Journal of Computational Physics 458 (2022) 111098

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Energy conserving successive multi-stage method for the linear wave equation

Jaemin Shin [a], June-Yub Lee [b,*]

[a] Department of Mathematics, Chungbuk National University, Cheongju 28644, Republic of Korea
[b] Department of Mathematics, Ewha Womans University, Seoul 03760, Republic of Korea

ARTICLE INFO

Article history:
Received 18 August 2021
Received in revised form 4 February 2022
Accepted 21 February 2022
Available online 25 February 2022

Keywords:
Linear wave equation
Energy conservation
High-order time accuracy
Successive Multi-Stage (SMS) method

ABSTRACT

We propose a new high-order multi-stage method to solve the linear wave equation in an unconditionally energy stable manner. This Successive Multi-Stage (SMS) method is extended from the Crank–Nicolson method and unconditional energy conservation is guaranteed. We develop up to the sixth-order SMS method using the order conditions for Runge–Kutta methods and provide mathematical arguments showing that the SMS method is a different branch from well-known high order energy preserving methods for Hamiltonian systems. We present a proof of the unique solvability and numerically demonstrate the accuracy and stability of the proposed methods compared with comparisons.

1. Introduction

Linear wave equations appear in many fields of physics and have played a significant role in mathematical modeling for transient physical phenomena. For instance, they arise when considering the Maxwell equations for electromagnetism [1,2], the acoustic equation for sound propagation [3,4], and the elastodynamic equation for wave propagation in solids [5,6]. One of the most important properties of wave equations is energy conservation.

In this paper, we introduce a high-order energy conserving numerical scheme for the linear wave equation,

$$\kappa u_{tt} = \nabla \cdot (\mathbf{M} \nabla u), \tag{1}$$

where $\mathbf{M} = \mathbf{M}(\mathbf{x})$ is a symmetric positive definite matrix and $\kappa = \kappa(\mathbf{x})$ is a strictly positive function. Equation (1) might be completed with specific boundary conditions in a bounded domain $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$. For simplicity, we consider the periodic boundary condition along the edges of computational domain where $\kappa(\mathbf{x})$ and $\mathbf{M}(\mathbf{x})$ are constants representing the homogeneous background medium. Homogeneous Neumann boundary condition $\mathbf{n} \cdot (\mathbf{M} \nabla u) = 0$ or homogeneous Dirichlet boundary condition $u = 0$ can be easily imposed using even or odd extension of the solution, respectively. The corresponding energy of (1) can be written as a sum of the kinetic and the potential energies of the system,

$$E(t) = \frac{1}{2} \int_{\Omega} \left(\kappa u_t^2 + \left| \mathbf{M}^{\frac{1}{2}} \nabla u \right|^2 \right) d\mathbf{x}, \tag{2}$$

* Corresponding author.
E-mail address: jyllee@ewha.ac.kr (J.-Y. Lee).

https://doi.org/10.1016/j.jcp.2022.111098
0021-9991/

where $\left|\mathbf{M}^{\frac{1}{2}}\nabla u\right|^2 = (\nabla u, \mathbf{M}\nabla u)$. It is worth noting that when considering the periodic or homogeneous boundary conditions, the energy (2) is conserved with respect to time, as demonstrated by integration by parts:

$$\frac{d}{dt}E(t) = \int_{\Omega} \left(\kappa u_t u_{tt} - u_t \nabla \cdot (\mathbf{M}\nabla u)\right) d\mathbf{x} = 0. \tag{3}$$

Note that non-zero Dirichlet or Neumann boundary conditions can be also considered using an additive patch to the solution, however, the energy defined in (2) is no longer conserved without further modification.

There have been plenty of discussions for numerical methods to solve the wave equations. For the long-time simulation, consequently, energy-conserving high-order methods have been attracting much attention to determine the phase and shape of the waves as accurately and stably as possible. Also, there has been intensive research on the numerical treatment of quasi-linear wave equations or even Hamiltonian partial differential equations (PDEs) over the past decades. For example, the average vector field method [7] has been applied to the Hamiltonian PDEs with constant symplectic structure. And, the discrete variational derivative method [8] has been proposed to the family of nonlinear wave equations by constructing the discrete version of the energy function. There are also studies focused on finding conditions for energy conservation in RK methods, for example, a symplectic RK method was developed in [9–11].

Our ultimate goal is to present a new RK-based high-order energy preserving method for quasi-linear wave equations or wide range of Hamiltonian systems. To demonstrate that our method is in a different branch from existing methods, we focus on the linear wave equations for the simplicity of argument in this paper. Considering energy-conserving methods for linear wave equations, it is noteworthy referring to the approach, based on the leap-frog method, to be compared with our method. The leap-frog method is a well-known multi-step method to solve the wave equation and has been used for many applications and simulations [12]. It is accurate to the second-order, but stability is not guaranteed with larger time step sizes. Many studies have introduced energy-conserving methods, such as theta methods [13–15], by generalizing the leap-frog method. On the other hand, it is well known that the standard Crank–Nicolson method preserves the conservation laws of the linear wave equation, but it just has the second-order accuracy. So we extend the Crank–Nicolson method to construct the high-order method with guaranteeing energy conservation and demonstrate that the Crank–Nicolson method based RK method is different from the existing RK ones.

In this paper, we propose a successive multi-stage (SMS) method as a high-order energy conserving numerical scheme. We start by constructing a framework that guarantees energy conservation. The proposed method can be considered as an RK method, making it is easy to find a proper coefficient set for high-order accuracy. We note that SMS methods are different from the existing symplectic RK methods. Specifically, our proposed methods do not satisfy the condition $b_i b_j - b_i a_{ij} - b_j a_{ji} = 0$ in [11]. That is, SMS methods constitute a new class of high-order multi-stage methods that guarantee energy conservation.

We first provide a proof that the Crank–Nicolson method is unconditionally energy conserving in Section 2, and we extend this idea to an s-stage SMS method in Section 3. In Sections 4 and 5, we numerically demonstrate the order of the accuracy as well as energy conservation in one and higher dimensions. Finally, conclusions are drawn in Section 6. In addition, we briefly introduce the derivation of order conditions for the linear Runge–Kutta method in the Appendix. It is worth noting that we only consider temporal accuracy in this paper, thus semi-discrete numerical schemes are presented. Fourier spectral methods are used for spatial derivatives, and all simulations are executed using the MATLAB program via a fast Fourier transform.

2. Classical Crank–Nicolson method

By defining an auxiliary variable $v = u_t$, we can represent the linear wave equation (1) in canonical form as

$$\begin{aligned} u_t &= v, \\ v_t &= \frac{1}{\kappa}\nabla \cdot (\mathbf{M}\nabla u) \end{aligned} \tag{4}$$

and the energy (2) in Hamiltonian form as

$$\mathcal{H}(u, v) = \frac{1}{2}\int_{\Omega} \left(\kappa v^2 + \left|\mathbf{M}^{\frac{1}{2}}\nabla u\right|^2\right) d\mathbf{x}. \tag{5}$$

For the semi-discrete formulation, we denote u^n and v^n as approximations of $u(\cdot, t^n)$ and $v(\cdot, t^n)$, where $t^n = n\Delta t$ and Δt is a time step size. We first consider the well-known Crank–Nicolson method

$$\begin{aligned} \frac{u^{n+1} - u^n}{\Delta t} &= \frac{v^{n+1} + v^n}{2}, \\ \frac{v^{n+1} - v^n}{\Delta t} &= \frac{1}{\kappa}\nabla \cdot \left(\mathbf{M}\nabla \frac{u^{n+1} + u^n}{2}\right). \end{aligned} \tag{6}$$

Lemma 1. *The Crank–Nicolson method (6) is unconditionally energy conserving, meaning that for any time step size Δt,*

$$\mathcal{H}\left(u^{n+1}, v^{n+1}\right) = \mathcal{H}\left(u^n, v^n\right). \tag{7}$$

Proof. We first calculate the difference in energy functionals,

$$\begin{aligned}&\mathcal{H}\left(u^{n+1}, v^{n+1}\right) - \mathcal{H}\left(u^n, v^n\right)\\ &\quad= \frac{1}{2}\int_\Omega \kappa\left(v^{n+1}\right)^2 - \kappa\left(v^n\right)^2 + \left|\mathbf{M}\nabla u^{n+1}\right|^2 - \left|\mathbf{M}\nabla u^n\right|^2 d\mathbf{x}.\end{aligned} \tag{8}$$

For the first two terms of (8), we can rearrange them as

$$\begin{aligned}&\frac{1}{2}\int_\Omega \kappa\left(v^{n+1}\right)^2 - \kappa\left(v^n\right)^2 d\mathbf{x} = \frac{1}{2}\int_\Omega \kappa\left(v^{n+1} - v^n\right)\left(v^{n+1} + v^n\right) d\mathbf{x}\\ &\quad= \frac{\Delta t}{4}\int_\Omega \left(v^{n+1} + v^n\right)\nabla\cdot\left(\mathbf{M}\nabla\left(u^{n+1} + u^n\right)\right) d\mathbf{x}\\ &\quad= -\frac{\Delta t}{4}\int_\Omega \nabla\left(v^{n+1} + v^n\right)\cdot\left(\mathbf{M}\nabla\left(u^{n+1} + u^n\right)\right) d\mathbf{x}.\end{aligned} \tag{9}$$

For the last two terms of (8), we can expand as

$$\begin{aligned}&\frac{1}{2}\int_\Omega \left|\mathbf{M}^{\frac{1}{2}}\nabla u^{n+1}\right|^2 - \left|\mathbf{M}^{\frac{1}{2}}\nabla u^n\right|^2 d\mathbf{x}\\ &\quad= \frac{1}{2}\int_\Omega \nabla\left(u^{n+1} - u^n\right)\cdot\left(\mathbf{M}\nabla\left(u^{n+1} + u^n\right)\right) d\mathbf{x}\\ &\quad= \frac{\Delta t}{4}\int_\Omega \nabla\left(v^{n+1} + v^n\right)\cdot\left(\mathbf{M}\nabla\left(u^{n+1} + u^n\right)\right) d\mathbf{x}.\end{aligned} \tag{10}$$

Adding (9) and (10), we obviously have $\mathcal{H}\left(u^{n+1}, v^{n+1}\right) - \mathcal{H}\left(u^n, v^n\right) = 0$. □

Remark 1. Lemma 1 implies that the energy corresponding to the numerical solution $\left(u^n, v^n\right)$ of the Crank–Nicolson method is a physical constant $E(t)$,

$$\mathcal{H}\left(u^n, v^n\right) = \mathcal{H}\left(u^0, v^0\right) = E\left(t^0\right). \tag{11}$$

3. Successive multi-stage methods

We now propose a successive multi-stage (SMS) method as an extension of the Crank–Nicolson method to an s-stage method, referred to as SMS(R_s), with a given coefficient vector

$$R_s = [r_1, r_2, \cdots, r_s]. \tag{12}$$

The SMS(R_s) is a one-step s-stage method that computes the next approximation $\left(u^{n+1}, v^{n+1}\right)$ from $\left(u^n, v^n\right)$. We set $u_0 = u^n$ and $v_0 = v^n$ for the initial-stage, and then calculate the intermediate value (u_i, v_i) by solving

$$\begin{aligned}\frac{u_i - u_{i-1}}{\Delta t} &= r_i\left(v_i + v_{i-1}\right),\\ \frac{v_i - v_{i-1}}{\Delta t} &= \frac{r_i}{\kappa}\nabla\cdot\left(\mathbf{M}\nabla\left(u_i + u_{i-1}\right)\right),\end{aligned} \tag{13}$$

for $i = 1, 2, \cdots, s$. Finally, we have $u^{n+1} = u_s$ and $v^{n+1} = v_s$. We note that SMS(R_1) with a coefficient vector

$$R_1 = \left[\tfrac{1}{2}\right] \tag{14}$$

is identical to the Crank–Nicolson method (6), which provides second-order accuracy. For higher order accuracy, we will describe the order conditions for the coefficient vector R_s in Section 3.3. Before introducing a specific example, we provide proofs of the unique solvability and energy conservation in Sections 3.1 and 3.2, respectively.

3.1. Unique solvability

Theorem 2. *The SMS(R_s) method (13) is uniquely solvable for any time step size Δt.*

Proof. For each stage $i = 1, 2, \cdots, s$, we need to solve

$$\begin{aligned} u_i - r_i \Delta t\, v_i &= \varphi_i, \\ v_i - \frac{r_i}{\kappa} \Delta t\, \nabla \cdot (\mathbf{M} \nabla u_i) &= \psi_i, \end{aligned} \tag{15}$$

where $\varphi_i = u_{i-1} + r_i \Delta t\, v_{i-1}$ and $\psi_i = v_{i-1} + r_i \Delta t\, \nabla \cdot (\mathbf{M} \nabla u_{i-1})$. The system (15) can be represented as

$$u_i - \frac{r_i^2 \Delta t^2}{\kappa} \nabla \cdot (\mathbf{M} \nabla u_i) = \varphi_i + r_i \Delta t\, \psi_i, \tag{16}$$

which has a unique solution u_i for any time step size Δt, since

$$I - \frac{r_i^2 \Delta t^2}{\kappa} \nabla \cdot (\mathbf{M} \nabla) \tag{17}$$

is an invertible operator with all eigenvalues bigger than 1 for given positive function κ and negative definite operator $\nabla \cdot (\mathbf{M} \nabla)$. Using the solution u_i from (16), we can easily obtain a unique solution of v_i by (15). □

3.2. Energy conservation

Theorem 3. *For each intermediate stage of the SMS(R_s) method (13), the energy is conserved for any time step size $r_i \Delta t$,*

$$\mathcal{H}(u_i, v_i) = \mathcal{H}(u_{i-1}, v_{i-1}). \tag{18}$$

Moreover, the proposed method (13) is unconditionally energy conserving, meaning that $\mathcal{H}\left(u^{n+1}, v^{n+1}\right) = \mathcal{H}\left(u^n, v^n\right)$ for any time step size Δt.

Proof. Similarly to the proof of Lemma 1, we first consider the difference of adjacent discrete energy functionals,

$$\begin{aligned} &\mathcal{H}(u_i, v_i) - \mathcal{H}(u_{i-1}, v_{i-1}) \\ &= \frac{1}{2} \int_\Omega \kappa v_i^2 - \kappa v_{i-1}^2 + \left| \mathbf{M}^{\frac{1}{2}} \nabla u_i \right|^2 - \left| \mathbf{M}^{\frac{1}{2}} \nabla u_{i-1} \right|^2 d\mathbf{x}. \end{aligned} \tag{19}$$

The two parts of (19) can be expanded as

$$\frac{1}{2} \int_\Omega \kappa v_i^2 - \kappa v_{i-1}^2\, d\mathbf{x} = -\frac{r_i \Delta t}{2} \int_\Omega \nabla (v_i + v_{i-1}) \cdot (\mathbf{M} \nabla (u_i + u_{i-1}))\, d\mathbf{x}, \tag{20}$$

$$\frac{1}{2} \int_\Omega \left| \mathbf{M}^{\frac{1}{2}} \nabla u_i \right|^2 - \left| \mathbf{M}^{\frac{1}{2}} \nabla u_{i-1} \right|^2 d\mathbf{x} = \frac{r_i \Delta t}{2} \int_\Omega \nabla (v_i + v_{i-1}) \cdot (\mathbf{M} \nabla (u_i + u_{i-1}))\, d\mathbf{x}. \tag{21}$$

Then we have $\mathcal{H}(u_i, v_i) = \mathcal{H}(u_{i-1}, v_{i-1})$ for any $r_i \Delta t$, $i = 1, \cdots, s$. Therefore,

$$\mathcal{H}\left(u^{n+1}, v^{n+1}\right) = \mathcal{H}(u_s, v_s) = \mathcal{H}(u_0, v_0) = \mathcal{H}\left(u^n, v^n\right) \tag{22}$$

which proves conservation of the discrete energy functional $\mathcal{H}\left(u^n, v^n\right)$. □

3.3. Temporal accuracy

We can consider a framework of the Runge–Kutta (RK) method to show the time accuracy for the SMS method. Summing (13) up to the i-th stage, the difference between i-th stage (u_i, v_i) and initial-stage (u_0, v_0) values can be written as a linear combination of the previous stage values:

$$\begin{aligned} \frac{u_i - u_0}{\Delta t} &= \sum_{j=1}^{i} r_j \left(v_j + v_{j-1} \right), \\ \frac{v_i - v_0}{\Delta t} &= \sum_{j=1}^{i} \frac{r_j}{\kappa} \nabla \cdot \left(\mathbf{M} \nabla \left(u_j + u_{j-1} \right) \right). \end{aligned} \tag{23}$$

Next, (23) can be further simplified as the RK method,

$$\begin{aligned}\frac{u_i-u_0}{\Delta t}&=\sum_{j=0}^{i}a_{ij}v_j,\\ \frac{v_i-v_0}{\Delta t}&=\sum_{j=0}^{i}\frac{a_{ij}}{\kappa}\nabla\cdot(\mathbf{M}\nabla u_i),\end{aligned}\tag{24}$$

where $a_{i0}=r_1$, $a_{ii}=r_i$, and $a_{ij}=r_j+r_{j+1}$ for $j=1,2,\cdots,i-1$. Furthermore, the SMS(R_s) method (13) can be described by the Butcher table for the RK method,

$$\begin{array}{c|c}\mathbf{c}&\mathbf{A}\\ \hline &\mathbf{b}^T\end{array}=\begin{array}{c|ccccc}0&0&0&0&\cdots&0\\ c_1&r_1&r_1&0&\cdots&0\\ c_2&r_1&r_1+r_2&r_2&\cdots&0\\ \vdots&\vdots&\vdots&\vdots&\ddots&\vdots\\ c_s&r_1&r_1+r_2&r_2+r_3&\cdots&r_s\\ \hline &r_1&r_1+r_2&r_2+r_3&\cdots&r_s\end{array},\tag{25}$$

where $\mathbf{A}\in\mathbb{R}^{(s+1)\times(s+1)}$, $\mathbf{b}\in\mathbb{R}^{s+1}$, and $\mathbf{c}=\mathbf{A1}$ with $\mathbf{1}=(1,1,\ldots,1)^T\in\mathbb{R}^{s+1}$.

Because of the linearity and autonomy of the given wave equation (1), we consider the linear RK method, which is briefly explained in Appendix. The order condition for the p-th order accuracy is $\mathbf{b}^T\mathbf{A}^{q-1}\mathbf{1}=1/q!$ for $1\le q\le p$.

For numerical simulations, we need a specific table for the desired accuracy. We now construct the coefficient vector R_s from the relationship between the SMS and linear RK methods. We first consider a single coefficient vector $R_1=[r_1]$. Considering the structure of the SMS method, we have the corresponding matrix $\mathbf{A}$ and vector $\mathbf{b}$ as

$$\mathbf{A}=\begin{bmatrix}0&0\\ r_1&r_1\end{bmatrix},\quad \mathbf{b}=\begin{bmatrix}r_1\\ r_1\end{bmatrix}.\tag{26}$$

To satisfy the first-order condition $\mathbf{b}^T\mathbf{1}=1$, we need $r_1=1/2$, which also satisfies the second-order condition $\mathbf{b}^T\mathbf{A1}=1/2$. Thus, the SMS method begins with second-order accuracy and $R_1=\left[\frac{1}{2}\right]$ is only one case of the second-order SMS method.

We note that there is no solution for the two-stage coefficient vector $R_2=[r_1,r_2]$ that satisfies the order conditions up to third-order accuracy. We now choose a possible choice of the coefficient vector $R_3=[r_1,r_2,r_3]$, which implies a three-stage SMS method, and consider corresponding matrix $\mathbf{A}$ and vector $\mathbf{b}$ as

$$\mathbf{A}=\begin{bmatrix}0&0&0&0\\ r_1&r_1&0&0\\ r_1&r_1+r_2&r_2&0\\ r_1&r_1+r_2&r_2+r_3&r_3\end{bmatrix},\quad \mathbf{b}=\begin{bmatrix}r_1\\ r_1+r_2\\ r_2+r_3\\ r_3\end{bmatrix}.\tag{27}$$

To satisfy the first-order condition $\mathbf{b}^T\mathbf{1}=1$, the coefficients should satisfy

$$r_1+r_2+r_3=\frac{1}{2},\tag{28}$$

which also satisfies the second-order condition $\mathbf{b}^T\mathbf{A1}=1/2$, just as in the case of the one-stage SMS method (26). With the identity (28) and the third-order condition $\mathbf{b}^T\mathbf{A}^2\mathbf{1}=1/6$, we have the following identity:

$$\begin{aligned}\frac{1}{2}\mathbf{b}^T\mathbf{A}^2\mathbf{1}&=r_1^3+r_2^3+r_3^3+2\left(r_1^2r_2+r_1^2r_3+r_2^2r_1+r_2^2r_3+r_3^2r_1+r_3^2r_2\right)+4r_1r_2r_3\\ &=(r_1+r_2+r_3)^3-(r_1+r_2)(r_2+r_3)(r_3+r_1)\\ &=\frac{r_1+r_2+r_3}{4}-\left(\frac{1}{2}-r_3\right)\left(\frac{1}{2}-r_1\right)\left(\frac{1}{2}-r_2\right)=\frac{1}{12}.\end{aligned}\tag{29}$$

Thus,

$$r_1r_2+r_1r_3+r_2r_3-2r_1r_2r_3=\frac{1}{12}.\tag{30}$$

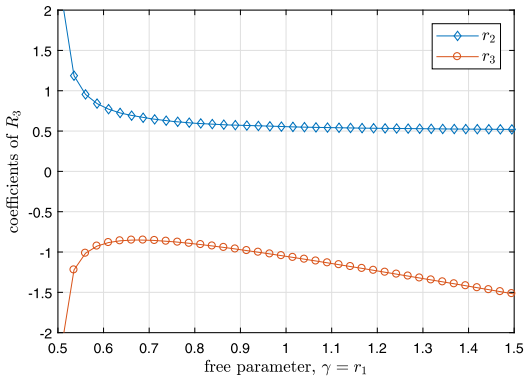

Fig. 1. Coefficients of R_3 for fourth-order accuracy.

We note that the coefficients satisfying up to third-order conditions also satisfy the fourth-order condition. In fact, the identity (28) and the condition $\mathbf{b}^T\mathbf{A}^3\mathbf{1} = 1/24$ induce the same result as in (30) since

$$\begin{aligned}\frac{1}{2}\mathbf{b}^T\mathbf{A}^3\mathbf{1} &= (r_1+r_2+r_3)^4 - 2(r_1+r_2+r_3)(r_1+r_2)(r_2+r_3)(r_3+r_1)\\ &= \frac{r_1+r_2+r_3}{4} - \frac{1}{16} - \left(\frac{1}{2}-r_3\right)\left(\frac{1}{2}-r_1\right)\left(\frac{1}{2}-r_2\right) = \frac{1}{48}.\end{aligned} \tag{31}$$

In conclusion, we need only two relations (28) and (30) to achieve the fourth-order accuracy.

Now, for finding R_3 to construct three-stage fourth-order SMS methods, we set a free parameter $r_1 = \gamma$. Then the parameters r_2 and r_3 satisfy

$$r_2 + r_3 = \frac{1-2\gamma}{2}, \quad r_2 r_3 = \frac{1-6\gamma+12\gamma^2}{12\,(1-2\gamma)}. \tag{32}$$

Fig. 1 shows the coefficients of R_3 with respect to the parameter $\gamma > 1/2$, which is the solvable region. This figure implies that there are infinitely many three-stage fourth-order SMS methods. Choosing $\gamma = 0.8$ for the numerical simulation, we have

$$R_3\,(\gamma) \approx [0.8,\ 0.599258893,\ -0.899258893]. \tag{33}$$

As we observe in the identity (28) and (30), any permutation of the solution r_1, r_2, and r_3 in R_3 satisfies the same order conditions up to four. In fact, the following theorem proves that this permutation invariant property is generally true for all s-stage SMS methods.

Theorem 4. *For an s-stage SMS method with a coefficient vector R_s, the value of $\mathbf{b}^T\mathbf{A}^p\mathbf{1}$ for $p \geq 0$ is invariant under the permutation of $R_s = [r_1, r_2, \cdots, r_s]$.*

Proof. Without loss of generality, we consider a swapped coefficient vector $\hat{R}_s := \left[r_1, \cdots, r_{k-1}, \overbrace{r_{k+1}, r_k}, r_{k+2}, \cdots, r_s\right]$ for $0 < k < s$ and prove that

$$\mathbf{b}^T\mathbf{A}^p\mathbf{1} = \hat{\mathbf{b}}^T\hat{\mathbf{A}}^p\mathbf{1} \tag{34}$$

where $\hat{\mathbf{A}}$, $\hat{\mathbf{b}}$, and $\hat{\mathbf{c}}$ are the matrix and the vectors in the Butcher table corresponding to the swapped coefficient vector $\hat{R}_s$. We observe that entries a_{ij} of $\mathbf{A}$ and $\hat{a}_{ij}$ of $\hat{\mathbf{A}}$ are differed by $d = r_{k+1} - r_k$ only at a few points, so we denote $\hat{\mathbf{A}} = \mathbf{A} + d\mathbf{D}$ where $\mathbf{D} \in \mathbb{R}^{(s+1)\times(s+1)}$ have zero entries except $D_{i,k-1} = 1, i \geq k$, $D_{k,k} = 1$, and $D_{i,k+1} = -1, i > k$. Also entries of $\hat{\mathbf{b}}$, $\hat{\mathbf{c}}$ are same with $\mathbf{b}$, $\mathbf{c}$ except

$$\hat{b}_{k-1} = b_{k-1} + d, \ \ \hat{b}_{k+1} = b_{k+1} - d, \ \text{ and } \ \hat{c}_k = c_k + 2d. \tag{35}$$

In the case of $p = 0$, (34) is obvious $\mathbf{b}^T\mathbf{1} = \hat{\mathbf{b}}^T\mathbf{1}$. Let $\mathbf{c}^{(p)} := \mathbf{A}^p\mathbf{1}$ and $\hat{\mathbf{c}}^{(p)} := \hat{\mathbf{A}}^p\mathbf{1}$, then we claim that the following identity holds for $p \geq 1$,

$$\hat{\mathbf{c}}^{(p)} = \mathbf{c}^{(p)} + \beta_p\mathbf{e}, \qquad \beta_p = \frac{d}{r_k + r_{k+1}}\left(c_{k+1}^{(p)} - c_{k-1}^{(p)}\right), \tag{36}$$

where $\mathbf{e} \in \mathbb{R}^{s+1}$ is the standard unit vector with $e_k = 1$. For $p = 1$, it is a trivial statement since $\hat{c}_k = c_k + 2d$, $\hat{c}_i = c_i$, $i \neq k$, and $\beta_1 = \frac{d}{r_k + r_{k+1}} 2\left(r_k + r_{k+1}\right) = 2d$. We prove this assertion (36) for $p > 1$ by mathematical induction.

$$\hat{\mathbf{c}}^{(p+1)} = (\mathbf{A} + d\mathbf{D})\left(\mathbf{c}^{(p)} + \beta_p \mathbf{e}\right) = \mathbf{c}^{(p+1)} + d\mathbf{D}\mathbf{c}^{(p)} + \beta_p \mathbf{A}\mathbf{e} + d\beta_p \mathbf{D}\mathbf{e}. \tag{37}$$

For $i < k$, $\hat{c}_i^{(p+1)} - c_i^{(p+1)} = 0$ since $D_{i,j} = a_{i,k} = 0$. In case of $i > k$, $\hat{c}_i^{(p+1)} - c_i^{(p+1)} = d\left(c_{k-1}^{(p)} - c_{k+1}^{(p)}\right) + \beta_p a_{i,k} = 0$ due to the mathematical induction assumption for β_p. For $i = k$, using $a_{k,k} + dD_{k,k} = r_{k+1}$, we get

$$\begin{aligned} \hat{c}_k^{(p+1)} - c_k^{(p+1)} &= d\left(c_{k-1}^{(p)} + c_k^{(p)}\right) + \beta_p r_{k+1} \\ &= \frac{d}{r_k + r_{k+1}}\left(r_k c_{k-1}^{(p)} + (r_k + r_{k+1}) c_k^{(p)} + r_{k+1} c_{k+1}^{(p)}\right). \end{aligned} \tag{38}$$

On the other hand, by the definition of $\mathbf{c}^{(p+1)} = \mathbf{A}\mathbf{c}^{(p)}$, we have

$$c_{k+1}^{(p+1)} - c_{k-1}^{(p+1)} = \sum_{j=0}^{s} \left(a_{k+1,j} - a_{k-1,j}\right) c_j^{(p)} = r_k c_{k-1}^{(p)} + (r_k + r_{k+1}) c_k^{(p)} + r_{k+1} c_{k+1}^{(p)}. \tag{39}$$

By combining (38) and (39), we conclude the mathematical induction (36) for $p + 1$

$$\beta_{p+1} := \hat{c}_k^{(p+1)} - c_k^{(p+1)} = \frac{d}{r_k + r_{k+1}}\left(c_{k+1}^{(p+1)} - c_{k-1}^{(p+1)}\right). \tag{40}$$

Finally, (35) and (36) implies the invariant,

$$\begin{aligned} \hat{\mathbf{b}}^T \hat{\mathbf{A}}^p \mathbf{1} - \mathbf{b}^T \mathbf{A}^p \mathbf{1} &= \sum_{i=0}^{s} \left(\hat{b}_i \hat{c}_i^{(p)} - b_i c_i^{(p)}\right) \\ &= (\hat{b}_{k-1} - b_{k-1}) c_{k-1}^{(p)} + b_k (\hat{c}_k^{(p)} - c_k^{(p)}) + (\hat{b}_{k+1} - b_{k+1}) c_{k+1}^{(p)} \\ &= d\left(c_{k-1}^{(p)} - c_{k+1}^{(p)}\right) + (r_k + r_{k+1})\beta_p = 0. \quad \square \end{aligned} \tag{41}$$

The p-th order SMS(R_s) involves the system of polynomial equations $\mathbf{b}^T \mathbf{A}^{q-1} \mathbf{1} = \frac{1}{q!}$ of degree $q = 1, \cdots, p$ with respect to $r_1, r_2, \cdots, r_s$, thus finding R_s by directly solving the system of p-th order polynomial equations is not a trivial task for $p > 2$. We now briefly explain how to find a s-stage SMS coefficients vector $R_s = [r_1, r_2, \cdots, r_s]$ for sixth-order accuracy.

The first-order condition is rather trivial. For $s \geq 1$, given $r_1, \cdots, r_{s-1}$, we choose $r_s = r_s(r_1, \cdots, r_{s-1}) = 1/2 - \sum_{i=1}^{s-1} r_i$, satisfying the first-order condition $\mathbf{b}^T \mathbf{1} = 1$. This is just a simple extension of (28). Then it is also easy to show that SMS(R_s) with $R_s = [r_1, \cdots, r_{s-1}, r_s(r_1, \cdots, r_{s-1})]$ satisfies the second-order conditions, $\mathbf{b}^T \mathbf{A} \mathbf{1} = 1/2$.

Next step is to find last two coefficients r_{s-1} and r_s for given $r_1, \cdots, r_{s-2}$ satisfying both of the first- and third-order conditions corresponding to $\mathbf{b}^T \mathbf{1} = 1$ and $\mathbf{b}^T \mathbf{A}^2 \mathbf{1} = 1/6$, respectively. If we choose r_s to satisfy the first-order condition then the third-order condition is just a polynomial equation for r_{s-1}, which is again an extension of (30). Here we do not show the details but we have an explicit formula for r_{s-1} and r_s as a function of $r_1, \cdots, r_{s-2}$ for $s \geq 3$ and the solution pair is unique up to permutation if it exists. This is similar to (32) which is the case of $s = 3$. A rather lengthy algebraic calculation proves that a third-order SMS(R_s) where $r_{s-1} = r_{s-1}(r_1, \cdots, r_{s-2})$ and $r_s = r_s(r_1, \cdots, r_{s-2})$ also satisfies the fourth-order condition, $\mathbf{b}^T \mathbf{A}^3 \mathbf{1} = 1/24$.

Finally we try to seek an s-stage SMS(R_s) satisfying the first-, third-, and fifth-order conditions. For given $r_1, \cdots, r_{s-3}$ with $s \geq 3$, we form a fifth polynomial equation as a function of $\gamma = r_{s-2}$ to solve $\mathbf{b}^T \mathbf{A}^4 \mathbf{1} = 1/120$ with

$$R_s = [r_1, \cdots, r_{s-3}, \gamma, r_{s-1}, r_s] \tag{42}$$

where $r_{s-1} = r_{s-1}(r_1, \cdots, r_{s-3}, \gamma)$ and $r_s = r_s(r_1, \cdots, r_{s-3}, \gamma)$ are given by the explicit formula for the first- and third-conditions. Though there is no simple algebraic formula on γ, we numerically implement this solver for double precision accuracy and figure out that there is no solution with $s \leq 4$. Fig. 2 shows the dotted plot of (r_1, r_2) where

$$R_5(r_1, r_2) = [r_1, r_2, r_3(r_1, r_2), r_4(r_1, r_2, r_3), r_5(r_1, r_2, r_3)] \tag{43}$$

is a solution for the fifth-order condition. For the numerical observation, we consider a domain for the coefficients as $(r_1, r_2) \in [-1.5, 2.5]^2$ and uniform spacing with 0.05. For examples, numerical values up to single precision accuracy

$$R_5(1, 1.5) \approx [1,\ 1.5,\ -1.474930077,\ -1.109591016,\ 0.584521093], \tag{44}$$

$$R_5(2, 1) \approx [2,\quad 1,\ -1.074758082,\ -1.995305362,\ 0.570063445], \tag{45}$$

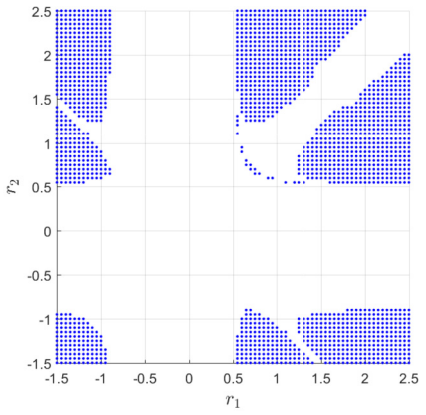

Fig. 2. Pairs of r_1 and r_2 where R_5 has a solution of the fifth-order accuracy.

and their permutations are the solutions satisfying all order conditions up to 5. We also numerically observe that a fifth-order method of SMS(R_5) always satisfies the sixth-order condition, $\mathbf{b}^T\mathbf{A}^5\mathbf{1} = 1/720$.

Remark 2. We have presented the numerical search algorithm for $R_3(\gamma)$ and $R_5(r_1, r_2)$ satisfying $\mathbf{b}^T\mathbf{A}^{q-1}\mathbf{1} = 1/q!$, $q = 1, \cdots, s$ for $s = 3, 5$. Similarly, one can try to search $R_7 = (r_1, r_2, r_3, r_4, r_5(r_1, \cdots, r_4), r_6(r_1, \cdots, r_5), r_7(r_1, \cdots, r_5))$ for given (r_1, r_2, r_3, r_4) satisfying $\mathbf{b}^T\mathbf{A}^{q-1}\mathbf{1} = 1/q!$, $q = 5, 6, 7$ while r_5, r_6, r_7 are chosen to fix $\mathbf{b}^T\mathbf{A}^{q-1}\mathbf{1} = 1/q!$, $q = 1, 2, 3, 4$. Instead of trying a brute force search for the numerical solutions of the system of 3 polynomial equations up to seventh order which may not exist, we will present more identities useful to simplify the system of polynomial equations related to the SMS method in future works.

Before we finish this section, we would like to introduce some properties, which are comparable with existing numerical methods. First, implicit RK methods usually consider the singly diagonally implicit type for the efficient computations, however, this strategy for the SMS methods is not available.

Lemma 5. *There is no a singly diagonal coefficient vector $R_s = [\gamma, \gamma, \cdots, \gamma]$ for higher than second-order SMS methods.*

Proof. Suppose that we have a singly diagonal coefficient vector $R_s = [\gamma, \gamma, \cdots, \gamma]$ and we try to find a specific value of γ to satisfy the order conditions. We denote as $r_1 = r_2 = \cdots = r_s = \gamma$ and we have corresponding coefficients $\mathbf{A}$ and $\mathbf{b}$ as in (25). For the first-order accuracy, we have

$$\mathbf{b}^T\mathbf{1} = 2\sum_{i=1}^{s} r_i = 2s\gamma = 1, \tag{46}$$

and thus $\gamma = 1/(2s)$. For the third-order accuracy, we have

$$\begin{aligned}\mathbf{b}^T\mathbf{A}(\mathbf{A1}) &= 2\gamma\sum_{i=1}^{s-1}\left(\sum_{j=1}^{i-1}4\gamma^2 j + 2\gamma^2 i\right) + \gamma\left(\sum_{j=1}^{s-1}4\gamma^2 j + 2\gamma^2 s\right)\\ &= 4\gamma^3\sum_{i=1}^{s-1} i^2 + 2\gamma^3 s^2 = \frac{2}{3}\gamma^3 s\left(2s^2+1\right) = \frac{2s^2+1}{12s^2} \neq \frac{1}{6}.\end{aligned} \tag{47}$$

Therefore, even the singly diagonal strategy for the SMS method can not surpass the third-order conditions. □

Remark 3. In fact, a SMS method with a singly diagonal coefficient vector $R_s = [\gamma, \gamma, \cdots, \gamma]$ where $\gamma = 1/(2s)$ also is the second-order accurate method, because it always satisfies the second-order condition $\mathbf{b}^T\mathbf{A1} = 1/2$.

Next, we need to remark that the proposed SMS methods differ from symplectic RK methods in [16,17]. It is well known that the algebraic stability condition $b_i b_j - b_i a_{ij} - b_j a_{ij} = 0$ is a sufficient condition for the symplectic RK methods. However, the SMS methods do not satisfy this condition. For the simplicity, we define the algebraic stability matrix as $\mathbf{G} = \mathbf{b}\mathbf{b}^T - \mathbf{B}\mathbf{A} - \mathbf{A}^T\mathbf{B}$, referred to as AG-matrix.

Theorem 6. *A SMS method does not satisfy the algebraic stability condition.*

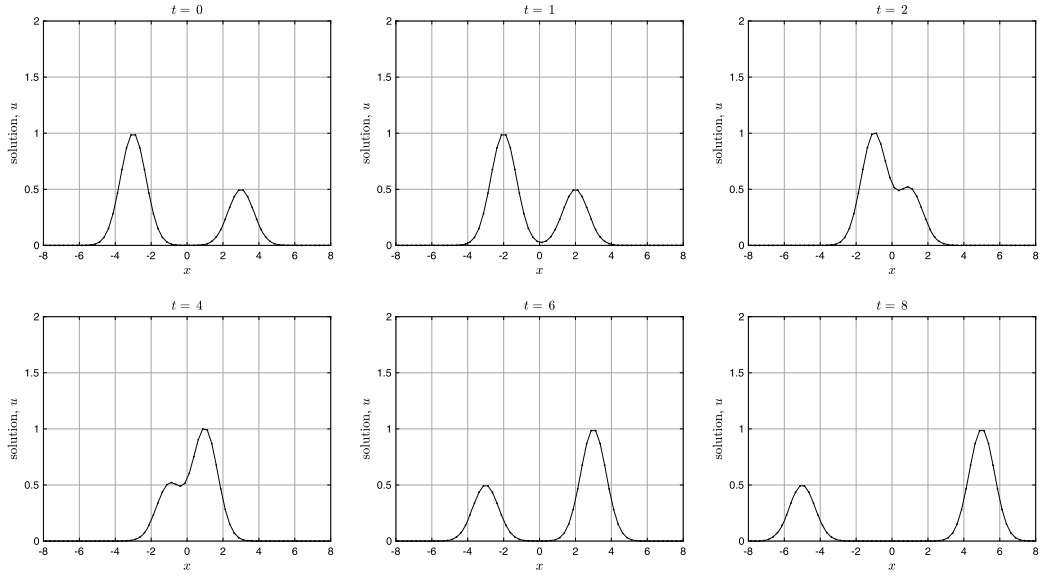

Fig. 3. Time evolution of the solution for the 1D wave equation.

Proof. We consider a coefficient vector $R_s = [r_1, r_2, \cdots, r_s]$ for a SMS method. Without loss of generality, we can suppose that $r_i \neq 0$ for all $i = 1, 2, \ldots, s$. Because $a_{ss} = r_s$ and $b_s = r_s$, we have

$$\mathbf{G}_{ss} = b_s^2 - 2b_s a_{ss} = -r_s^2 \neq 0. \tag{48}$$

Therefore, the SMS method has at least one nonzero entry in AG-matrix **G**, and it means that the algebraic stability condition is not satisfied. □

Remark 4. In fact, the AG-matrix **G** of a SMS method is a diagonal matrix, $G_{ij} = 0$ for $i \neq j$ and $G_{ii} \neq 0$ unless $r_i = r_{i+1}$ for $0 < i < s$.

4. Numerical results in one-dimensional homogeneous medium

In this section, we apply the proposed SMS methods to the one-dimensional scalar wave equation in a homogeneous medium,

$$u_{tt} = u_{xx} \tag{49}$$

to numerically demonstrate the high-order accuracy and energy conservation. We begin by showing the solution of the wave equation (49) with a periodic boundary condition and the following initial conditions:

$$u(x, 0) = e^{-(x+3)^2} + 0.5e^{-(x-3)^2}, \tag{50}$$

$$u_t(x, 0) = 2(x+3)\, e^{-(x+3)^2} - (x-3)\, e^{-(x-3)^2} \tag{51}$$

on a domain $\Omega = [-8, 8]$. The Fourier spectral method is used for spatial derivatives in the numerical computations.

Fig. 3 shows the time evolution of the solution using the fourth-order method, SMS($R_3(0.8)$), with sufficiently small step sizes $\Delta t = T_f/2^{12}$ and $\Delta x = 1/4$. At the earlier stages, two solitary waves move toward each other and merge into one wave. Because the linear waveforms do not interfere, they eventually separate in their original shapes. After finishing this separation, the two solitary solutions move toward both ends at a constant wave speed.

To highlight the energy conservation of the proposed methods, we present numerical results for the second- and fourth-order methods in Sections 4.1 and 4.2, respectively.

4.1. Comparison with classical second-order methods

The leap-frog method $u^{n+1} - 2u^n + u^{n-1} = \Delta t^2 \Delta u^n$ is a well-known multi-step method that can be applied to the scalar wave equation (49). This method achieves second-order accuracy in time, however, there is a time step restriction. The theta method [14] is defined as

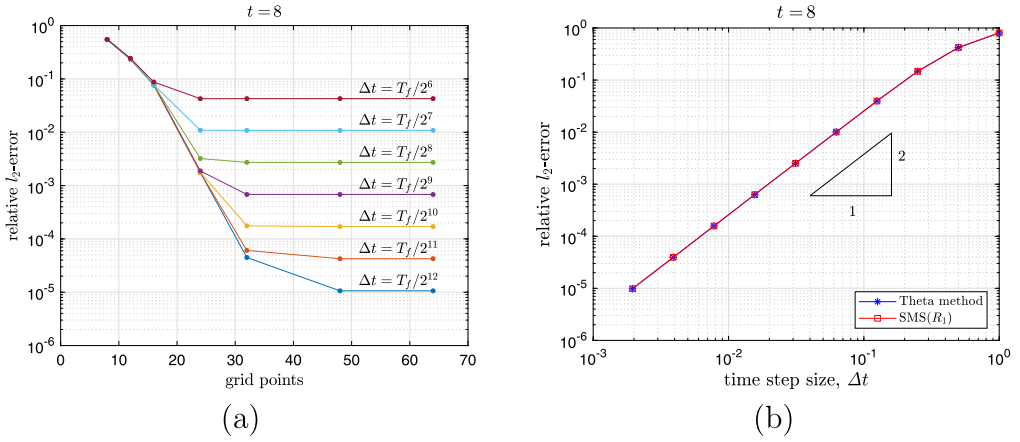

Fig. 4. Relative l_2-norm errors of solutions with second-order methods.

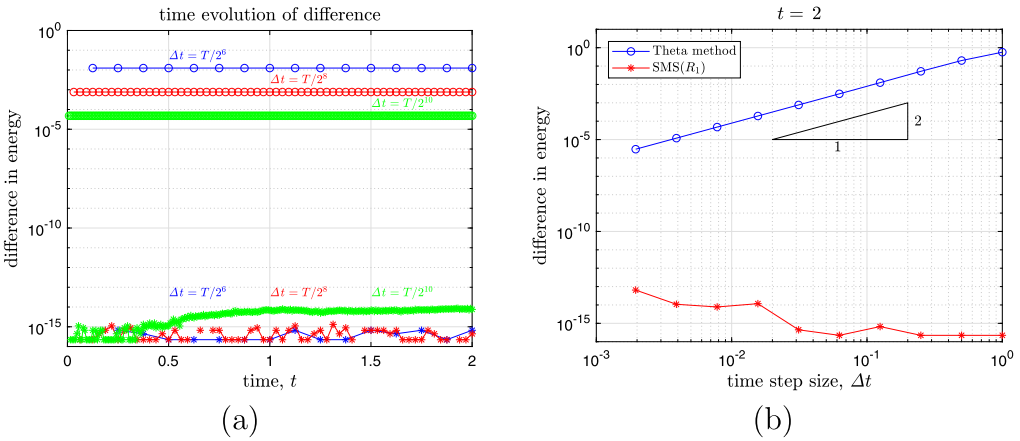

Fig. 5. Absolute difference of the discrete energy with second-order methods.

$$\frac{u^{n+1}-2u^n+u^{n-1}}{\Delta t^2}=\Delta\left(\theta u^{n+1}+(1-2\theta)u^n+\theta u^{n-1}\right), \tag{52}$$

and is a generalization of the leap-frog method (with $\theta=0$). The theta method (52) achieves second-order accuracy. Note that for $\theta\geq 1/4$, energy conservation is guaranteed for the discrete energy defined for the theta method. Specially,

$$\mathcal{E}_\theta\left(u^{n+1},u^n\right)=\frac{1}{2}\left(M_\theta^h\frac{u^{n+1}-u^n}{\Delta t},\frac{u^{n+1}-u^n}{\Delta t}\right)+\frac{1}{2}\left(-\Delta\frac{u^{n+1}+u^n}{2},\frac{u^{n+1}+u^n}{2}\right), \tag{53}$$

where $M_\theta^h=I-\left(\theta-\frac{1}{4}\right)\Delta t^2\Delta$.

We demonstrate the numerical convergence of the theta method and the proposed second-order SMS(R_1) method using the same conditions and parameters as in the beginning of this section. For the convergence result, simulations are performed by varying grid points and time steps. We choose $\theta=1/4$ for the numerical simulation of the theta method.

Fig. 4 shows the numerical results for spatial and temporal convergence. Fig. 4(a) shows the relative l_2-error of $u\,(x,t=8)$ computed by the proposed SMS(R_1) method with respect to various grid points. The Fourier spectral method is used for spatial discretization, meaning that the spectral convergence is well demonstrated and 64 grid points are enough to resolve the numerical solution. Fig. 4(b) shows the relative l_2-error of the solution via the theta and SMS(R_1) methods with respect to various time step sizes $\Delta t=T_f/2^{12},T_f/2^{11},\ldots,T_f/2^3$. It is observed that the methods provide the desired second-order accuracy in time. Here, the error is computed by comparison with a reference solution, which is a numerical solution of the SMS(R_1) method with $\Delta t=T_f/2^{14}$ and $\Delta x=1/8$.

Fig. 5 shows the difference in energy for the theta method (circled line) and the SMS(R_1) method (star line). Fig. 5(a) is the time evolution of energy difference and Fig. 5(b) is the difference of energy at $t=2$ with respect to the various time steps Δt. Conservation of the discrete energy (53) for the theta method is demonstrated, however the energy varies according to the time step Δt. Since $\mathcal{E}_\theta\left(u^1,u^0\right)=E\left(t^0\right)+O\left(\Delta t^2\right)$ and $\mathcal{E}_\theta\left(u^{n+1},u^n\right)=\mathcal{E}_\theta\left(u^n,u^{n-1}\right)$ for all $n>1$, we have

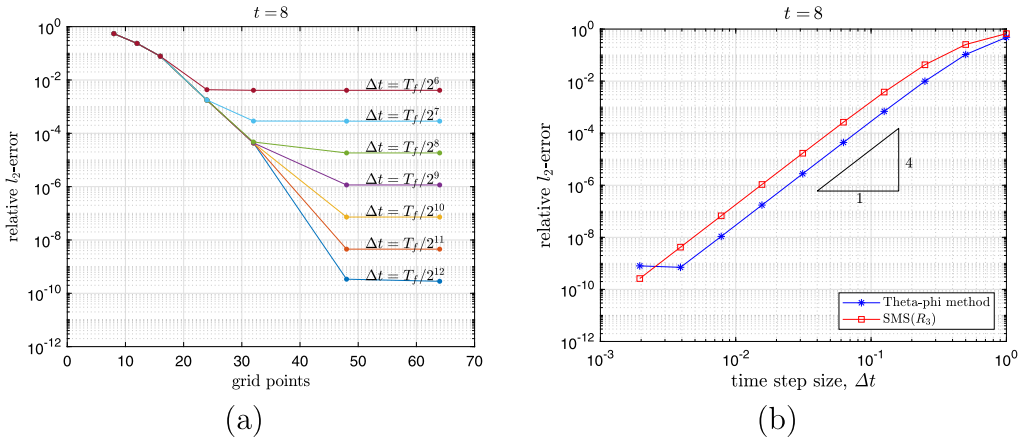

Fig. 6. Relative l_2-norm errors of solutions with fourth-order methods.

$$\mathcal{E}_\theta\left(u^{n+1},u^n\right)=E\left(t^0\right)+O\left(\Delta t^2\right). \tag{54}$$

This is why a second-order convergence rate is shown for the theta method in Fig. 5(b). Meanwhile, the proposed SMS(R_1) method exhibits the energy conservation regardless of the time step size except the machine precision and the round-off calculation errors.

4.2. Comparison with classical fourth-order methods

The theta-phi method [13] for the scalar wave equation (49),

$$\begin{aligned}\frac{u^{n+1}-2u^n+u^{n-1}}{\Delta t^2}&=\Delta\left(\theta u^{n+1}+(1-2\theta)u^n+\theta u^{n-1}\right)\\&\quad-\left(\theta-\frac{1}{12}\right)\Delta t^2\Delta^2\left(\varphi u^{n+1}+(1-2\varphi)u^n+\varphi u^{n-1}\right)\end{aligned} \tag{55}$$

achieves fourth-order accuracy and exhibits discrete energy conservation in appropriate parameters θ and φ. The corresponding discrete energy for the theta-phi method is

$$\begin{aligned}\mathcal{E}_{\theta,\varphi}\left(u^{n+1},u^n\right)=&\frac{1}{2}\left(M^h_{\theta,\varphi}\frac{u^{n+1}-u^n}{\Delta t},\frac{u^{n+1}-u^n}{\Delta t}\right)\\&+\frac{1}{2}\left(K^h_{\theta,\varphi}\frac{u^{n+1}+u^n}{2},\frac{u^{n+1}+u^n}{2}\right),\end{aligned} \tag{56}$$

where

$$M^h_{\theta,\varphi}=I-\left(\theta-\frac{1}{4}\right)\Delta t^2\Delta+\left(\theta-\frac{1}{12}\right)\left(\varphi-\frac{1}{4}\right)\Delta t^4\Delta^2, \tag{57}$$

$$K^h_{\theta,\varphi}=-\Delta+\left(\theta-\frac{1}{12}\Delta t^2\Delta^2\right). \tag{58}$$

We demonstrate numerical convergence using the same conditions and parameters as those used in the beginning of this section. To show the convergence result, simulations are performed by varying grid points and time steps. We choose $\theta=1/4$ and $\varphi=1/4$ for numerical simulation of the theta-phi method.

Fig. 6 shows the numerical convergence results for spatial and temporal accuracy. Fig. 6(a) shows the relative l_2-error of $u(x,t=8)$ with respect to various grid points for the proposed fourth-order SMS(R_3) method and Fig. 6(b) shows the relative l_2-error with respect to various time steps Δt. The error is computed by comparing with a reference solution that is a numerical solution of the SMS(R_3) method with $\Delta t=T_f/2^{14}$ and $\Delta x=1/8$. As in Fig. 4 the spectral convergence is well demonstrated and 64 grid points ($\Delta x=1/4$) are enough to resolve the numerical solution. It is observed that these methods provide the desired fourth-order accuracy.

Fig. 7 shows the difference in energy for the theta-phi method (circled line) and the SMS(R_3) method (star line). Fig. 7 (a) shows the time evolution of this difference and Fig. 7 (b) shows the difference in energy with respect to the various time steps Δt. We can observe that the theta-phi method has a second-order convergence rate regarding the difference in

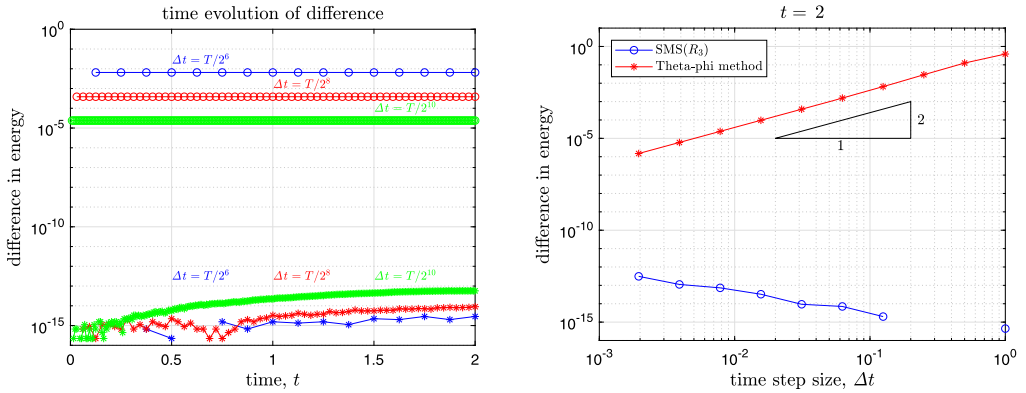

Fig. 7. Absolute errors of the discrete energy with fourth-order methods.

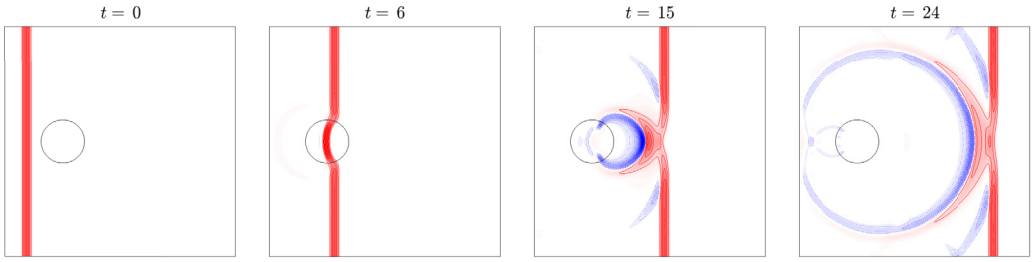

Fig. 8. Time evolution of the solution for the 2D wave equation.

energy. This can be explained by noting the discrete energy (56) is just second-order accurate integration, i.e., $\mathcal{E}_{\theta,\varphi}\left(u^1, u^0\right) = E\left(t^0\right) + O\left(\Delta t^2\right)$. Therefore, in spite of energy conservation, $\mathcal{E}_{\theta,\varphi}\left(u^{n+1}, u^n\right) = \mathcal{E}_{\theta,\varphi}\left(u^n, u^{n-1}\right)$ for all $n > 1$, we have

$$\mathcal{E}_{\theta,\varphi}\left(u^{n+1}, u^n\right) = E\left(t^0\right) + O\left(\Delta t^2\right). \tag{59}$$

Meanwhile, the proposed SMS(R_3) method shows the energy conservation under machine precision and round-off errors.

5. Numerical results in higher space dimensions

To demonstrate the applicability, we use the proposed SMS methods to solve the two-dimensional inhomogeneous wave equation and the three-dimensional scalar wave equation.

5.1. Time evolution in a two-dimensional inhomogeneous media

We now consider the two-dimensional wave equation in an inhomogeneous medium,

$$u_{tt} = \nabla \cdot (a(x, y) \nabla u), \tag{60}$$

with a zero Neumann boundary condition. Here, the variable coefficient is

$$a(x, y) = \frac{3}{4} - \frac{1}{4} \tanh\left(\frac{3 - \sqrt{(x-8)^2 + (y-16)^2}}{0.2}\right), \tag{61}$$

which represents an experimental obstacle for the numerical simulations. The initial states are

$$u(x, y, 0) = e^{-4(x+4)^2}, \tag{62}$$

$$u_t(x, y, 0) = 8(x+4)\, e^{-4(x+4)^2} \tag{63}$$

in the domain $\Omega = [0, 32] \times [0, 32]$. For the numerical simulations, the solution is evolved to time $T_f = 24$.

Fig. 8 shows the time evolution of the solution with sufficiently small step sizes $\Delta t = T_f/2^{11}$ and $\Delta x = 1/8$ using the sixth-order method, SMS(R_5) with $R_5 = R_5(1, 1.5)$. The black circle indicates the contour line of the variable coefficient $a(x, y)$ at the level of $z = 0.75$. In each snapshot, the red and the blue regions indicate $0 < u \leq 1.5$ and $-0.5 \leq u < 0$,

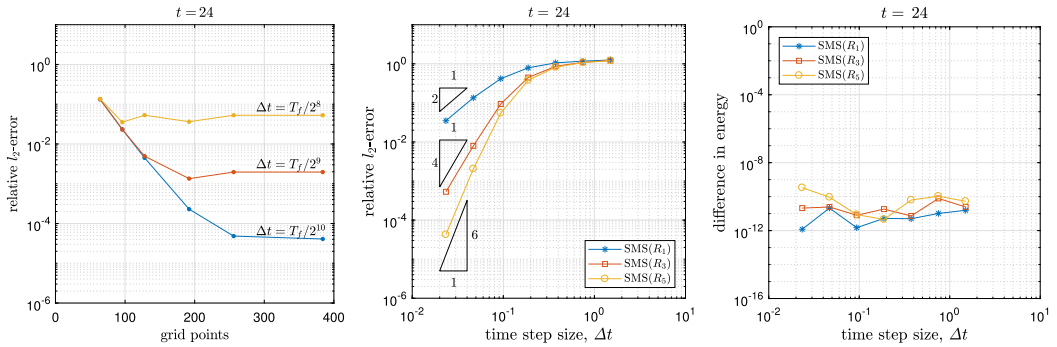

Fig. 9. Relative l_2-errors of the numerical solutions and energy difference in 2D.

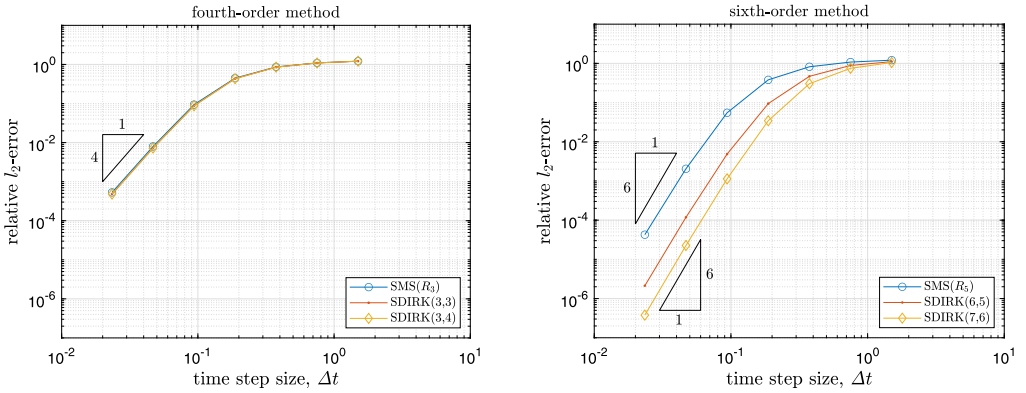

Fig. 10. Relative l_2-norm errors of solutions with fourth- and sixth-order methods.

respectively. To highlight the magnitude, we added red contour lines in 0.2 increments and blue contour lines in 0.05 increments.

Fig. 9 shows the relative l_2-error and energy difference of $u\,(x, y, t = 24)$ with various space steps $\Delta x = 64, 96, 128, \ldots,$ 386 and various time steps $\Delta t = T_f/2^{10}, T_f/2^9, \ldots, T_f/2^4$. Here, the errors are computed by comparison with the reference solution, which is a numerical solution with small step sizes $\Delta t = T_f/2^{12}$ and $\Delta x = 1/16$. It is observed that the spectral accuracy is well demonstrated and the methods give the desired order accuracy both in space and time and inherit the energy conservation.

For the comparison to the existing Symplectic RK methods, we employ the coefficients of the Symplectic Diagonally Implicit RK schemes of q-stages and p-order in [11], referred as in SDIRK(q, p). Note that our proposed SMS methods for fourth- and sixth-order accuracy can be represented as the three- and five-stage methods, respectively.

Fig. 10 shows the numerical results for the temporal convergence of the fourth- and sixth-order methods with a well-resolved step size $\Delta x = 1/8$. It shows the relative l_2-error of the solution with respect to various time step sizes $\Delta t = T_f/2^{10}, T_f/2^9, \ldots, T_f/2^4$. Here, the error is computed through comparison with the reference numerical solution obtained the sixth-order method with $\Delta t = T_f/2^{12}$. We need to remark that, in the case of the linear problem, SDIRK(3, 3) becomes the fourth-order accuracy and SDIRK(6, 5) the sixth-order accuracy. It is observed that the methods provide the desired order of accuracy in time.

5.2. Time evolution in a three-dimensional homogeneous medium

We consider a three-dimensional scalar wave equation

$$u_{tt} = u_{xx} + u_{yy} + u_{zz} \tag{64}$$

in a domain $\Omega = [-8, 8] \times [-8, 8] \times [-8, 8]$ with a periodic boundary condition. The initial states are

$$u\,(x, y, z, 0) = e^{-r^2} \text{ and } u_t\,(x, y, z, 0) = 0, \tag{65}$$

where $r = \sqrt{x^2 + y^2 + z^2}$. The numerical solution is evolved to time $T_f = 4$.

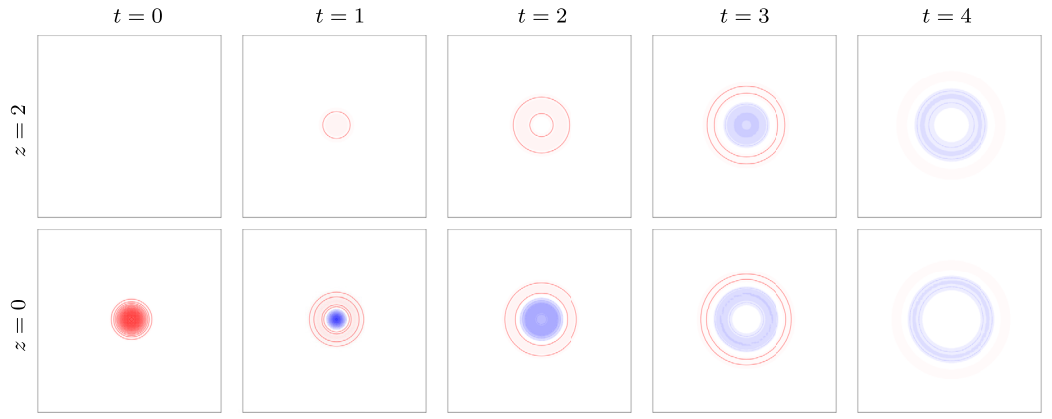

Fig. 11. Time evolution of the solution for the 3D wave equation.

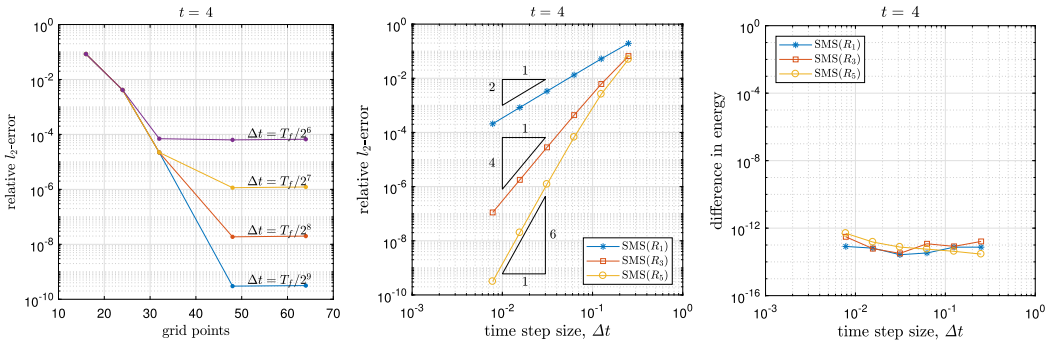

Fig. 12. Relative l_2-errors and energy difference of the numerical 3D solutions.

Fig. 11 shows the time evolution of the solution with sufficiently small step sizes $\Delta t = T_f/2^{10}$ and $\Delta x = 1/4$ using the sixth-order SMS(R_5) method with $R_5(1, 1.5)$. In each snapshot, the red, white, and blue regions indicate $u = 1$, 0, and -1, respectively. To highlight the magnitude, we added red and blue contour lines in 0.05 increments.

Fig. 12 shows the relative l_2-error and the energy difference of $u(x, y, z, t = 4)$ with various space and time steps. Here, the errors are computed by comparison with the reference solution which is a numerical solution with small step sizes $\Delta t = T_f/2^{11}$ and $\Delta x = 1/8$. The spectral accuracy is well demonstrated and 64 grid points ($\Delta x = 1/4$) are enough to resolve the numerical solution. It is observed that these methods provide the desired order accuracy and the energy conservation.

6. Conclusions

We proposed the successive multi-stage (SMS) method for the linear wave equation with the energy conservation and high-order accuracy in time. The SMS methods are generalized from the Crank–Nicolson method, which exhibits energy conservation. For high-order accuracy, we implement the SMS method into the framework of the Runge–Kutta method, in particular, the linear Runge–Kutta method. Note that the proposed SMS methods are different from the symplectic Runge–Kutta method, which is a well-known energy conserving method for wave equations. We provided mathematical proofs of unconditional energy conservation and unique solvability of a semi-discrete scheme. We also numerically demonstrated the accuracy and the energy conservation for various dimensions, including the case of an inhomogeneous medium.

In this paper, we focus on the linear wave equations to emphasize the proposed method is different from well-known existing RK methods but our goal is solving more general class of problems such as quasi-linear or Hamiltonian PDEs. The SMS method could be easily extended to solve the wave equation with forcing terms or quasi-linear wave equations in the form of $u_{tt} = \nabla \cdot (\mathbf{M}\nabla u) + f(u, \mathbf{x}, t)$, $u_{tt} = \nabla \cdot (\mathbf{M}\nabla u) + g(\nabla u, u_t, u, \mathbf{x}, t)$. Such extensions will be reported in other manuscripts [J. Shin, J.-Y. Lee, Energy conserving successive multi-stage method for the linear wave equation with forcing terms, in preparation].

CRediT authorship contribution statement

Jaemin Shin: Conceptualization, Numerical Simulation, and Original draft Writing, **June-Yub Lee:** Formal analysis, Validation, and Reviewing & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Korean Government MSIP (2017R1E1A1A03070161, 2020R1C1C1A01013468).

Appendix A. Linear Runge–Kutta method

For self-consistency, we briefly introduce the linear Runge–Kutta method for a linear system. Compared to the general version of the RK method, we only need compact order conditions because of linearity and autonomy. Let us consider the following initial value problems for the autonomous differential equation with a linear operator $\mathcal{L}$:

$$\frac{\partial \phi}{\partial t} = \mathcal{L}(\phi) \text{ and } \phi(0) = \phi^0. \tag{A.1}$$

Let $\mathcal{L}^{m+1}(\phi) = \mathcal{L}\left(\mathcal{L}^m(\phi)\right)$ for $m \geq 1$, then the Taylor expansion of the exact solution of the system (A.1) can be written as follows using $\mathcal{L}^m(\phi) = \frac{\partial^m \phi}{\partial t^m}$,

$$\phi(h) = \phi_0 + \Delta t \mathcal{L}(\phi_0) + \frac{\Delta t^2}{2!}\mathcal{L}^2(\phi_0) + \cdots + \frac{\Delta t^m}{m!}\mathcal{L}^m(\phi_0) + O\left(\Delta t^{m+1}\right). \tag{A.2}$$

Let ϕ^n be an approximation of $\phi\left(t^n\right)$ of the RK method computing ϕ^{n+1} at $t^{n+1} = t^n + \Delta t$. Starting with $\phi_0 = \phi^n$ and $k_0 = \mathcal{L}(\phi_0)$, we calculate ϕ_i for each stage $i = 1, 2, \ldots, s$,

$$\phi_i = \phi_0 + \Delta t \sum_{j=0}^{s} a_{ij} k_j, \tag{A.3}$$

where a_{ij} are real coefficients and $k_i = \mathcal{L}(\phi_i)$. Finally, we evaluate the next time approximation ϕ^{n+1} as

$$\phi^{n+1} = \phi_0 + \Delta t \sum_{j=0}^{s} b_j k_j, \tag{A.4}$$

where b_j are weights. For the simple description, we denote the coefficients as a matrix $\mathbf{A} \in \mathbb{R}^{(s+1)\times(s+1)}$ and a vector $\mathbf{b} \in \mathbb{R}^{s+1}$. Defining a vector-wise operator evaluation as $\mathcal{L}(\boldsymbol{\phi}) = (\mathcal{L}(\phi_0), \mathcal{L}(\phi_1), \cdots, \mathcal{L}(\phi_s))^T$ with $\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_s)^T$, we can rewrite $k_0, k_1, \ldots, k_s$ as $\mathbf{k} = \mathcal{L}(\phi_0 \mathbf{1} + \Delta t \mathbf{A}\mathbf{k})$. The linear RK method (A.4) can be rewritten as

$$\phi^{n+1} = \phi_0 + \Delta t \mathbf{b} \cdot \mathbf{k}, \tag{A.5}$$

where

$$\mathbf{k} = \mathcal{L}(\phi_0)\mathbf{1} + \Delta t \mathcal{L}^2(\phi_0)\mathbf{A}\mathbf{1} + \cdots + \Delta t^{s-1}\mathcal{L}^s(\phi_0)\mathbf{A}^{s-1}\mathbf{1} + O\left(\Delta t^s\right). \tag{A.6}$$

By equating the coefficients of the elementary differential (A.2) with the Taylor expansion (A.5), we obtain that the order condition for the s-th order accuracy is $\mathbf{b} \cdot \mathbf{A}^{q-1}\mathbf{1} = 1/q!$ for $q = 1, 2, \ldots, s$.

References

[1] B. Gallinet, J. Butet, O.J. Martin, Numerical methods for nanophotonics: standard problems and future challenges, Laser Photonics Rev. 9 (6) (2015) 577–603.
[2] A. Taflove, S.C. Hagness, Computational Electrodynamics: the Finite-Difference Time-Domain Method, Artech House, 2005.
[3] P. Filippi, A. Bergassoli, D. Habault, J.P. Lefebvre, Acoustics: Basic Physics, Theory, and Methods, Elsevier, 1998.
[4] G.C. Cohen, Higher-Order Numerical Methods for Transient Wave Equations, 2003.
[5] D. Komatitsch, F. Coutel, P. Mora, Tensorial formulation of the wave equation for modelling curved interfaces, Geophys. J. Int. 127 (1) (1996) 156–168.
[6] L. Guillot, Y. Capdeville, J.-J. Marigo, 2-d non-periodic homogenization of the elastic wave equation: Sh case, Geophys. J. Int. 182 (3) (2010) 1438–1454.
[7] E. Celledoni, V. Grimm, R.I. McLachlan, D. McLaren, D. O'Neale, B. Owren, G. Quispel, Preserving energy resp. dissipation in numerical pdes using the "average vector field" method, J. Comput. Phys. 231 (20) (2012) 6770–6789.
[8] D. Furihata, Finite-difference schemes for nonlinear wave equation that inherit energy conservation property, J. Comput. Appl. Math. 134 (1–2) (2001) 37–57.
[9] T.J. Bridges, S. Reich, Numerical methods for hamiltonian pdes, J. Phys. A, Math. Gen. 39 (19) (2006) 5287.
[10] L. Brugnano, G.F. Caccia, F. Iavernaro, Energy conservation issues in the numerical solution of the semilinear wave equation, Appl. Math. Comput. 270 (2015) 842–870.

[11] M.A. Sánchez, C. Ciuca, N.C. Nguyen, J. Peraire, B. Cockburn, Symplectic hamiltonian HDG methods for wave propagation phenomena, J. Comput. Phys. 350 (2017) 951–973.
[12] J. Diaz, M.J. Grote, Energy conserving explicit local time stepping for second-order wave equations, SIAM J. Sci. Comput. 31 (3) (2009) 1985–2014.
[13] J. Chabassier, S. Imperiale, Introduction and study of fourth order theta schemes for linear wave equations, J. Comput. Appl. Math. 245 (2013) 194–212.
[14] S. Britt, E. Turkel, S. Tsynkov, A high order compact time/space finite difference scheme for the wave equation with variable speed of sound, J. Sci. Comput. 76 (2) (2018) 777–811.
[15] F. Smith, S. Tsynkov, E. Turkel, Compact high order accurate schemes for the three dimensional wave equation, J. Sci. Comput. 81 (3) (2019) 1181–1209.
[16] J. Sanz-Serna, L. Abia, Order conditions for canonical runge–kutta schemes, SIAM J. Numer. Anal. 28 (4) (1991) 1081–1096.
[17] E. Hairer, C. Lubich, G. Wanner, Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations, Vol. 31, Springer Science & Business Media, 2006.