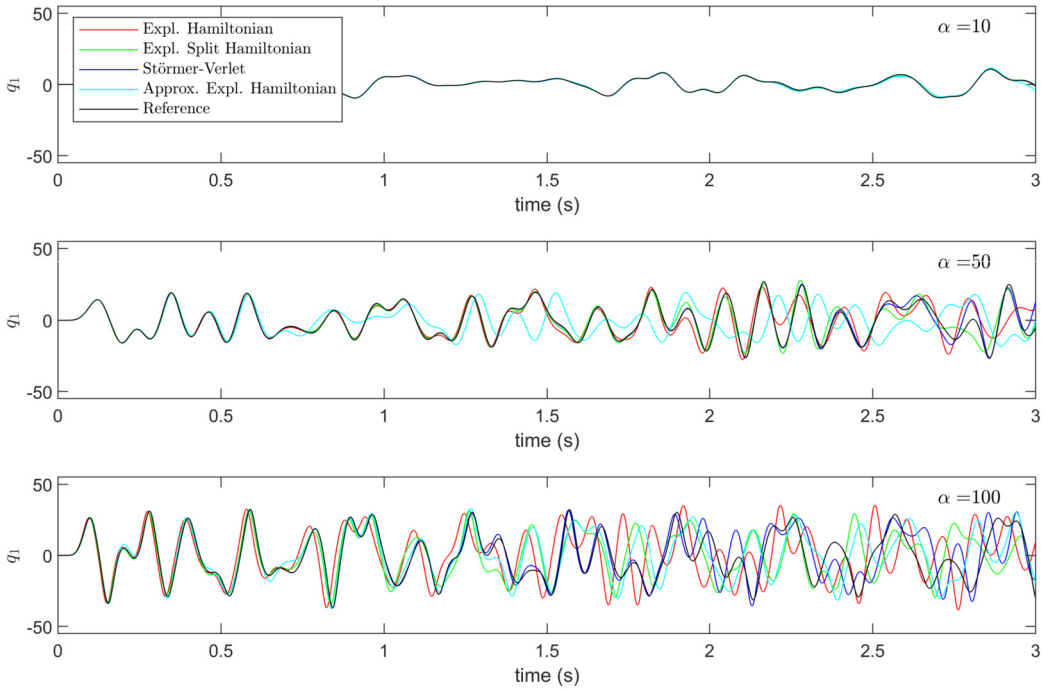

Fig. 1. Time histories of q_1, under the initial condition $q_4 = \alpha$, and under different values of α as indicated. Results for schemes (27), (38) (20), and (22) using midpoint quadrature are plotted against a reference solution. (For interpretation of the colours in the figures, the reader is referred to the web version of this article.)

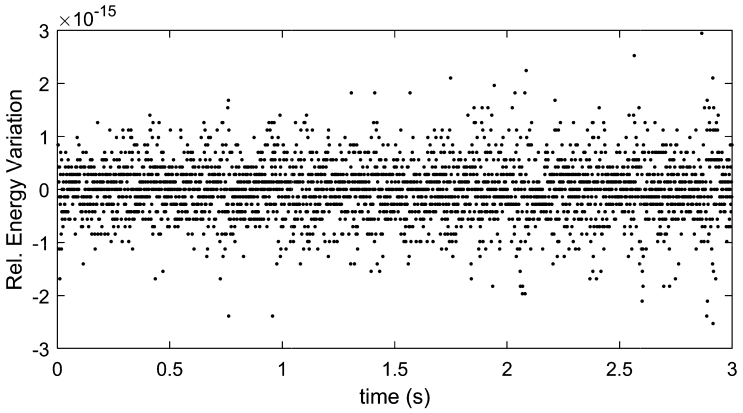

Fig. 2. Relative variation in numerical energy, as defined by (51), for scheme (27) for the Fermi-Pasta-Ulam system.

during motion. One may obtain a geometrically exact model by considering large strains, and applying Hooke's law [35]. This model is written compactly as

$$\rho A \partial_t^2 u = \partial_x \left(\frac{\partial \mathcal{V}}{\partial \zeta} \right), \quad \rho A \partial_t^2 v = \partial_x \left(\frac{\partial \mathcal{V}}{\partial \eta} \right). \tag{56}$$

Here, $u = u(x,t) : \mathcal{D} \times \mathbb{R}_0^+ \to \mathbb{R}$ represents the transverse displacement, for a spatial domain $\mathcal{D} = [0, L]$, where L is the length of the unstretched string. Similarly, $v = v(x,t)$ is the longitudinal displacement. In (56), ∂_t and ∂_x represent partial differentiation with respect to time t and spatial coordinate x. Furthermore, we define $\zeta = \partial_x u$, $\eta = \partial_x v$. The function $\mathcal{V} = \mathcal{V}(\zeta, \eta) : \mathbb{R}^2 \to \mathbb{R}_0^+$ is a non-negative potential density, which may be given in two equivalent forms as

$$\mathcal{V}(\zeta,\eta) = \begin{cases} \dfrac{EA}{2}\left(\zeta^2+\eta^2\right) - (EA - T_0)\left(\sqrt{(1+\eta)^2+\zeta^2} - 1 - \eta\right), & \text{(a)} \\ \dfrac{T_0}{2}\left(\zeta^2+\eta^2\right) + \dfrac{EA - T_0}{2}\left(\sqrt{(1+\eta)^2+\zeta^2} - 1\right)^2. & \text{(b)} \end{cases} \tag{57}$$

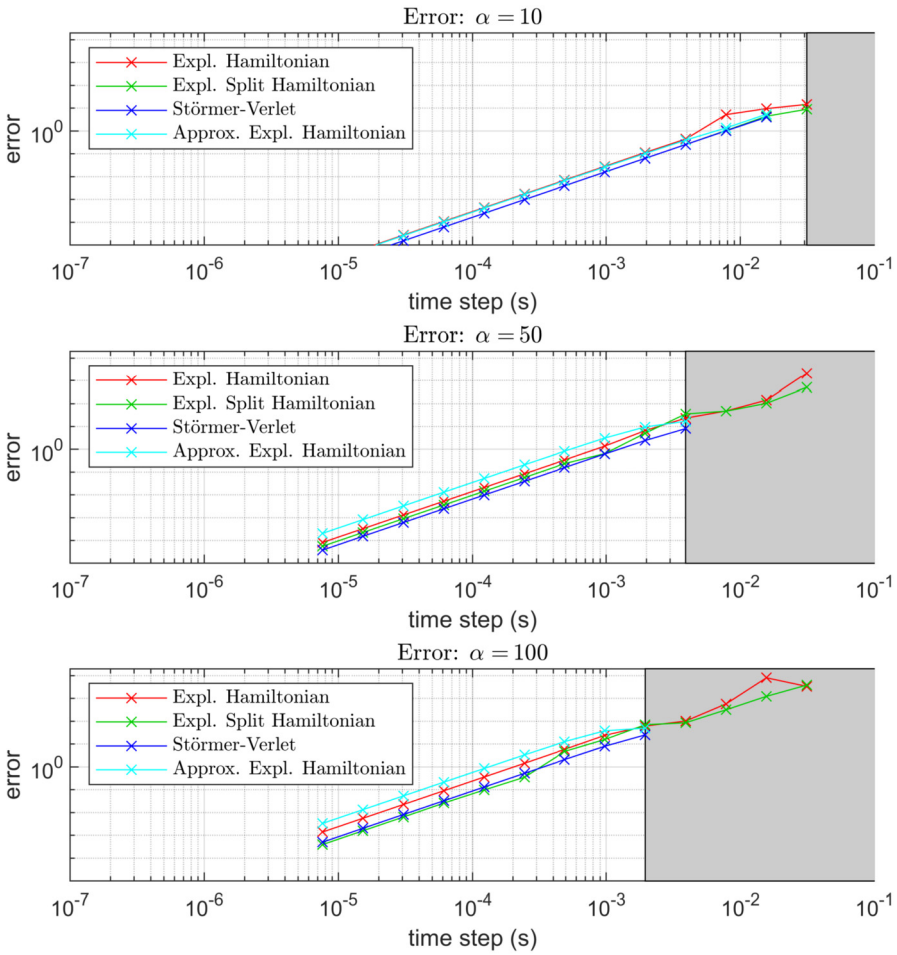

Fig. 3. Error, as defined in (55), as a function of time step k for the explicit Hamiltonian scheme (27), the split potential form (38), Störmer-Verlet (20), and the Hamiltonian scheme (22) using a midpoint quadrature approximation, for the Fermi-Pasta-Ulam problem, where $\mathbf{q}^{(0)} = [0, 0, 0, \alpha, 0, 0]$, and for different values of α as indicated. (For interpretation of the colours in the figures, the reader is referred to the web version of this article.)

Here, the various constants that appear are: ρ, the volume density in kg·m^{-3}; E, Young's modulus, in kg·s^{-2}·m^{-1}; A, the area of the string cross section in m^2; and T_0, the applied tension in kg·m·s^{-2}. It is easy to show that (57a) is the same as (57b), and that $\mathcal{V} \geq 0 \; \forall \; (\zeta, \eta)$. In (57b), one may easily split $\mathcal{V}$ into a quadratic part, plus a non-negative nonlinear term, provided that $EA > T_0$ (a condition that is generally satisfied, for instance by all strings of interest in musical acoustics): ultimately, this allows for numerical solution using a split potential form as described in Section 3.4. Both forms have been used in previous works: (57a) in e.g. [36,11]; (57b) in [37]. System (56) is Hamiltonian, with the total energy defined by

$$H = \int_{\mathcal{D}} \frac{\rho A}{2} \left((\partial_t u)^2 + (\partial_t v)^2 \right) + \mathcal{V} \, \mathrm{d}x. \tag{58}$$

Energy conservation holds under a suitable set of boundary conditions. Here, conditions of fixed type are considered, such that $u = v = 0$ at $x = \{0, L\}$.

4.2.1. Semi-discrete form

The domain $\mathcal{D}$ may be divided into segments of length h, the grid spacing. Let M be the total number of segments, yielding $M-1$ grid points, not including the end points, where the solution is fixed to zero. The continuous functions $u(x,t)$, $v(x,t)$ may then be approximated by grid functions $u_l(t)$, $v_l(t)$, at the grid point $x_l = lh$, $l = 1, ..., M-1$. Approximations to ∂_x may be given as difference operators, expressed here in terms of their action on a grid function $u_l(t)$:

$$D_+ u_l = \frac{1}{h} \left(u_{l+1} - u_l \right), \quad D_- u_l = \frac{1}{h} \left(u_l - u_{l-1} \right). \tag{59}$$

From these, one may also define the second difference operator as $D_2 = D_+ D_-$. Furthermore, let $\zeta_l = D_- u_l$, $\eta_l = D_- v_l$. A semi-discrete form for (56) is then obtained as

$$\rho A \ddot{u}_l = D_+ \left(\frac{\partial \mathcal{V}_l}{\partial \zeta_l} \right), \qquad \rho A \ddot{v}_l = D_+ \left(\frac{\partial \mathcal{V}_l}{\partial \eta_l} \right), \tag{60}$$

where $\mathcal{V}_l \triangleq \mathcal{V}(\zeta_l, \eta_l)$. It is convenient to write system (60) compactly using the consolidated state vector $\mathbf{q} = [\mathbf{u}^\mathsf{T}, \mathbf{v}^\mathsf{T}]^\mathsf{T}$. In this form, the difference operators are represented by matrices, such that D_- becomes the $M \times (M-1)$ matrix $\mathbf{D}_-$ given in terms of its action on e.g. $\mathbf{u}$ as

$$\mathbf{D}_- \mathbf{u} = \frac{1}{h}([\mathbf{u}^\mathsf{T}, 0] - [0, \mathbf{u}^\mathsf{T}]). \tag{61}$$

Then, define $\mathbf{D}_+ = -\mathbf{D}_-^\mathsf{T}$, and $\mathbf{D}_2 = \mathbf{D}_+ \mathbf{D}_-$. In vector form, (60) becomes

$$\rho A \ddot{\mathbf{q}} = \frac{1}{h} \begin{bmatrix} \mathbf{D}_+ \nabla_{\boldsymbol{\zeta}} V \\ \mathbf{D}_+ \nabla_{\boldsymbol{\eta}} V \end{bmatrix}, \tag{62}$$

where $\boldsymbol{\zeta} = \mathbf{D}_- \mathbf{u}$, $\boldsymbol{\eta} = \mathbf{D}_- \mathbf{v}$. This system conserves the semi-discrete energy of the form (1), with

$$\mathbf{M} = \rho A h \mathbf{I}_{2M-2}, \quad V = h \sum_{l=1}^{M} \mathcal{V}_l. \tag{63}$$

Here $\mathbf{I}_{2M-2}$ is the $(2M-2) \times (2M-2)$ identity matrix. A proof is obtained immediately by noting that

$$\dot{V} = \dot{\boldsymbol{\zeta}}^\mathsf{T} \nabla_{\boldsymbol{\zeta}} V + \dot{\boldsymbol{\eta}}^\mathsf{T} \nabla_{\boldsymbol{\eta}} V = -\dot{\mathbf{u}}^\mathsf{T} \mathbf{D}_+ \nabla_{\boldsymbol{\zeta}} V - \dot{\mathbf{v}}^\mathsf{T} \mathbf{D}_+ \nabla_{\boldsymbol{\eta}} V = -\dot{\mathbf{q}}^\mathsf{T} \begin{bmatrix} \mathbf{D}_+ \nabla_{\boldsymbol{\zeta}} V \\ \mathbf{D}_+ \nabla_{\boldsymbol{\eta}} V \end{bmatrix}. \tag{64}$$

Thus, left-multiplying (62) by $h\dot{\mathbf{q}}^\mathsf{T}$ yields (63). A split potential form is also available, via (57b). In this case,

$$\mathbf{K} = -T_0 h \begin{bmatrix} \mathbf{D}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{bmatrix}, \quad V' = \frac{h(EA - T_0)}{2} \sum_{l=1}^{M} \left(\sqrt{(1+\eta_l)^2 + \zeta_l^2} - 1 \right)^2 \tag{65}$$

4.2.2. Numerical methods

System (62) may be integrated in time in a number of ways. First, introduce the discrete time vector $\mathbf{q}^n$, approximating $\mathbf{q}(t)$ at the time $t_n = kn$. Furthermore, let $V^n = V(\boldsymbol{\zeta}^n, \boldsymbol{\eta}^n)$. An explicit time stepping scheme is obtained by application of the Störmer-Verlet algorithm:

$$\mathbf{q}^{n+1} = 2\mathbf{q}^{n+1} - \mathbf{q}^{n-1} + \frac{k^2}{\rho A h} \begin{bmatrix} \mathbf{D}_+ \nabla_{\boldsymbol{\zeta}^n} V^n \\ \mathbf{D}_+ \nabla_{\boldsymbol{\eta}^n} V^n \end{bmatrix}. \tag{66}$$

While simple, this scheme does not conserve a positive discrete energy, and instabilities may occur at large displacements. A stable scheme may be obtained by an energy-conserving discretisation of the gradient [38,39]:

$$\mathbf{q}^{n+1} = 2\mathbf{q}^{n+1} - \mathbf{q}^{n-1} + \frac{k^2}{\rho A} \begin{bmatrix} \mathbf{D}_+ \mathfrak{g}_{\boldsymbol{\zeta}}^n \\ \mathbf{D}_+ \mathfrak{g}_{\boldsymbol{\eta}}^n \end{bmatrix} \tag{67}$$

where the discrete gradients are defined as

$$(\mathfrak{g}_{\boldsymbol{\zeta}}^n)_l = \frac{\mathcal{V}(\zeta_l^{n+1}, \eta_l^n) - \mathcal{V}(\zeta_l^{n-1}, \eta_l^n)}{\zeta_l^{n+1} - \zeta_l^{n-1}}, \quad (\mathfrak{g}_{\boldsymbol{\eta}}^n)_l = \frac{\mathcal{V}(\zeta_l^n, \eta_l^{n+1}) - \mathcal{V}(\zeta_l^n, \eta_l^{n-1})}{\eta_l^{n+1} - \eta_l^{n-1}}, \tag{68}$$

with $l = 1, .., M$. This scheme leads to discrete energy conservation, and unconditional stability, but it is fully implicit, and will generally require the use of iterative root finding routines such as Newton-Raphson [39].

Finally, scheme (27) results from Hamiltonian (1) with definitions (63). The split-potential form (38) is also available, via (65). In the latter case, a stability condition arises as per (44), such that

$$k \leq \sqrt{\rho A / T_0}\, h. \tag{69}$$

This is the standard Courant-Friedrichs-Lewy stability condition for the one-dimensional linear wave equation [40].

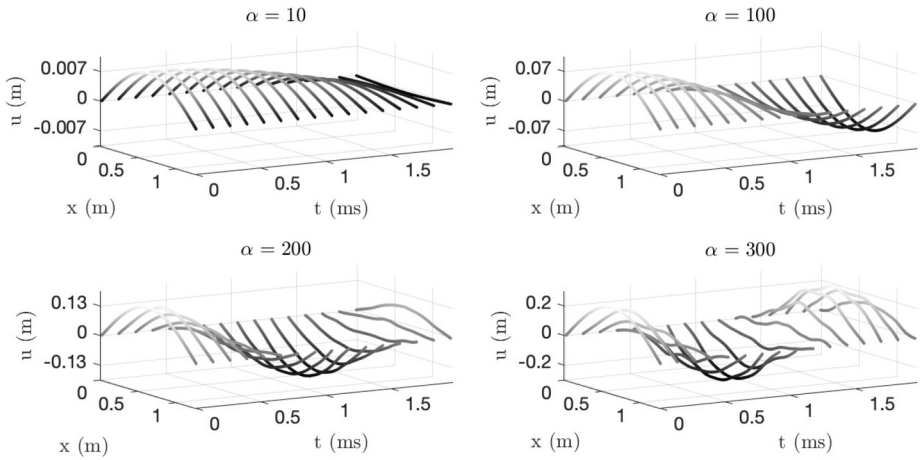

Fig. 4. Snapshots of the geometrically exact nonlinear string, computed using scheme (38), with split potential form as per (65). Here, $k = 2.6 \cdot 10^{-7}$, and the grid spacing is chosen as $h = 1.05\sqrt{E/\rho}\,k$. The string's initial normalised amplitude α is as indicated.

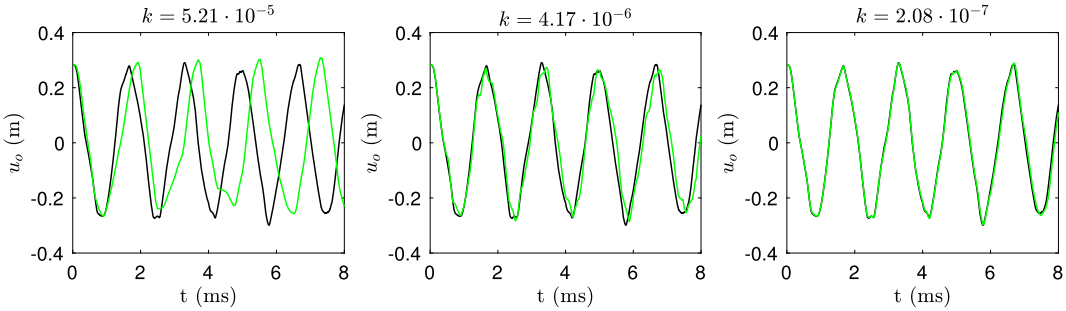

Fig. 5. Transverse displacement u (in green) for a string initialised with $\alpha = 300$, at $x = 0.5L$. In black is the reference solution, computed using scheme (66) with a time step $k = 1.04 \cdot 10^{-7}$. In green are the waveforms computed using (38), and with time steps as indicated. The grid spacing is chosen as $h = 1.05\sqrt{E/\rho}\,k$. (For interpretation of the colours in the figures, the reader is referred to the web version of this article.)

4.2.3. Numerical results

As a first example, let the string be initialised in its first linear mode of vibration in the transverse direction, that is

$$u(x,0) = \alpha\sqrt{A}\sin(\pi x/L), \quad v(x,0) = 0, \tag{70}$$

and let the initial velocity be zero for both transverse and longitudinal motion. The amplitude parameter α is nondimensional. The string parameters are taken from [41], for the C3 piano string, and are: $\rho = 7850$ kg m^{-3}; $A = 8.87 \cdot 10^{-7}$ m^2; $L = 1.259$ m; $E = 2.02 \cdot 10^{11}$ kg·s^{-2}·m^{-1}; $T_0 = 759$ kg·m·s^{-2}. Fig. 4 shows snapshots of the computed solution using scheme (38), and under various initial amplitudes α. In this case, a regularisation of the potential energy V' is employed (see Section 2.5), with a shift of $\epsilon = 10^8$. Typical amplitude-dependent phenomena are visible: the frequency of vibration increases with the initial amplitude, and the initial shape deforms progressively. In Fig. 5, the output waveform of scheme (38) are checked against a reference solution obtained using the Störmer-Verlet algorithm (66), for the large input amplitude $\alpha = 300$, indicating that the output of the two schemes converges to a common solution.

The relative energy error, as defined in (51), is shown in Fig. 6, showing conservation to near machine accuracy. The larger range of variation here, compared with the case of the Fermi-Pasta-Ulam system is a result of the much larger state size and the resulting accumulation of errors. Notice in particular that here, in contrast to the case of the Fermi-Pasta-Ulam system, even though the energy variation is extremely small, there is now a clear correlation with the numerical solution. Such an effect is highly dependent on finite wordlength effects in double precision floating point, including the precise order of operations in the final update, and within the expression used to calculate the relative error (51), and is not well understood.

Fig. 7 presents the compute times for schemes (66), (67) and (38). The fully implicit, conservative scheme (67) is the slowest, requiring a few iterations of the Newton-Raphson routine per time step. The proposed scheme (38) has compute times of the same order of magnitude as the fully explicit Störmer-Verlet algorithm, and a few orders of magnitude smaller than the fully implicit scheme. All simulations were run in Matlab, using a 2016 MacBook Pro with a 2.9 GHz Quad-Core Intel Core i7 chip. Notice that for Störmer-Verlet and the explicit Hamiltonian scheme, compute time scales with $1/k^2$ in the limit of small k, as expected in this case, as the grid spacing h has been chosen to scale directly with k as $h = 1.05\sqrt{E/\rho}k$.

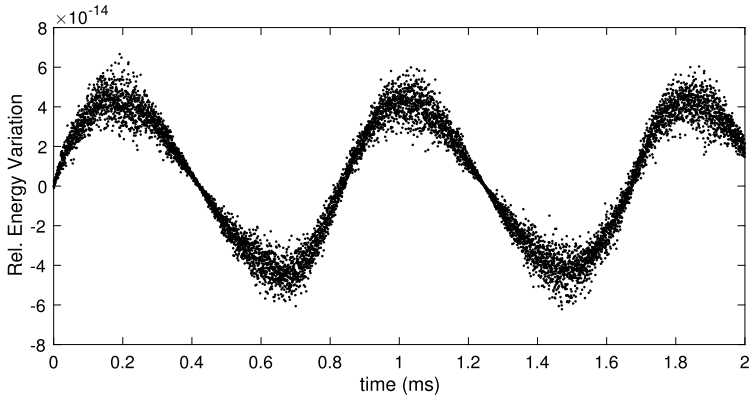

Fig. 6. Energy variation of scheme (38) for the geometrically exact nonlinear string, using the split potential form as per (65). Here, the energy error is as per (51). Here, the time step is $k = 2.4 \cdot 10^{-7}$, and the grid spacing is chosen as $h = 1.05\sqrt{E/\rho}\,k$. The string is initialised with using $\alpha = 300$.

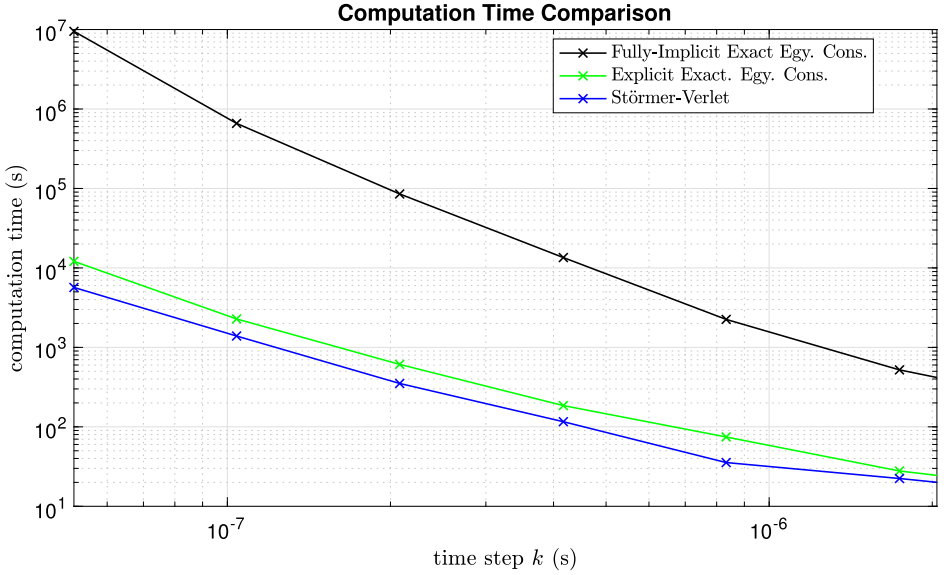

Fig. 7. Computation times, in s, for the geometrically exact nonlinear string, using the fully implicit scheme (67) (black), Störmer-Verlet (66) (blue) and the proposed scheme (38) (green). The grid spacing for a given time step is chosen as $h = 1.05\sqrt{E/\rho}\,k$. For the fully implicit scheme, Newton-Raphson is run with a tolerance of 10^{-13}, and the maximum number of iterations is limited to 20. (For interpretation of the colours in the figure, the reader is referred to the web version of this article.)

In the limit of large k, computational cost reaches a plateau, due to the small problem size in this case (i.e., the number of degrees of freedom becomes so small that computational cost can no longer be attributed to a single dominating factor).

4.3. Nonlinear plate vibration: the Föppl-von Kármán system

As a final example, consider the problem of the transverse vibration of a thin plate at high amplitudes. A commonly used model is the so-called dynamic analogue of the Föppl-von Kármán equations (see, e.g., [42–44]), that have been used extensively recently in studies of wave turbulence [45–47]. Time-stepping methods have been developed [48], including a linearly-implicit energy-conserving method [49].

Though the dynamic Föppl-von Kármán equations can be written directly in Hamiltonian form, they are most commonly presented as the following pair of coupled partial differential equations:

$$\rho\xi\,\partial_t^2 q = -\mathrm{Q}\,\Delta\Delta q + \mathcal{L}(q,F) \qquad \frac{2}{E\xi}\Delta\Delta F = -\mathcal{L}(q,q)\,. \tag{71}$$

Here, $q(x,y,t)$ and $F(x,y,t)$ are the transverse displacement of the plate and Airy stress function respectively; both are functions of spatial coordinates $(x,y) \in \mathcal{D} \subset \mathbb{R}^2$, and time $t \geq 0$. In this simple example, the spatial domain $\mathcal{D}$ will be taken to be the square of side length L m, so that $\mathcal{D} = [0,L]^2$. ∂_t and Δ represent partial differentiation with respect to time t

and the 2D Laplacian operator, respectively. $\Delta\Delta$ is thus the biharmonic operator. For simplicity, boundary conditions are assumed to be of simply supported type over the boundary $\partial\mathcal{D}$ of $\mathcal{D}$, so that

$$q = \Delta q = 0 \qquad F = \Delta F = 0 \qquad \text{over} \quad \partial\mathcal{D}\,. \tag{72}$$

The various constants that appear in (71) are: ρ, the material density, in kg·m^{-3}; ξ, the plate thickness, in m; E, Young's modulus, in kg·s^{-2}·m^{-1}; and the flexural rigidity $Q = E\xi^3/12(1-\nu^2)$, where ν is Poisson's ratio for the plate material. $\mathcal{L}$ is a bilinear operator, defined in terms of its action on two functions $f(x,y)$ and $g(x,y)$ as

$$\mathcal{L}(f,g) = \partial_x^2 f\, \partial_y^2 g + \partial_y^2 f\, \partial_x^2 g - 2\partial_x\partial_y f\, \partial_x\partial_y g\,, \tag{73}$$

where ∂_x and ∂_y represent partial differentiation with respect to x and y, respectively. Notice that only the first of (71) is dynamic; the pair of equations (71) could be rewritten as a single equation in displacement q alone, and would constitute a second order in time cubic nonlinear PDE.

System (71) is Hamiltonian, with the total energy defined by

$$H = \iint\limits_{\mathcal{D}} \frac{\rho\xi}{2}(\partial_t q)^2 + \frac{Q}{2}(\Delta q)^2 + \frac{1}{2E\xi}(\Delta F)^2 d\sigma\,. \tag{74}$$

This particular form of the energy holds under fixed edge boundary conditions (such as simply supported (72)), and must be modified under other types of conditions, such as free-edge. Notice that the final two terms under the integral above, which correspond to the potential energy, separate into a quadratic form in q, representing stored energy due to linear effects, and a quadratic form in F, representing additional nonlinear effects; both terms are individually non-negative, signalling that in numerical design, a splitting of the form described in Section 2.4 is available.

4.3.1. Semi-discrete form

For the square region $\mathcal{D}$, of side length L, one may start by defining grid locations $x_l = lh$, $y_m = mh$, where $l, m = 1, \ldots, M-1$, for some integer M such that $M = L/h$, where h is a grid spacing. The semi-discrete grid functions $q_{l,m}(t)$ and $F_{l,m}(t)$ thus represent approximations to $q(x,y,t)$ and $F(x,y,t)$ at $x = x_l$ and $y = y_m$, respectively.

Approximations to the spatial derivative operators ∂_x and ∂_y may be written, in terms of their action on a grid function $u_{l,m}(t)$ (such as $q_{l,m}$ or $F_{l,m}$ as defined above), as

$$D_{x\pm}u_{l,m} = \mp\frac{1}{h}\left(u_{l,m} - u_{l\pm1,m}\right) \qquad D_{y\pm}u_{l,m} = \mp\frac{1}{h}\left(u_{l,m} - u_{l,m\pm1}\right)\,. \tag{75}$$

Under simply supported conditions, when grid points outside the range $l, m = 1, \ldots, M-1$ are referred to in the definitions above, such values are assumed to be zero. These are the most basic forward and backward approximations to derivatives, but are sufficient for the present purposes—more elaborate approximations (such as those of spectral type [50]) are available.

From the basic operations defined in (75), centred approximations to the Laplacian Δ and biharmonic operator $\Delta\Delta$ follow as

$$D_\Delta = D_{x+}D_{x-} + D_{y+}D_{y-} \qquad D_{\Delta\Delta} = D_\Delta D_\Delta\,. \tag{76}$$

It is important to note that this particular construction of the biharmonic approximation $D_{\Delta\Delta}$, through a product of Laplacian approximations D_Δ under fixed conditions ensures that the simply supported conditions (72) are satisfied.

A semi-discrete approximation to (71) then follows as

$$\rho\xi\ddot{q}_{l,m} = -Q\, D_{\Delta\Delta}q_{l,m} + \ell(q,F) = 0 \qquad \frac{2}{E\xi}D_{\Delta\Delta}F_{l,m} = -\ell(q,q)\,. \tag{77}$$

The operator $\ell(\cdot,\cdot)$ approximates $\mathcal{L}(\cdot,\cdot)$, as defined in (73). One useful centred approximation, operating on two grid functions $f_{l,m}$ and $g_{l,m}$ is [49]:

$$\begin{aligned}\ell(f,g) = {} & D_{x+}D_{x-}f\, D_{y+}D_{y-}g + D_{y+}D_{y-}f\, D_{x+}D_{x-}g \\ & -\tfrac{1}{2}D_{x+}D_{y+}f\, D_{x+}D_{y+}g - \tfrac{1}{2}D_{x+}D_{y-}f\, D_{x+}D_{y-}g - \tfrac{1}{2}D_{x-}D_{y+}f\, D_{x-}D_{y+}g - \tfrac{1}{2}D_{x-}D_{y-}f\, D_{x-}D_{y-}g\,.\end{aligned} \tag{78}$$

It is useful to represent the semi-discrete ODE system (77) in vector form, using $(M-1)^2 \times 1$ vectors $\mathbf{q}$ and $\mathbf{F}$. In this representation, the operators D_{x+} and D_{y+} become $(M-1)^2 \times M(M-1)$ matrices $\mathbf{D}_{x+}$ and $\mathbf{D}_{y+}$, and $\mathbf{D}_{x-} = -\mathbf{D}_{x+}^\mathsf{T}$ and $\mathbf{D}_{y-} = -\mathbf{D}_{y+}^\mathsf{T}$. The operators D_Δ and $D_{\Delta\Delta}$ become $(M-1)^2 \times (M-1)^2$ matrices $\mathbf{D}_\Delta$ and $\mathbf{D}_{\Delta\Delta}$ respectively. One arrives at the form

$$\rho\xi\ddot{\mathbf{q}} = -Q\,\mathbf{D}_{\Delta\Delta}\mathbf{q} + \ell(\mathbf{q},\mathbf{F}) = 0 \qquad \frac{2}{E\xi}\mathbf{D}_{\Delta\Delta}\mathbf{F} = -\ell(\mathbf{q},\mathbf{q})\,. \tag{79}$$

This second order in time ODE system serves as the starting point for methods such as Störmer-Verlet and other linearly-implicit energy-conserving methods, as described below.

By introducing the momentum variable $\mathbf{p} = \mathbf{M}\dot{\mathbf{q}}$, the first order system (3) equivalent to (79) results from a Hamiltonian of the form of (1), with $N = (M-1)^2$, and where

$$\mathbf{M} = \rho\xi h^2 \mathbf{I}_{(M-1)^2} \qquad V = \frac{Q h^2}{2} \|\mathbf{D}_\Delta \mathbf{q}\|^2 + \frac{h^2}{2E\xi} \|\mathbf{D}_\Delta \mathbf{F}\|^2 . \tag{80}$$

Here, $\mathbf{I}_{(M-1)^2}$ is the $(M-1)^2 \times (M-1)^2$ identity matrix. A natural splitting of the potential energy V as in (16) follows, with

$$\mathbf{K} = Q h^2 \mathbf{D}_{\Delta\Delta} \qquad V' = \frac{h^2}{2E\xi} \|\mathbf{D}_\Delta \mathbf{F}\|^2 . \tag{81}$$

4.3.2. Numerical methods

Beginning from the second order system (79), one may introduce the discrete time vectors $\mathbf{q}^n$ and $\mathbf{F}^n$, representing approximations to q and F at $t = nk$, for integer n, and where k is the time step in s. Störmer-Verlet integration results immediately in:

$$\mathbf{q}^{n+1} = \left(2\mathbf{I}_{(M-1)^2} - \frac{Q k^2}{\rho\xi} \mathbf{D}_{\Delta\Delta}\right) \mathbf{q}^n - \mathbf{q}^{n-1} + \frac{k^2}{\rho\xi} \ell(\mathbf{q}^n, \mathbf{F}^n) \qquad \frac{2}{E\xi} \mathbf{D}_{\Delta\Delta} \mathbf{F}^n = -\ell(\mathbf{q}^n, \mathbf{q}^n) . \tag{82}$$

The first of these updates is explicit, but relies on $\mathbf{F}^n$, which must be obtained from the second equation through the solution of a linear system involving the biharmonic operator $\mathbf{D}_{\Delta\Delta}$.

The Störmer-Verlet scheme above is not conservative, and is prone to numerical instability. A slight variant, however, leads to a scheme with exact energy conservation:

$$\mathbf{q}^{n+1} = \left(2\mathbf{I}_{(M-1)^2} - \frac{Q k^2}{\rho\xi} \mathbf{D}_{\Delta\Delta}\right) \mathbf{q}^n - \mathbf{q}^{n-1} + \frac{k^2}{2\rho\xi} \ell(\mathbf{q}^n, \mathbf{F}^{n+1} + \mathbf{F}^{n-1}) \qquad \frac{1}{E\xi} \mathbf{D}_{\Delta\Delta} \left(\mathbf{F}^{n+1} + \mathbf{F}^n\right) = -\ell(\mathbf{q}^{n+1}, \mathbf{q}^n) . \tag{83}$$

Due to the bilinearity of the operator ℓ, this scheme is linearly implicit—at each time step, $\mathbf{q}^{n+1}$ and $\mathbf{F}^{n+1}$ must be solved simultaneously, using a linear system constructed anew at each time step; Störmer-Verlet, in contrast, requires only the solution of a linear system in $\mathbf{D}_{\Delta\Delta}$, which is of a known form. Thus it may be expected that scheme (83), while energy-conserving and provably numerically stable [49], is significantly slower to execute than the Störmer-Verlet scheme (82).

Finally, from the Hamiltonian form given in (3), with V and $\mathbf{M}$ as given in (80), as well as the splitting of the potential energy as in (81), a time-interleaved scheme of the form (38) results. This scheme possesses a non-negative exactly conserved numerical energy under the condition (44) which, in this case, reduces to a lower bound on the grid spacing h in terms of the time step k:

$$h \geq h_{\min} = 2\sqrt{k} \left(D/\rho\xi\right)^{\frac{1}{4}} . \tag{84}$$

The scheme, like Störmer-Verlet, relies on a linear system solution involving the biharmonic operator $\mathbf{D}_{\Delta\Delta}$, but is otherwise explicit.

4.3.3. Numerical examples

As an example, consider initialisation of the Föppl-von Kármán system (71) through its lowest linear mode shape:

$$q(x, y, 0) = \alpha\xi \sin(\pi x/L) \sin(\pi y/L) \qquad \partial_t q|_{x,y,t=0} = 0 , \tag{85}$$

where the dimensionless parameter α controls the maximum amplitude of the initial condition relative to the plate thickness ξ. Furthermore, the plate is assumed to be made of steel with $E = 2 \times 10^{11}$ Pa, $\rho = 7850$ kg·m^{-3}, and $\nu = 0.3$, and to be of thickness $\xi = 2$ mm and side length $L = 0.5$ m.

Using the scheme (38), with the grid spacing and time step chosen according to (84), typical amplitude-dependent behaviour is observed. See Fig. 8. At a low initial condition amplitude of $\alpha = 0.01$, behaviour is essentially linear. At amplitudes near the plate thickness at $\alpha = 2$, the period of oscillation decreases, and spontaneous mode generation is observed, and for large amplitudes, such as $\alpha = 10$, turbulent behaviour is observed. Under such stringent high-amplitude conditions ($\alpha = 10$), scheme (38) remains stable, and results converge to those of the reference solution, computed using Störmer-Verlet with a time step of $k = 2.5 \times 10^{-6}$ s. No regularisation (see Section 2.5) is used in this case. See Fig. 9. Under these conditions, Störmer-Verlet is unstable for $k > 2 \times 10^{-5}$ s.

The simulation of the Föppl-von Kármán system is computationally intensive. Most interesting in this case is a comparison of computation times between Störmer-Verlet (82), the linearly-implicit energy-conserving method (83), and the

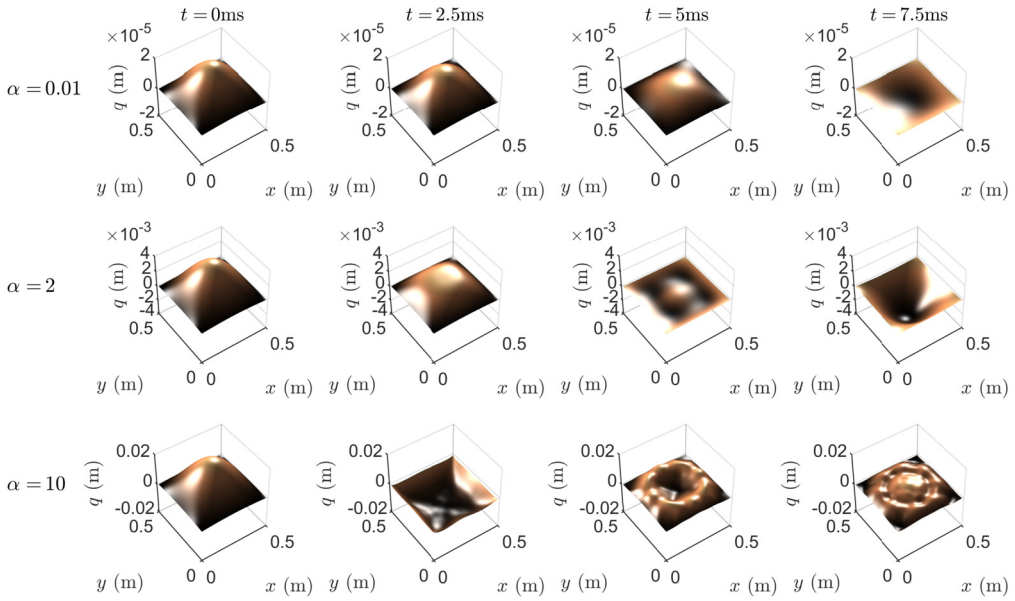

Fig. 8. Snapshots of the time evolution of a Föppl-von Kármán plate, at times as indicated, and for different initial condition amplitudes $\alpha = 0.01$, $\alpha = 2$ and $\alpha = 10$. Results are computed using scheme (38), with a time step of $k = 5 \times 10^{-6}$.

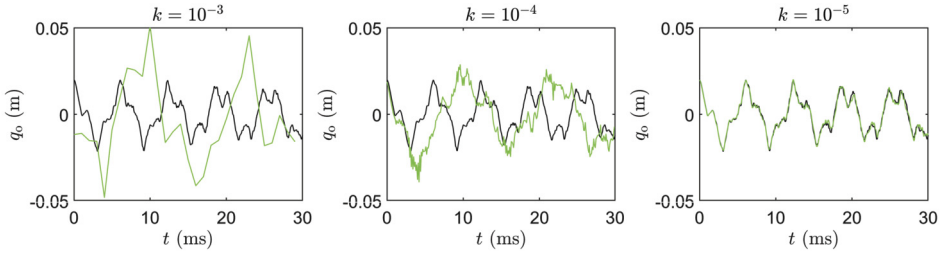

Fig. 9. The output waveform $q_o(t)$ drawn from a plate vibrating at high amplitude, with $\alpha = 10$. The reference solution, generated using Störmer-Verlet with $k = 2.5 \times 10^{-6}$ s is shown in black; results are shown for scheme (38), in green, at different time steps k, as indicated. (For interpretation of the colours in the figures, the reader is referred to the web version of this article.)

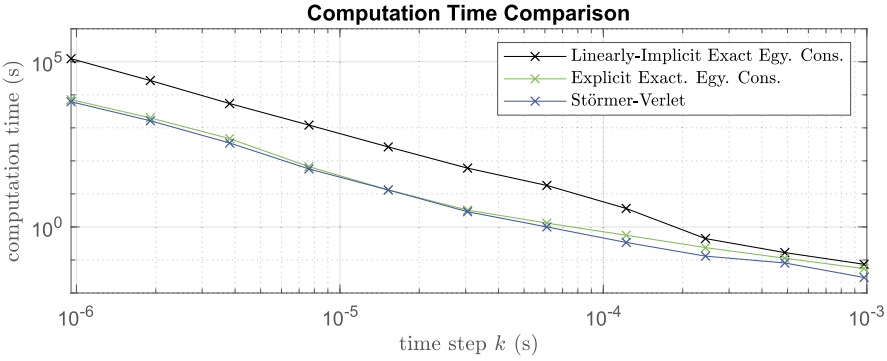

Fig. 10. Computation times, in s, for the Föppl-von Kármán plate, for a 1 s simulation duration, using the linearly implicit exact energy conserving method (83), the explicit exact energy conserving method (38), and Störmer-Verlet (82). (For interpretation of the colours in the figure, the reader is referred to the web version of this article.)

explicit energy-conserving scheme (38). Unavoidable in all cases is some form of linear system solution; for (82) and (38), this will involve the biharmonic operator, which is constant over the course of a simulation, and thus amenable to factorisation techniques (e.g. Cholesky) to decrease solution times. For the linearly-implicit exact energy conserving method (83), however, this is not the case—the linear system to be solved must be constructed anew at each time step. This is reflected in timings, as shown in Fig. 10, for simulations of 1 second for different choices of time step k. Computation was performed

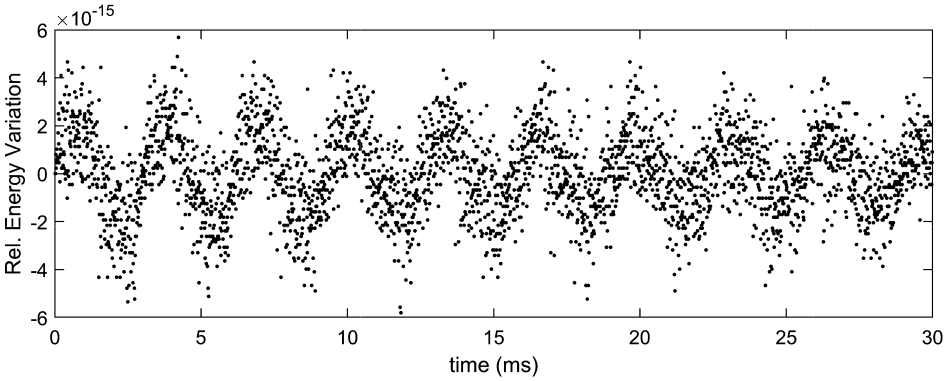

Fig. 11. Relative numerical energy variation, as defined in (51), for scheme (38) for the Föppl-von Kármán plate, with a high initial condition of the form of (85), with $\alpha = 10$. The time step is $k = 10^{-5}$ s.

in Matlab on a Lenovo P50 with an Intel Xeon E3 v5. As can be seen, computation time for the explicit scheme (38) is far lower than for scheme (83), and very nearly on par with Störmer-Verlet.

Finally, see Fig. 11, illustrating the relative energy variation for scheme (38) for the Föppl-von Kármán plate; as in the case of nonlinear string vibration, the energy variation is of the order of 10^{-15}, with some correlation with the numerical solution visible—see Fig. 6 for comparison.

5. Concluding remarks

The design of exactly energy-conserving methods for Hamiltonian systems has progressed from fully implicit designs through, more recently, to explicit methods for which exact energy conservation can be attained through an approximation to a continuous integral of the potential energy, or, more importantly, to linearly implicit designs based on invariant energy quadratisation. The main novelty in this paper is to call attention to structure within such linearly implicit designs—structure that can be exploited in order arrive at fully explicit methods. These exactly conservative methods are of roughly the same computational cost as the most efficient non-conservative explicit methods—with the additional feature of a clear means of ensuring numerical stability, either unconditionally, or, if a splitting of the potential energy is employed, under well-defined conditions on the time step that are independent of the initial conditions. Accuracy is of second order, and is plainly evident in the centred (but interleaved) discretisation approach, and borne out by simulation results (see Section 4.1). It is not clear whether it is possible to extend this framework to obtain higher order accuracy—on the other hand, there is some flexibility to explore more accurate approximation to at least part of the problem, with the explicit energy-conserving framework presented here. See the remark "Generalized Update for $\mathbf{q}^n$" in Section 3.5. In the more general context of SAV schemes, higher-order accurate schemes have been proposed recently [51].

These methods are not completely general, and require, additionally, a condition of non-negativity on the potential energy, as per invariant energy quadratisation approaches. More generally, given that the dynamics of a system are independent of shifts in the potential energy by a constant (i.e., a gauge), a more general condition is that the potential energy is bounded from below. The useful technique of splitting of the potential energy introduced here is a further restriction. But the restriction (8) mentioned above to non-negative expressions for potential energy V (or bounded from below) is slightly more strict than necessary. More general is a restriction to expressions $V(\mathbf{q})$ that are single-signed for all $\mathbf{q}$—and even more generally, bounded either from above or below. An important example here is the N-body problem, under a gravitational potential, for which $V(\mathbf{q}) \leq 0$. In this case, one may set, instead of (11), $V = -\frac{1}{2}\psi^2$, and the main development follows as above, with this sign change, and an explicit exactly energy conserving method follows as before. In this case, however, global bounds on solution size are not available, as the total energy itself is no longer necessarily non-negative. See the comments in [26] regarding general dynamic stability for Hamiltonian systems.

Only briefly alluded to here, in Section 3.5, is the extension to the case of variable time steps—useful in modelling systems with slow/fast dynamics, as discussed in [11]. The behaviour of such schemes remains unexplored. Another open question is the need for regularization, as introduced in Section 2.5. In two of the three cases presented here, the Fermi-Pasta-Ulam problem, and the Föppl-von Kármán system, good results were obtained without the use of such regularisation. In the case of string vibration, however, such regularisation was necessary. It would be of great interest to know the origin of this distinction. In their work on the Navier-Stokes equations, Lin et al. noted that the choice of the shift constant has an influence on the behaviour of the global error [29], though no indication on the choice of such constant is given. In all cases, however, regularisation has no impact on the explicit nature of the schemes proposed here, and only negligible impact on computational cost. It is certainly clear that, if very large values of the gauge are required, there will be a tradeoff with regard to numerical precision and roundoff errors in the resulting calculations.

As a final note, in all of the example problems described in Section 4, the potential energy function is a smooth function of the coordinates $\mathbf{q}$. This is not always the case, with collisions being a prime example. Though the methods presented

here (and most others) do extend to this case, the ramifications of the order of differentiability of the potential function on convergence rates remain unclear. For collisions, a comparison between fully-implicit methods and methods based on energy quadratisation was offered in [52], showing comparable convergence curves, though only through an ad-hoc modification of the gradient of the auxiliary variable ψ.

CRediT authorship contribution statement

Stefan Bilbao: Conceptualization, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft. **Michele Ducceschi:** Conceptualization, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft. **Fabiana Zama:** Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Michele Ducceschi reports financial support was provided by European Research Council.

Data availability

No data was used for the research described in the article.

Acknowledgements

M. Ducceschi was supported by the European Research Council (ERC), under grant 2020-StG-950084-NEMUS. For the purpose of open access, the first author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising.

References

[1] R. LaBudde, D. Greenspan, Energy and momentum conserving methods of arbitrary order for the numerical integration of equations of motion I. Motion of a single particle, Numer. Math. 25 (1976) 323–346.
[2] A. Marciniak, Energy conserving, arbitrary order numerical solutions of the N-body problem, Numer. Math. 45 (1984) 207–218.
[3] W. Strauss, L. Vazquez, Numerical solution of a nonlinear Klein-Gordon equation, J. Comput. Phys. 28 (1978) 271–278.
[4] D. Greenspan, Conservative numerical methods for $\ddot{x} = f(x)$, J. Comput. Phys. 56 (1984) 28–41.
[5] J. Simo, N. Tarnow, K. Wong, Exact energy-momentum conserving algorithms for symplectic schemes for nonlinear dynamics, Comput. Methods Appl. Mech. Eng. 100 (1992) 63–116.
[6] T. Hughes, T. Caughey, W. Liu, Finite-element methods for nonlinear elastodynamics which conserve energy, J. Appl. Mech. 45 (1978) 366–370.
[7] G. Zhong, J. Marsden, Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators, Phys. Lett. A 133 (1988) 134–139.
[8] E. Hairer, Long-time energy conservation of numerical integrators, in: L. Pardo, A. Pinkus, E. Suli, M. Todd (Eds.), Foundations of Computational Mathematics: Santander 2005, Cambridge University Press, Cambridge, UK, 2006, pp. 162–180.
[9] R. McLachlan, M. Perlmutter, Energy drift in reversible time integration, J. Phys. A, Math. Gen. 37 (2004) L593–L598.
[10] E. Faou, E. Hairer, T. Pham, Energy conservation with non-symplectic methods: examples and counter-examples, BIT Numer. Math. 44 (2004) 699–709.
[11] F. Marazzato, A. Ern, C. Mariotti, L. Monasse, An explicit pseudo-energy conserving time-integration scheme for Hamiltonian dynamics, Comput. Methods Appl. Mech. Eng. 347 (2019) 906–927.
[12] L. Vu-Quoc, S. Li, Invariant-conserving finite difference algorithms for the nonlinear Klein-Gordon equation, Comput. Methods Appl. Mech. Eng. 107 (1993) 341–391.
[13] J. Diaz, M. Grote, Energy conserving explicit local time stepping for second-order wave equations, SIAM J. Sci. Comput. 31 (2009) 1985–2014.
[14] Y. Chin, C. Qin, Explicit energy-conserving schemes for the three-body problem, J. Comput. Phys. 83 (1989) 485–493.
[15] B. Shadwick, J. Bowman, P. Morrison, Exactly conservative integrators, SIAM J. Appl. Math. 59 (1998) 1112–1133.
[16] G. Quispel, D. McLaren, A new class of energy-preserving numerical integration methods, J. Phys. A, Math. Theor. 41 (2008) 1–7.
[17] L. Brugnano, F. Iavernaro, D. Trigiante, A two-step, fourth-order method with energy preserving properties, Comput. Phys. Commun. 183 (2012) 1860–1868.
[18] X. Yang, D. Han, Linearly first- and second-order, unconditionally energy stable schemes for the phase field crystal model, J. Comput. Phys. 330 (2016) 1116–1134.
[19] X. Yang, J. Zhao, Q. Wang, Numerical approximations for the molecular beam epitaxial growth model based on the invariant energy quadratization method, J. Comput. Phys. 333 (2017) 104–127.
[20] J. Shen, J. Xu, J. Yang, The scalar auxiliary variable (SAV) approach for gradient flows, J. Comput. Phys. 353 (2018) 407–416.
[21] Y. Gong, J. Zhao, Energy-stable Runge–Kutta schemes for gradient flow models using the energy quadratization approach, Appl. Math. Lett. 94 (2019) 224–231.
[22] H. Zhang, X. Qian, S. Song, Novel high-order energy-preserving diagonally implicit Runge–Kutta schemes for nonlinear Hamiltonian ODEs, Appl. Math. Lett. 102 (2020) 1–9.
[23] S. Sato, Y. Miyatake, J. Butcher, High-order linearly implicit schemes conserving quadratic invariants, available at: arXiv:2203.00944v1, 2021.
[24] J. Sherman, W. Morrison, Adjustment of an inverse matrix corresponding to a change in one element of a given matrix, Ann. Math. Stat. 21 (1950) 124–127.
[25] S. Bilbao, M. Ducceschi, Fast explicit algorithms for Hamiltonian numerical integration, in: Proceedings of the European Nonlinear Dynamics Conference, Lyon, France, 2022.
[26] O. Gonzalez, J. Simo, On the stability of symplectic and energy-momentum algorithms for non-linear Hamiltonian systems with symmetry, Comput. Methods Appl. Mech. Eng. 134 (1996) 197–222.
[27] W. Terrell, Stability and Stabilization, Princeton University Press, Princeton, NJ, 2009.

[28] Z. Liu, X. Li, The exponential scalar auxiliary variable (e-sav) approach for phase field models and its explicit computing, SIAM J. Sci. Comput. 42 (2020) B630–B655.
[29] L. Lin, Z. Yang, S. Dong, Numerical approximation of incompressible Navier-Stokes equations based on an auxiliary energy variable, J. Comput. Phys. 388 (2019) 1–22.
[30] E. Hairer, C. Lubich, G. Wanner, Geometric numerical integration illustrated by the Störmer–Verlet method, Acta Numer. 12 (2003) 399–450.
[31] J. Zhao, A revisit of the energy quadratization method with a relaxation technique, Appl. Math. Lett. 120 (2021) 107331.
[32] M. Jiang, Z. Zhang, J. Zhao, Improving the accuracy and consistency of the scalar auxiliary variable (SAV) method with relaxation, J. Comput. Phys. 456 (2022) 110954.
[33] E. Fermi, J. Pasta, S. Ulam, Studies of nonlinear problems, Technical Report Los Alamos LA-1940, 1955.
[34] E. Hairer, C. Lubich, G. Wanner, Geometric Numerical Integration, Springer Series in Computational Mathematics, 2006.
[35] P. Morse, U. Ingard, Theoretical Acoustics, Princeton University Press, Princeton, NJ, USA, 1968.
[36] J. Chabassier, P. Joly, Energy preserving schemes for nonlinear Hamiltonian systems of wave equations: application to the vibrating piano string, Comput. Methods Appl. Mech. Eng. 199 (2010) 2779–2795.
[37] M. Ducceschi, S. Bilbao, Simulation of the geometrically exact nonlinear string via energy quadratisation, J. Sound Vib. 534 (2022) 117021.
[38] T. Itoh, K. Abe, Hamiltonian-conserving discrete canonical equations based on variational difference quotients, J. Comput. Phys. 76 (1988) 85–102.
[39] M. Ducceschi, S. Bilbao, Non-iterative, conservative schemes for geometrically exact nonlinear string vibration, in: Proceedings of the International Conference on Acoustics (ICA 2019), Aachen, Germany, 2019.
[40] R. Courant, K. Friedrichs, H. Lewy, On the partial differential equations of mathematical physics, Math. Ann. 100 (1928) 32–74 (in German).
[41] J. Chabassier, M. Duruflé, Physical parameters for piano modeling, Technical Report, 2012, available at: https://hal.inria.fr/hal-00688679v1/document.
[42] T. von Kármán, Festigkeitsprobleme im Maschinenbau, in: Encyklopädie der Mathematischen Wissenschaften, vol. 4, 1910, pp. 311–385.
[43] A. Föppl, Vorlesungen über technische Mechanik, Druck und Verlag von B.G. Teubner, Leipzig, 1907.
[44] A. Nayfeh, D. Mook, Nonlinear Oscillations, John Wiley and Sons, New York, NY, 1979.
[45] M. Ducceschi, O. Cadot, C. Touzé, S. Bilbao, Dynamics of the wave turbulence spectrum in vibrating plates: a numerical investigation using a conservative finite difference scheme, Phys. D: Nonlinear Phenom. 280–281 (2014) 73–85.
[46] N. Yokoyama, M. Takaoka, Weak and strong wave turbulence spectra for elastic thin plate, Phys. Rev. Lett. 110 (2013) 105501.
[47] G. Düring, C. Josserand, S. Rica, Wave turbulence theory of elastic plates, Phys. D: Nonlinear Phenom. 347 (2017) 42–73.
[48] R. Kirby, Z. Yosibash, Solution of von Kármán dynamic non-linear plate equations using a pseudo-spectral method, Comput. Methods Appl. Mech. Eng. 193 (2004) 575–599.
[49] S. Bilbao, A family of conservative finite difference schemes for the dynamical von Kármán plate equations, Numer. Methods Partial Differ. Equ. 24 (2008) 193–216.
[50] R. Vichnevetsky, J. Bowles, Fourier Analysis of Numerical Approximations of Hyperbolic Equations, SIAM, Philadelphia, PA, 1982.
[51] Y. Gong, J. Zhao, Q. Wang, Arbitrarily high-order unconditionally energy stable SAV schemes for gradient flow models, Comput. Phys. Commun. 249 (2020) 107033.
[52] M. Ducceschi, S. Bilbao, S. Willemsen, S. Serafin, Linearly-implicit schemes for collisions in musical acoustics based on energy quadratisation, J. Acoust. Soc. Am. 149 (2021) 3502–3516.