Discrete and Continuous Dynamical Systems - Series B
Vol. 28, No. 6, June 2023, pp. 3428-3452
doi:10.3934/dcdsb.2022225

EFFICIENT SAV–HERMITE METHODS FOR THE NONLINEAR DIRAC EQUATION

Zhe Yu[1] and Jie Shen[*,2]

[1]School of Science, Harbin Institute of Technology (Shenzhen), Guangdong, 518055, China

[2]Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA

(Communicated by Zhimin Zhang)

Abstract. We construct two SAV/CN–Hermite schemes for one and two dimensional nonlinear Dirac equation in this paper. Both SAV/CN methods preserve the energy and mass and an efficient and accurate scheme is proposed by Hermite approximation. The numerical analysis of error estimates is established on SAV/CN scheme. Numerical experiments confirm the expected convergence order. In addition, they are given to show the simulations of binary and ternary collisions.

1. **Introduction.** The Dirac equation [18] plays an important role in particle physics. As a relativistic wave equation, it can describe all spin–1/2 massive particles for which parity is a symmetry, such as atomic physics, heavy–ion collisions, laser physics, condensed matter physics, and astrophysics [7]. In contrast to the Schrödinger equation which describes the wave function with only one complex value, the wave functions in the Dirac theory are vectors constructed by four complex numbers (known as bispinors). Besides, the Dirac equation degenerates to the Weyl equation in the limit of zero mass.

A main feature of the nonlinear Dirac equation is that it allows solitary wave solutions or particle–like solutions, which are stable local solutions with limited energy and mass [13]. In other words, particles appear in a strong local electric field region, which can be regarded as the basic component in the extended object description of quantum field theory [19]. Moreover, it also appears in the Einstein–Cartan–Sciama–Kibble theory of gravity, which extends general relativity to matter with intrinsic angular momentum [9].

Due to its wide applications, a large effort has been devoted to develop efficient and accurate numerical methods for solving the Dirac equation. For examples, Fillion–Gourdeau et al. [6] developed an unconditionally stable step–splitting numerical method for time–dependent Dirac equation in coordinate space. Xu et al. [20] investigated the behaviour of some standard finite difference schemes and the operator splitting scheme with respect to conservation, reversibility, and accuracy.

2020 *Mathematics Subject Classification.* Primary: 35Q41, 65M70; Secondary: 65M15.

Key words and phrases. Nonlinear Dirac equation; SAV approximation; Spectral method; Error estimate.

The first author is supported by [Guangdong Basic and Applied Basic Research Foundation No. 2020B1515310006 and the China Scholarship Council].

*Corresponding author: Jie Shen.

Recently, Bao et al. [1] proposed three methods for the nonlinear Dirac equation: finite difference time domain method, symmetric exponential wave integrator Fourier pseudospectral method and time–splitting Fourier pseudospectral method, along with error estimates and comparisons.

The scalar auxiliary variable (SAV) approach [16, 17] has been frequently used to construct efficient energy stable schemes dissipative/conservative nonlinear systems. In particular, the schemes based on the SAV approach only requires solving decoupled linear systems with constant coefficients at each time step, while other energy stable schemes require solving either a nonlinear system as with convex splitting approach [5], or coupled linear systems with variable coefficients as with the IEQ approach [22]. Many work have been devoted to the SAV approach for dissipative systems: for examples, Xu and Shen established the convergence of the SAV method for L^2– and H^{-1}–gradient flows; Cheng and Shen [2] developed the multiple SAV approach in order to deal with systems with disparate nonlinearities; Hou et al [10] present a variant of SAV approaches for gradient flows, which leads different convergence orders with different parameters, and Jiang et al. [12] proposed the relaxed SAV approach which could provide improved accuracy. On the other hand, the SAV approach has also been applied to solve conservative nonlinear systems: Antoine et al. developed SAV–Crank–Nicolson schemes for the nonlinear Schrödinger equation (see also [11, 23]), and its error analysis was carried out in [4]; Fu et al. [8] constructed High–order structure–preserving algorithms for the fractional nonlinear Schrödinger equation while Xu et al. [21] developed high–order energy–preserving exponential time differencing method for nonlinear Hamiltonian PDEs, both based on the SAV approach; Shen and Zheng developed SAV schemes for the generalized Zakharov systems.

In this paper, we construct two energy preserving SAV methods for one and two dimensional nonlinear Dirac equation. One is based on a Lagrange multiplier approach [3] which conserves the original energy and mass, the other is based on a penalty approach [24, 3] which conserves mass and a modified energy. The main cost of these two schemes are solving three decoupled linear systems with constant coefficients although an extra nonlinear algebraic system needs to be solved with the former. Since the Dirac equation is set in the whole space, we adapt the Hermite–spectral method in space which enjoys efficient implementation and spectral convergence [15]. Due to the difficulty introduced by the nonlinear algebraic system in the first scheme, we only carry out an error analysis for the second scheme.

The rest of the paper is organized as follows. In Section 2, we derive some mathematical properties of the Dirac equation. In Section 3, we construct two SAV schemes based on Lagrange multiplier and Penalization. We establish error estimates of Penalized MSAV approach in Section 4, and describe the full discretization by using the Hermite spectral method in Section 5. Numerical results are presented in Section 6 followed by a summary of the paper in the last section.

2. **Nonlinear dirac equation.**

We consider the following one and two dimensional Dirac equation [1]

$$\partial_t \Psi = -\left[\sum_{j=1}^{d} \sigma_j \partial_j + \mathrm{i}\gamma\sigma_0\right]\Psi + \mathrm{i}\big(\bar{\Psi}^T\sigma_0\Psi\big)\sigma_0\Psi, \qquad (\mathbf{x}, t) \in \mathbb{R}^d \times (0, T), \tag{1a}$$

$$\Psi(\mathbf{x}, 0) = \Psi_0(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^d, \tag{1b}$$

$$\lim_{|\mathbf{x}|\to\infty} \Psi(\mathbf{x},t) = 0, \qquad t \in (0,T), \tag{1c}$$

where $d = 1, 2$, $\mathrm{i} = \sqrt{-1}$ and

$$\Psi = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}, \quad \bar{\Psi}^T = \begin{pmatrix} \bar{\psi}_1 & \bar{\psi}_2 \end{pmatrix}, \quad \partial_j = \frac{\partial}{\partial x_j},$$
$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \quad \sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The unknown function Ψ is complex–valued, and $\bar{\psi}$ represents the conjugate of ψ. The nonlinear Dirac equation conserves the total mass, i.e.,

$$\|\Psi(\cdot,t)\| \equiv \|\Psi(\cdot,0)\|, \tag{2}$$

and the energy, i.e.,

$$\mathbb{E}[\Psi] = \int_{\mathbb{R}} -\left(\sum_{j=1}^{d} \bar{\Psi}^T \sigma_j \partial_j \Psi + \mathrm{i}\gamma \bar{\Psi}^T \sigma_0 \Psi \right) + \frac{1}{2} \mathrm{i} \left(\bar{\Psi}^T \sigma_0 \Psi \right)^2 \mathrm{d}x \equiv \mathbb{E}[\Psi_0]. \tag{3}$$

For the sake of convenience, we separate the real part and the imaginary part of (1) as follows. Denote $\psi_1 = u_1 + \mathrm{i}u_2$, $\psi_2 = u_3 + \mathrm{i}u_4$, $\mathbf{u} = \begin{pmatrix} u_1 & u_2 & u_3 & u_4 \end{pmatrix}^T$, and set

$$G = \begin{pmatrix} & -1 & & \\ 1 & & & -1 \\ & & 1 & \end{pmatrix}, \quad A = \begin{pmatrix} -1 & & & \\ & -1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix},$$
$$D_1 = \begin{pmatrix} & & & -1 \\ & & 1 & \\ & -1 & & \\ 1 & & & \end{pmatrix}, \quad D_2 = \begin{pmatrix} & & 1 & \\ & & & 1 \\ -1 & & & \\ & -1 & & \end{pmatrix}.$$

We define the inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$ in $L^2(\mathbb{R})^4$ as

$$\left(\mathbf{u},\mathbf{v}\right) = \int_{\mathbb{R}^d} (u_1 v_1 + u_2 v_2 + u_3 v_3 + u_4 v_4) \mathrm{d}\mathbf{x}, \quad \|\mathbf{u}\| = \left(\mathbf{u},\mathbf{u}\right)^{\frac{1}{2}}.$$

Then, the energy functional can be written as

$$\mathcal{E}[\mathbf{u}] := \mathcal{E}_0[\mathbf{u}] + \mathcal{E}_1[\mathbf{u}], \tag{4}$$

where

$$\mathcal{E}_0[\mathbf{u}] = \frac{1}{2} \sum_{j=1}^{d} \left(D_j \partial_j \mathbf{u}, \mathbf{u} \right) + \frac{\gamma}{2} \left(A\mathbf{u}, \mathbf{u} \right), \quad \mathcal{E}_1[\mathbf{u}] = \frac{1}{4} \left\| \mathbf{u}^T A \mathbf{u} \right\|^2. \tag{5}$$

Thanks to the energy conservation, we can rewrite (1) as the following gradient system:

$$\partial_t \mathbf{u} = G\boldsymbol{\mu}, \qquad (\mathbf{x},t) \in \mathbb{R}^d \times (0,T), \tag{6a}$$
$$\boldsymbol{\mu} = \frac{\delta \mathcal{E}}{\delta \mathbf{u}} = \sum_{j=1}^{d} D_j \partial_j \mathbf{u} + \gamma A\mathbf{u} + \frac{\delta \mathcal{E}_1}{\delta \mathbf{u}}, \qquad (\mathbf{x},t) \in \mathbb{R}^d \times (0,T), \tag{6b}$$
$$\mathbf{u}(\mathbf{x},0) = \mathbf{u}^0(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^d, \tag{6c}$$
$$\lim_{|\mathbf{x}|\to\infty} \mathbf{u}(\mathbf{x},t) = \mathbf{0}, \qquad t \in (0,T), \tag{6d}$$

and the energy conservation (3) becomes

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} = \left\langle \frac{\delta \mathcal{E}}{\delta \mathbf{u}}, \frac{\partial \mathbf{u}}{\partial t} \right\rangle = (\boldsymbol{\mu}, G\boldsymbol{\mu}) = 0. \tag{7}$$

Remark 2.1. It is easy to check that the matrices G, D_1, D_2, A satisfy the following properties for any $\mathbf{v}, \mathbf{w} \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$:

$$(G\mathbf{v}, G\mathbf{w}) = (D_1\mathbf{v}, D_1\mathbf{w}) = (D_2\mathbf{v}, D_2\mathbf{w}) = (A\mathbf{v}, A\mathbf{w}) = (\mathbf{v}, \mathbf{w}),$$
$$(G\mathbf{v}, \mathbf{v}) = 0, \quad (G\mathbf{v}, A\mathbf{v}) = 0.$$

3. SAV approach.

3.1. The Lagrange multiplier SAV approach.

The original SAV approach [17] can only conserve a modified energy. In order to conserve both the original energy and mass, we adopt the Lagrange multiplier SAV (LagSAV) approach introduced in [3].

Let us denote $\mathcal{Q}[\mathbf{u}] = \frac{1}{2}\|\mathbf{u}\|^2$, since the Dirac equation is mass and energy conserving, we can introduce two Lagrange multipliers $\eta_1(t)$, $\eta_2(t)$ to expend (6) as

$$\partial_t \mathbf{u} = G\boldsymbol{\mu}, \tag{8a}$$

$$\boldsymbol{\mu} = \frac{\delta \mathcal{E}}{\delta \mathbf{u}} = \sum_{j=1}^{d} D_j \partial_j \mathbf{u} + \gamma A\mathbf{u} + \eta_1(t)\, \frac{\delta \mathcal{E}_1}{\delta \mathbf{u}} + \eta_2(t) \frac{\delta \mathcal{Q}}{\delta \mathbf{u}}, \tag{8b}$$

$$\frac{\mathrm{d}\mathcal{E}_1}{\mathrm{d}t} = \eta_1(t) \left\langle \frac{\delta \mathcal{E}_1}{\delta \mathbf{u}}, \frac{\partial \mathbf{u}}{\partial t} \right\rangle + \eta_2(t) \left\langle \frac{\delta \mathcal{Q}}{\delta \mathbf{u}}, \frac{\partial \mathbf{u}}{\partial t} \right\rangle, \tag{8c}$$

$$\frac{\mathrm{d}\mathcal{Q}}{\mathrm{d}t} = 0. \tag{8d}$$

With $\eta_1(0) = 1$ and $\eta_2(0) = 0$, the above system is equivalent to (6) with $(\eta_1(t), \eta_2(t)) \equiv (1, 0)$.

Instead of (6), we shall construct numerical approximation for (8). More precisely, a second–order scheme (8) based on Crank–Nicolson (CN) is as follows:

$$\mathbf{u}^{n+1} - \mathbf{u}^n = \tau G \boldsymbol{\mu}^{n+\frac{1}{2}}, \tag{9a}$$

$$\begin{aligned} \boldsymbol{\mu}^{n+\frac{1}{2}} = {} & \sum_{j=1}^{d} D_j \partial_j \mathbf{u}^{n+\frac{1}{2}} + \gamma A \mathbf{u}^{n+\frac{1}{2}} \\ & + \eta_1^{n+\frac{1}{2}} \frac{\delta \mathcal{E}_1}{\delta \mathbf{u}}[\mathbf{u}^{*,n+\frac{1}{2}}] + \eta_2^{n+\frac{1}{2}} \frac{\delta \mathcal{Q}}{\delta \mathbf{u}}[\mathbf{u}^{*,n+\frac{1}{2}}], \end{aligned} \tag{9b}$$

$$\begin{aligned} \mathcal{E}_1[\mathbf{u}^{n+1}] - \mathcal{E}_1[\mathbf{u}^n] = {} & \eta_1^{n+\frac{1}{2}} \left(\frac{\delta \mathcal{E}_1}{\delta \mathbf{u}}[\mathbf{u}^{*,n+\frac{1}{2}}], \mathbf{u}^{n+1} - \mathbf{u}^n \right) \\ & + \eta_2^{n+\frac{1}{2}} \left(\frac{\delta \mathcal{Q}}{\delta \mathbf{u}}[\mathbf{u}^{*,n+\frac{1}{2}}], \mathbf{u}^{n+1} - \mathbf{u}^n \right), \end{aligned} \tag{9c}$$

$$\mathcal{Q}[\mathbf{u}^{n+1}] = \mathcal{Q}^0, \tag{9d}$$

where, for any sequence of functions $\{\mathbf{v}^k\}$,

$$\mathbf{v}^{n+\frac{1}{2}} = \frac{1}{2}(\mathbf{v}^{n+1} + \mathbf{v}^n), \quad \mathbf{v}^{*,n+\frac{1}{2}} = \frac{3}{2}\mathbf{v}^n - \frac{1}{2}\mathbf{v}^{n-1}.$$

Theorem 3.1. *The scheme (9) preserves the mass, i.e.,*

$$\mathcal{Q}^n := \mathcal{Q}(\mathbf{u}^n) \equiv \mathcal{Q}^0, \qquad n \geqslant 1, \tag{10}$$

and the energy, i.e.,

$$\mathcal{E}^n := \mathcal{E}[\mathbf{u}^n] \equiv \mathcal{E}^0, \qquad n \geqslant 1. \tag{11}$$

Proof. The equation (9d) guarantees the conservation of mass.

Taking the inner products of (9a) and (9b) respectively with $\boldsymbol{\mu}^{n+\frac{1}{2}}$ and $-(\mathbf{u}^{n+1}-\mathbf{u}^n)$, summing up the results with (9c), we get

$$\begin{aligned}\tau(G\boldsymbol{\mu}^{n+\frac{1}{2}}, \boldsymbol{\mu}^{n+\frac{1}{2}}) = \frac{1}{2}\sum_{j=1}^{d}\left(D_j\partial_j\mathbf{u}^{n+1} + D_j\partial_j\mathbf{u}^n, \mathbf{u}^{n+1}-\mathbf{u}^n\right)\\ + \frac{1}{2}\gamma\left(A\mathbf{u}^{n+1} + A\mathbf{u}^n, \mathbf{u}^{n+1}-\mathbf{u}^n\right) + \mathcal{E}_1^{n+1} - \mathcal{E}_1^n.\end{aligned}$$

Since $(G\boldsymbol{\mu}^{n+\frac{1}{2}}, \boldsymbol{\mu}^{n+\frac{1}{2}}) = 0$, and by integration by parts,

$$\left(D_j\partial_j\mathbf{u}^{n+1}, \mathbf{u}^n\right) = \left(D_j\partial_j\mathbf{u}^n, \mathbf{u}^{n+1}\right), \quad \left(A\mathbf{u}^{n+1}, \mathbf{u}^n\right) = \left(A\mathbf{u}^n, \mathbf{u}^{n+1}\right), \quad j = 1, 2, \tag{12}$$

we derive that

$$\begin{aligned}0 = \tau(G\boldsymbol{\mu}^{n+\frac{1}{2}}, \boldsymbol{\mu}^{n+\frac{1}{2}}) &= \frac{1}{2}\sum_{j=1}^{d}\left(D_j\partial_j\mathbf{u}^{n+1}, \mathbf{u}^{n+1}\right) - \frac{1}{2}\sum_{j=1}^{d}\left(D_j\partial_j\mathbf{u}^n, \mathbf{u}^n\right)\\ &\quad + \frac{1}{2}\gamma\left(A\mathbf{u}^{n+1}, \mathbf{u}^{n+1}\right) - \frac{1}{2}\gamma\left(A\mathbf{u}^n, \mathbf{u}^n\right) + \mathcal{E}_1[\mathbf{u}^{n+1}] - \mathcal{E}_1[\mathbf{u}^n]\\ &= \mathcal{E}^{n+1} - \mathcal{E}^n.\end{aligned}$$

□

We now describe how to efficiently solve the coupled system (9). Writing

$$\mathbf{u}^{n+1} = \mathbf{u}_1^{n+1} + \eta_1^{n+\frac{1}{2}}\mathbf{u}_2^{n+1} + \eta_2^{n+\frac{1}{2}}\mathbf{u}_3^{n+1}, \tag{13}$$

in the first two equations of (9) and collecting the terms without $\eta_i^{n+\frac{1}{2}}$, with $\eta_1^{n+\frac{1}{2}}$ and with $\eta_2^{n+\frac{1}{2}}$ respectively, we obtain the following decoupled systems:

$$\Big(I - \frac{1}{2}\tau\sum_{j=1}^{d}GD_j\partial_j - \frac{1}{2}\tau\gamma GA\Big)\mathbf{u}_1^{n+1} = \Big(I + \frac{1}{2}\tau\sum_{j=1}^{d}GD_j\partial_j + \frac{1}{2}\tau\gamma GA\Big)\mathbf{u}^n, \tag{14a}$$

$$\Big(I - \frac{1}{2}\tau\sum_{j=1}^{d}GD_j\partial_j - \frac{1}{2}\tau\gamma GA\Big)\mathbf{u}_2^{n+1} = \tau G\frac{\delta\mathcal{E}_1}{\delta\mathbf{u}}[\mathbf{u}^{*,n+\frac{1}{2}}], \tag{14b}$$

$$\Big(I - \frac{1}{2}\tau\sum_{j=1}^{d}GD_j\partial_j - \frac{1}{2}\tau\gamma GA\Big)\mathbf{u}_3^{n+1} = \tau G\frac{\delta\mathcal{Q}}{\delta\mathbf{u}}[\mathbf{u}^{*,n+\frac{1}{2}}] = \tau G\mathbf{u}^{*,n+\frac{1}{2}}. \tag{14c}$$

The above systems are all linear differential equations with constant coefficients, and can be efficiently solved by using the spectral–Galerkin method described in Section 5.

Once we obtain $\mathbf{u}_1^{n+1}, \mathbf{u}_2^{n+1}, \mathbf{u}_3^{n+1}$ from the above, we plug (13) into (9c) and (9d) to get

$$\begin{aligned}&\mathcal{E}_1[\mathbf{u}_1^{n+1} + \eta_1^{n+\frac{1}{2}}\mathbf{u}_2^{n+1} + \eta_2^{n+\frac{1}{2}}\mathbf{u}_3^{n+1}] - \mathcal{E}_1[\mathbf{u}^n] \\ &\quad = \left(\eta_1^{n+\frac{1}{2}}\frac{\delta \mathcal{E}_1}{\delta \mathbf{u}}[\mathbf{u}^{*,n+\frac{1}{2}}] + \eta_2^{n+\frac{1}{2}}\mathbf{u}^{*,n+\frac{1}{2}}, \mathbf{u}_1^{n+1} + \eta_1^{n+\frac{1}{2}}\mathbf{u}_2^{n+1} + \eta_2^{n+\frac{1}{2}}\mathbf{u}_3^{n+1} - \mathbf{u}^n\right),\end{aligned} \tag{15}$$

$$\mathcal{Q}[\mathbf{u}_1^{n+1} + \eta_1^{n+\frac{1}{2}}\mathbf{u}_2^{n+1} + \eta_2^{n+\frac{1}{2}}\mathbf{u}_3^{n+1}] = \mathcal{Q}^0, \tag{16}$$

which is nonlinear algebraic system for $\eta_1^{n+\frac{1}{2}}$ and $\eta_2^{n+\frac{1}{2}}$.

Remark 3.2. We observe that the equations (15) and (16) are quartic and quadratic equations whose solution, if it exists, can be obtained using a Newton iteration at a negligible cost. However, it is still an open question whether the nonlinear algebraic equation admits a solution close to the exact solution $(\eta_1, \eta_2) = (1, 0)$, particularly for τ not sufficiently small.

3.2. **Penalized MSAV approach.** The Lagrange multiplier SAV approach in the last section can preserve exactly the mass and original energy, but it leads to a nonlinear algebraic system that may subject to a restrictive time step constraint, or may fail to admit a solution close to the exact solution $(\eta_1, \eta_2) = (1, 0)$. In this section, we present an alternative approach which relaxes the mass conservation by adding a penalization term in the energy, and then introduce two SAVs to deal with the disparity between the nonlinear term and the penalization term which involves a small parameter. This procedure has shown to be very successful for the vesicle membrane problem in [2].

We consider the following penalized total energy:

$$\tilde{\mathcal{E}}[\mathbf{u}] = \mathcal{E}_0[\mathbf{u}] + \mathcal{E}_1[\mathbf{u}] + \mathcal{E}_2[\mathbf{u}], \tag{17}$$

where $\mathcal{E}_0$ and $\mathcal{E}_1$ are given in (5), and $\mathcal{E}_2[\mathbf{u}] = \frac{1}{4\varepsilon}(\|\mathbf{u}\|^2 - \|\mathbf{u}^0\|^2)^2$ is the additional penalization term to approximate the mass conservation.

We introduce two SAVs

$$r_1(t) = \sqrt{\mathcal{E}_1[\mathbf{u}]}, \quad r_2(t) = \|\mathbf{u}\|^2 - \|\mathbf{u}^0\|^2,$$

and rewrite the gradient system (8) with the new total energy in (17) as follows:

$$\partial_t \mathbf{u} = G\boldsymbol{\mu}, \tag{18a}$$

$$\boldsymbol{\mu} = \frac{\delta \tilde{\mathcal{E}}}{\delta \mathbf{u}} = \sum_{j=1}^{d} D_j \partial_j \mathbf{u} + \gamma A\mathbf{u} + \frac{r_1(t)}{\sqrt{\mathcal{E}_1[\mathbf{u}]}}\frac{\delta \mathcal{E}_1}{\delta \mathbf{u}} + \frac{1}{\varepsilon} r_2(t)\mathbf{u}, \tag{18b}$$

$$r_1'(t) = \frac{1}{2\sqrt{\mathcal{E}_1[\mathbf{u}]}}\left\langle \frac{\delta \mathcal{E}_1}{\delta \mathbf{u}}, \frac{\partial \mathbf{u}}{\partial t}\right\rangle, \tag{18c}$$

$$r_2'(t) = 2(\mathbf{u}, \partial_t \mathbf{u}). \tag{18d}$$

Obviously, we have

$$\frac{\mathrm{d}\tilde{\mathcal{E}}}{\mathrm{d}t} = \left\langle \frac{\delta \tilde{\mathcal{E}}}{\delta \mathbf{u}}, \frac{\partial \mathbf{u}}{\partial t}\right\rangle = (\boldsymbol{\mu}, G\boldsymbol{\mu}) = 0.$$

Then, a second–order scheme based on Crank–Nicolson for the above system is

$$\mathbf{u}^{n+1} - \mathbf{u}^n = \tau G\boldsymbol{\mu}^{n+\frac{1}{2}}, \tag{19a}$$

$$\boldsymbol{\mu}^{n+\frac{1}{2}} = \sum_{j=1}^{d} D_j\partial_j \mathbf{u}^{n+\frac{1}{2}} + \gamma A\mathbf{u}^{n+\frac{1}{2}} + \frac{r_1^{n+\frac{1}{2}}}{\sqrt{\mathcal{E}_1[\mathbf{u}^{*,n+\frac{1}{2}}]}}\frac{\delta \mathcal{E}_1}{\delta \mathbf{u}}[\mathbf{u}^{*,n+\frac{1}{2}}] + \frac{1}{\varepsilon}\mathbf{u}^{*,n+\frac{1}{2}} r_2^{n+\frac{1}{2}}, \tag{19b}$$

$$r_1^{n+1} - r_1^n = \frac{1}{2\sqrt{\mathcal{E}_1[\mathbf{u}^{*,n+\frac{1}{2}}]}}\left(\frac{\delta \mathcal{E}_1}{\delta \mathbf{u}}[\mathbf{u}^{*,n+\frac{1}{2}}], \mathbf{u}^{n+1} - \mathbf{u}^n\right), \tag{19c}$$

$$r_2^{n+1} - r_2^n = 2(\mathbf{u}^{*,n+\frac{1}{2}}, \mathbf{u}^{n+1} - \mathbf{u}^n). \tag{19d}$$

Theorem 3.3. *The scheme* (19) *is unconditionally energy stable in the sense that*

$$\tilde{\mathcal{E}}^{n+1} - \tilde{\mathcal{E}}^n = \tau\big(G\boldsymbol{\mu}^{n+\frac{1}{2}}, \boldsymbol{\mu}^{n+\frac{1}{2}}\big) = 0, \tag{20}$$

where the modified energy is defined as

$$\tilde{\mathcal{E}}^n = \frac{1}{2}\sum_{j=1}^{d}(D_j\partial_j\mathbf{u}^n, \mathbf{u}^n) + \frac{\gamma}{2}(A\mathbf{u}^n, \mathbf{u}^n) + (r_1^n)^2 + \frac{1}{4\varepsilon}(r_2^n)^2, \qquad n \geqslant 1. \tag{21}$$

Proof. Taking the inner product of (19a) and (19b) with $\boldsymbol{\mu}^{n+\frac{1}{2}}$ and $-(\mathbf{u}^{n+1} - \mathbf{u}^n)$ respectively, and multiplying (19c) and (19d) by $2r_1^{n+1/2} = r_1^{n+1} + r_1^n$ and $\frac{1}{2\varepsilon}r_2^{n+1/2} = \frac{1}{4\varepsilon}(r_2^{n+1} + r_2^n)$ respectively, summing up the results, we obtain

$$\begin{aligned} 0 = \tau\left(G\boldsymbol{\mu}^{n+\frac{1}{2}}, \boldsymbol{\mu}^{n+\frac{1}{2}}\right) &= \frac{1}{2}\sum_{j=1}^{d}\left(D_j\partial_j\mathbf{u}^{n+1} + D_j\partial_j\mathbf{u}^n, \mathbf{u}^{n+1} - \mathbf{u}^n\right) \\ &+ \frac{1}{2}\gamma\left(A\mathbf{u}^{n+1} + A\mathbf{u}^n, \mathbf{u}^{n+1} - \mathbf{u}^n\right) + (r_1^{n+1})^2 - (r_1^n)^2 + \frac{1}{4\varepsilon}((r_2^{n+1})^2 - (r_2^n)^2). \end{aligned}$$

Thanks to (12), we obtain the desired result. □

We now describe how to efficiently solve the coupled system (19). Writing

$$\mathbf{u}^{n+1} = \mathbf{u}_1^{n+1} + r_1^{n+1}\mathbf{u}_2^{n+1} + r_2^{n+1}\mathbf{u}_3^{n+1}, \tag{22}$$

in (19a) and (19b), and collecting the terms without r_i^{n+1}, with r_1^{n+1} and with r_2^{n+1} respectively, we find that $\mathbf{u}_i^{n+1}$, $(i = 1, 2, 3)$ can be determined from the following decoupled systems:

$$\Big(I - \frac{1}{2}\tau\sum_{j=1}^{d} GD_j\partial_j - \frac{1}{2}\tau\gamma GA\Big)\mathbf{u}_1^{n+1} = \mathbf{b}^n, \tag{23a}$$

$$\Big(I - \frac{1}{2}\tau\sum_{j=1}^{d} GD_j\partial_j - \frac{1}{2}\tau\gamma GA\Big)\mathbf{u}_2^{n+1} = \frac{1}{2}\tau G\mathbf{p}^n, \tag{23b}$$

$$\Big(I - \frac{1}{2}\tau\sum_{j=1}^{d} GD_j\partial_j - \frac{1}{2}\tau\gamma GA\Big)\mathbf{u}_3^{n+1} = \frac{1}{2\varepsilon}\tau G\mathbf{u}^{*,n+\frac{1}{2}}, \tag{23c}$$

where

$$\mathbf{p}^n = \frac{1}{\sqrt{\mathcal{E}_1[\mathbf{u}^{*,n+\frac{1}{2}}]}}\frac{\delta \mathcal{E}_1}{\delta \mathbf{u}}[\mathbf{u}^{*,n+\frac{1}{2}}]$$

$$\mathbf{b}^n = \Big(I + \frac{1}{2}\tau\sum_{j=1}^{d} GD_j\partial_j + \frac{1}{2}\tau\gamma GA\Big)\mathbf{u}^n + \frac{1}{2}\tau r_1^n G\mathbf{p}^n + \frac{1}{2\varepsilon}\tau r_2^n G\mathbf{u}^{*,n+\frac{1}{2}}.$$

The linear systems in (23) can be efficiently solved by using the Hermite–Galerkin method described in Section 5. Once we obtain $\mathbf{u}_1^{n+1}, \mathbf{u}_2^{n+1}, \mathbf{u}_3^{n+1}$, we can plug (22) into (19c) and (19d) to obtain the following 2×2 linear algebraic system for r_1^{n+1} and r_2^{n+1}:

$$\Big(1-\frac{1}{2}(\mathbf{p}^n,\mathbf{u}_2^{n+1})\Big)r_1^{n+1}-\frac{1}{2}(\mathbf{p}^n,\mathbf{u}_3^{n+1})r_2^{n+1}=\frac{1}{2}(\mathbf{p}^n,\mathbf{u}_1^{n+1}-\mathbf{u}^n), \tag{24a}$$

$$-2(\mathbf{u}^{*,n+\frac{1}{2}},\mathbf{u}_2^{n+1})r_1^{n+1}+\big(1-2(\mathbf{u}^{*,n+\frac{1}{2}},\mathbf{u}_2^{n+1})\big)r_2^{n+1}=2(\mathbf{u}^{*,n+\frac{1}{2}},\mathbf{u}_1^{n+1}-\mathbf{u}^n). \tag{24b}$$

Hence, the system (19) can be efficiently solved.

4. **Error estimate.** In this section, we establish an error estimate for the scheme (19). Let $\mathbf{u}(t^n)$, $\boldsymbol{\mu}(t^n)$, $r_1(t^n)$, $r_2(t^n)$ be the exact solution of (18) at time t^n, and $\mathbf{u}^n$, $\boldsymbol{\mu}^n$, r_1^n, r_2^n be the solution of the scheme (19). Denote

$$\mathbf{e}^n=\mathbf{u}^n-\mathbf{u}(t^n),\ \boldsymbol{\omega}^n=\boldsymbol{\mu}^n-\boldsymbol{\mu}(t^n),\ s_1^n=r_1^n-r_1(t^n),\ s_2^n=r_2^n-r_2(t^n).$$

Setting

$$\mathscr{L}=\sum_{j=1}^d D_j\partial_j+\gamma A,\quad \mathbf{p}[\mathbf{u}]=\frac{1}{\sqrt{\mathcal{E}_1[\mathbf{u}]}}\frac{\delta\mathcal{E}_1[\mathbf{u}]}{\delta\mathbf{u}}, \tag{25}$$

and subtracting (18) from (19), we obtain

$$\mathbf{e}^{n+1}-\mathbf{e}^n=\tau G\boldsymbol{\omega}^{n+\frac{1}{2}}-T_1^n[\mathbf{u}], \tag{26a}$$

$$\begin{aligned}\boldsymbol{\omega}^{n+\frac{1}{2}}=&\frac{1}{2}\mathscr{L}(\mathbf{e}^{n+1}+\mathbf{e}^n)+\mathscr{L}T_2^n[\mathbf{u}]+s_1^{n+\frac{1}{2}}\mathbf{p}^n+T_2^n[r_1]\mathbf{p}^n+r_1(t^{n+\frac{1}{2}})T_3^n\\&+\frac{1}{\varepsilon}\big(s_2^{n+\frac{1}{2}}\mathbf{u}^{*,n+\frac{1}{2}}+T_2^n[r_2]\mathbf{u}^{*,n+\frac{1}{2}}+r_2(t^{n+\frac{1}{2}})T_4^n\big),\end{aligned} \tag{26b}$$

$$s_1^{n+1}-s_1^n=\frac{1}{2}(\mathbf{p}^n,\mathbf{e}^{n+1}-\mathbf{e}^n)+\frac{1}{2}(\mathbf{p}^n,T_1^n[\mathbf{u}])+\frac{1}{2}\Big(T_3^n,\tau\frac{\partial\mathbf{u}}{\partial t}(t^{n+\frac{1}{2}})\Big)-T_1^n[r_1], \tag{26c}$$

$$\begin{aligned}s_2^{n+1}-s_2^n=&2(\mathbf{u}^{*,n+\frac{1}{2}},\mathbf{e}^{n+1}-\mathbf{e}^n)+2(\mathbf{u}^{*,n+\frac{1}{2}},T_1^n[\mathbf{u}])\\&+2\Big(T_4^n,\tau\frac{\partial\mathbf{u}}{\partial t}(t^{n+\frac{1}{2}})\Big)-T_1^n[r_2],\end{aligned} \tag{26d}$$

where for any function ϕ,

$$T_1^n[\phi]=\phi(t^{n+1})-\phi(t^n)-\tau\partial_t\phi(t^{n+\frac{1}{2}}),$$
$$T_2^n[\phi]=\frac{1}{2}\phi(t^{n+1})-\phi(t^{n+\frac{1}{2}})+\frac{1}{2}\phi(t^n),$$

and

$$T_3^n=\mathbf{p}^n-\mathbf{p}[\mathbf{u}(t^{n+\frac{1}{2}})],\quad T_4^n=\mathbf{u}^{*,n+\frac{1}{2}}-\mathbf{u}(t^{n+\frac{1}{2}}).$$

Lemma 4.1. *For $\mathbf{v},\mathbf{w}\in H^1(\mathbb{R}^d)$, we have*

$$(\mathbf{v},\mathscr{L}\mathbf{w})=(\mathscr{L}\mathbf{v},\mathbf{w}), \tag{27a}$$

$$(G\mathbf{v},\mathscr{L}\mathbf{v})=0. \tag{27b}$$

Proof. First, it is obvious that

$$(\mathbf{v}, A\mathbf{w}) = (A\mathbf{v}, \mathbf{w}).$$

Since $\mathbf{v}, \mathbf{w} \in H^1(\mathbb{R}^d)$, we have $\lim_{|\mathbf{x}|\to\infty} \mathbf{v}(\mathbf{x}) = 0$, $\lim_{|\mathbf{x}|\to\infty} \mathbf{w}(\mathbf{x}) = 0$. Using integration by parts with boundary conditions, we obtain

$$\begin{aligned}(\mathbf{v}, D_1\partial_1\mathbf{w}) &= \int_{\mathbb{R}^d} (v_1\partial_1 w_4 - v_2\partial_1 w_3 + v_3\partial_1 w_2 - v_4\partial_1 w_1)\mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^d} (-\partial_1 v_1 w_4 + \partial_1 v_2 w_3 - \partial_1 v_3 w_2 + \partial_1 v_4 w_1)\mathrm{d}\mathbf{x} = (D_1\partial_1\mathbf{v}, \mathbf{w});\end{aligned}$$

$$\begin{aligned}(\mathbf{v}, D_2\partial_2\mathbf{w}) &= \int_{\mathbb{R}^d} (v_1\partial_2 w_3 + v_2\partial_2 w_4 - v_3\partial_2 w_1 - v_4\partial_2 w_2)\mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^d} (\partial_2 v_3 w_1 + \partial_2 v_4 w_2 - \partial_2 v_1 w_3 - \partial_2 v_2 w_4)\mathrm{d}\mathbf{x} = (D_2\partial_2\mathbf{v}, \mathbf{w});\end{aligned}$$

$$\begin{aligned}(G\mathbf{v}, D_1\partial_1\mathbf{v}) &= \int_{\mathbb{R}^d} (-v_2\partial_1 v_4 - v_1\partial_1 v_3 - v_4\partial_1 v_2 - v_3\partial_1 v_1)\mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^d} \partial_1(-v_1 v_3 - v_2 v_4)\mathrm{d}\mathbf{x} = 0;\end{aligned}$$

$$\begin{aligned}(G\mathbf{v}, D_2\partial_2\mathbf{v}) &= \int_{\mathbb{R}^d} (-v_2\partial_2 v_3 + v_1\partial_2 v_4 + v_4\partial_2 v_1 - v_3\partial_2 v_2)\mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^d} \partial_2(-v_2 v_3 + v_1 v_4)\mathrm{d}\mathbf{x} = 0.\end{aligned}$$

Then (27) hold thanks to the definition of $\mathscr{L}$ in (25) and Remark 2.1. □

Theorem 4.2. *Let* $\mathbf{u}$ *be the exact solution of* (19)*, we assume that for each* $t \in (0, T)$*,*

$$\begin{aligned}&\mathbf{u}(\cdot) \in H^1(\mathbb{R}^d) \cap W^{1,\infty}(\mathbb{R}^d),\ \partial_t\mathbf{u}(\cdot) \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d), \\ &\partial_{tt}\mathbf{u}(\cdot) \in H^2(\mathbb{R}^d),\ \partial_{ttt}\mathbf{u}(\cdot) \in H^1(\mathbb{R}^d).\end{aligned} \tag{28}$$

Then we have for all $n < T/\tau$*,*

$$\begin{aligned}&\frac{1}{2}\big(\|\nabla\mathbf{e}^n\|^2 + \gamma^2\|\mathbf{e}^n\|^2\big) + (s_1^n)^2 + \frac{1}{4\varepsilon}(s_2^n)^2 \\ &\quad\lesssim \tau^4 \exp\big((1-2\tau)^{-1}t^n\big) \int_0^{t^n} \Big(\big\|\partial_{ttt}\mathbf{u}(s)\big\|^2_{H^1(\mathbb{R}^d)} \\ &\quad + \big\|\partial_{tt}\mathbf{u}(s)\big\|^2_{H^2(\mathbb{R}^d)} + \big\|r_1'''(s)\big\|^2 + \frac{1}{\varepsilon^2}\big\|r_2'''(s)\big\|^2 + \big\|r_1''(s)\big\|^2 + \frac{1}{\varepsilon^2}\big\|r_2''(s)\big\|^2\Big)\mathrm{d}s.\end{aligned}$$

Here, $A \lesssim B$ *presents* $A \leqslant cB$ *where* c *is a positive constant independent of* τ*.*

Proof. By the assumption, there exists $C > 0$ such that for all $t \in (0, T)$, we obtain that

$$\begin{aligned}&\|\mathbf{u}\|_{W^{1,\infty}(\mathbb{R}^d)},\ \|\mathbf{u}\|_{H^1(\mathbb{R}^d)},\ \|\partial_t\mathbf{u}\|,\ \|\partial_t\mathbf{u}\|_{L^\infty(\mathbb{R}^d)}, \\ &\|\partial_{tt}\mathbf{u}\|_{H^1(\mathbb{R}^d)},\ \|\partial_{tt}\mathbf{u}\|_{H^2(\mathbb{R}^d)},\ \|\partial_{ttt}\mathbf{u}\|_{H^1(\mathbb{R}^d)} \leqslant C.\end{aligned}$$

Setting $\mathbf{U} = (\mathbf{u}^T A\mathbf{u})A\mathbf{u}$, we can derive the following

$$\begin{aligned}
&\|\mathbf{U}\| \leqslant \left\|(\mathbf{u}^T\mathbf{u})\mathbf{u}\right\| \leqslant C\|\mathbf{u}\|^2_{L^\infty(\mathbb{R}^d)}\|\mathbf{u}\|,\\
&\|\mathbf{U}\|_{L^\infty(\mathbb{R}^d)} \leqslant C\|\mathbf{u}\|^3_{L^\infty(\mathbb{R}^d)}, \quad \left\|\frac{\delta\mathbf{U}}{\delta\mathbf{u}}\right\|_{L^\infty(\mathbb{R}^d)} \leqslant C\|\mathbf{u}\|^2_{L^\infty(\mathbb{R}^d)}.
\end{aligned}$$

We can also derive from the above that

$$\|\mathbf{p}\| \leqslant \left\|(\mathbf{u}^T A\mathbf{u})A\mathbf{u}\right\| \leqslant C\|\mathbf{u}\|^2_{L^\infty(\mathbb{R}^d)}\|\mathbf{u}\|, \quad \|\nabla\mathbf{p}\| \leqslant C\|\mathbf{u}\|^2_{L^\infty(\mathbb{R}^d)}\|\nabla\mathbf{u}\|.$$

On the other hand, we can also derive the following properties for $r_1(t)$, $r_2(t)$:

$$\begin{aligned}
r_1'(t) &= \frac{1}{2}r_1^{-1}(\mathbf{u}^T A\mathbf{u}, \mathbf{u}^T A\partial_t\mathbf{u}),\\
|r_1'(t)| &\leqslant \left\|\mathbf{u}^T A\partial_t\mathbf{u}\right\| \leqslant \|\mathbf{u}\|_{L^\infty(\mathbb{R}^d)}\|\partial_t\mathbf{u}\|,\\
r_1''(t) &= -\frac{1}{4}r_1^{-3}(\mathbf{u}^T A\mathbf{u}, \mathbf{u}^T A\partial_t\mathbf{u})^2 + \frac{1}{2}r_1^{-1}(\mathbf{u}^T A\mathbf{u}, \partial_t\mathbf{u}^T A\partial_t\mathbf{u})\\
&\quad + r_1^{-1}\|\mathbf{u}^T A\partial_t\mathbf{u}\|^2 + \frac{1}{2}r_1^{-1}(\mathbf{u}^T A\mathbf{u}, \mathbf{u}^T A\partial_{tt}\mathbf{u})\\
|r_1''(t)| &\leqslant \frac{1}{4}\left\|\mathbf{u}^T A\partial_t\mathbf{u}\right\|^2 + \frac{1}{2}\left\|\partial_t\mathbf{u}^T A\partial_t\mathbf{u}\right\| + \frac{1}{2}\left\|\mathbf{u}^T A\partial_{tt}\mathbf{u}\right\|\\
&\lesssim \|\mathbf{u}\|^2_{L^\infty(\mathbb{R}^d)}\|\partial_t\mathbf{u}\|^2 + \|\partial_t\mathbf{u}\|_{L^\infty(\mathbb{R}^d)}\|\partial_t\mathbf{u}\| + \|\mathbf{u}\|_{L^\infty(\mathbb{R}^d)}\|\partial_{tt}\mathbf{u}\|,\\
r_1'''(t) &= \frac{3}{8}r_1^{-5}(\mathbf{u}^T A\mathbf{u}, \mathbf{u}^T A\partial_t\mathbf{u})^3 - \frac{3}{2}r_1^{-3}\|\mathbf{u}^T A\partial_t\mathbf{u}\|^2(\mathbf{u}^T A\mathbf{u}, \mathbf{u}^T A\partial_t\mathbf{u})\\
&\quad - \frac{1}{2}r_1^{-3}(\mathbf{u}^T A\mathbf{u}, \mathbf{u}^T A\partial_t\mathbf{u})(\mathbf{u}^T A\mathbf{u}, \partial_t\mathbf{u}^T A\partial_t\mathbf{u})\\
&\quad - \frac{1}{2}r_1^{-3}(\mathbf{u}^T A\mathbf{u}, \mathbf{u}^T A\partial_t\mathbf{u})(\mathbf{u}^T A\mathbf{u}, \mathbf{u}^T A\partial_{tt}\mathbf{u})\\
&\quad + r_1^{-1}(\mathbf{u}^T A\partial_t\mathbf{u}, \partial_t\mathbf{u}^T A\partial_t\mathbf{u}) + 3r_1^{-1}(\mathbf{u}^T A\partial_t\mathbf{u}, \mathbf{u}^T A\partial_{tt}\mathbf{u})\\
&\quad + \frac{3}{2}r_1^{-1}(\mathbf{u}^T A\mathbf{u}, \partial_t\mathbf{u}^T A\partial_{tt}\mathbf{u}) + \frac{1}{2}r_1^{-1}(\mathbf{u}^T A\mathbf{u}, \partial_t\mathbf{u}^T A\partial_{ttt}\mathbf{u})\\
&\quad + 2r_1^{-1}(\mathbf{u}^T A\mathbf{u}, \partial_t\mathbf{u}^T A\partial_t\mathbf{u}),\\
|r_1'''(t)| &\leqslant \frac{15}{8}\left\|\mathbf{u}^T A\partial_t\mathbf{u}\right\|^3 + \frac{3}{2}\left\|\mathbf{u}^T A\partial_t\mathbf{u}\right\|\cdot\left\|\partial_t\mathbf{u}^T A\partial_t\mathbf{u}\right\|\\
&\quad + \frac{7}{2}\left\|\mathbf{u}^T A\partial_t\mathbf{u}\right\|\cdot\left\|\mathbf{u}^T A\partial_{tt}\mathbf{u}\right\| + \frac{3}{2}\left\|\partial_t\mathbf{u}^T A\partial_{tt}\mathbf{u}\right\|\\
&\quad + \frac{1}{2}\left\|\partial_t\mathbf{u}^T A\partial_{ttt}\mathbf{u}\right\| + 2\left\|\partial_t\mathbf{u}^T A\partial_t\mathbf{u}\right\|\\
&\lesssim \|\mathbf{u}\|^3_{L^\infty(\mathbb{R}^d)}\|\partial_t\mathbf{u}\|^3 + \|\mathbf{u}\|_{L^\infty(\mathbb{R}^d)}\|\partial_t\mathbf{u}\|_{L^\infty(\mathbb{R}^d)}\|\partial_t\mathbf{u}\|^2\\
&\quad + \|\mathbf{u}\|^2_{L^\infty(\mathbb{R}^d)}\|\partial_t\mathbf{u}\|\cdot\|\partial_{tt}\mathbf{u}\| + \|\partial_t\mathbf{u}\|_{L^\infty(\mathbb{R}^d)}\|\partial_{tt}\mathbf{u}\|\\
&\quad + \|\partial_t\mathbf{u}\|_{L^\infty(\mathbb{R}^d)}\|\partial_t\mathbf{u}\| + \|\partial_t\mathbf{u}\|_{L^\infty(\mathbb{R}^d)}\|\partial_{ttt}\mathbf{u}\|,\\
r_2'(t) &= (\mathbf{u}, \partial_t\mathbf{u}), \qquad |r_2'(t)| \leqslant \|\mathbf{u}\|\cdot\|\partial_t\mathbf{u}\|,\\
r_2''(t) &= (\mathbf{u}, \partial_{tt}\mathbf{u}) + \|\partial_t\mathbf{u}\|^2, \qquad |r_2''(t)| \leqslant \|\mathbf{u}\|\cdot\|\partial_{tt}\mathbf{u}\| + \|\partial_t\mathbf{u}\|^2,\\
r_2'''(t) &= (\mathbf{u}, \partial_{ttt}\mathbf{u}) + 3(\partial_t\mathbf{u}, \partial_{tt}\mathbf{u}), \ |r_2'''(t)| \leqslant \|\mathbf{u}\|\cdot\|\partial_{ttt}\mathbf{u}\| + 3\|\partial_t\mathbf{u}\|\cdot\|\partial_{tt}\mathbf{u}\|.
\end{aligned}$$

Taking the inner products of (26a) and (26b) with $\mathscr{L}\boldsymbol{\omega}^{n+\frac{1}{2}}$ and $-\mathscr{L}(\mathbf{e}^{n+1}-\mathbf{e}^n)$ respectively, and summing up the results, thanks to (27a) and (27b), we obtain

$$\begin{aligned}\frac{1}{2}\big(\|\nabla\mathbf{e}^{n+1}\|^2+\gamma^2\|\mathbf{e}^{n+1}\|^2-\|\nabla\mathbf{e}^n\|^2-\gamma^2\|\mathbf{e}^n\|^2\big)+(s_1^{n+1})^2+\frac{1}{4\varepsilon}(s_2^{n+1})^2\\ -(s_1^n)^2-\frac{1}{4\varepsilon}(s_2^n)^2=I_1+I_2+I_3+I_4+\frac{1}{\varepsilon}(I_5+I_6+I_7),\end{aligned}\tag{29}$$

where, with the two SAVs $r_1(t)\equiv 1$ and $r_2(t)\equiv 0$ in the continuous case for $t\in(0,T)$, there hold

$$\begin{aligned}I_1&=-\langle\mathscr{L}T_1^n[\mathbf{u}],\boldsymbol{\omega}^{n+\frac{1}{2}}\rangle-\tau\langle\mathscr{L}^2T_2^n[\mathbf{u}],G\boldsymbol{\omega}^{n+\frac{1}{2}}\rangle+(\mathscr{L}T_2^n[\mathbf{u}],\mathscr{L}T_1^n[\mathbf{u}]),\\ I_2&=\tau s_1^{n+\frac{1}{2}}\langle\mathbf{p}^n-\mathscr{L}\mathbf{p}^n,G\boldsymbol{\omega}^{n+\frac{1}{2}}\rangle,\\ I_3&=s_1^{n+\frac{1}{2}}(\mathscr{L}\mathbf{p}^n,T_1^n[\mathbf{u}])+\tau s_1^{n+\frac{1}{2}}\Big(T_3^n,\frac{\partial\mathbf{u}}{\partial t}(t^{n+\frac{1}{2}})\Big)-2s_1^{n+\frac{1}{2}}T_1^n[r_1],\\ I_4&=-\tau T_2^n[r_1](\mathscr{L}\mathbf{p}^n,G\boldsymbol{\omega}^{n+\frac{1}{2}})+T_2^n[r_1](\mathscr{L}\mathbf{p}^n,T_1^n[\mathbf{u}])\\ &\quad-\tau(\mathscr{L}T_3^n,G\boldsymbol{\omega}^{n+\frac{1}{2}})+(\mathscr{L}T_3^n,T_1^n[\mathbf{u}]),\\ I_5&=\tau s_2^{n+\frac{1}{2}}(\mathbf{u}^{*,n+\frac{1}{2}}-\mathscr{L}\mathbf{u}^{*,n+\frac{1}{2}},G\boldsymbol{\omega}^{n+\frac{1}{2}})\\ I_6&=s_2^{n+\frac{1}{2}}(\mathscr{L}\mathbf{u}^{*,n+\frac{1}{2}},T_1^n[\mathbf{u}])+\tau s_2^{n+\frac{1}{2}}\Big(T_4^n,\frac{\partial\mathbf{u}}{\partial t}(t^{n+\frac{1}{2}})\Big)-\frac{1}{2}s_2^{n+\frac{1}{2}}T_1^n[r_2]\\ I_7&=-\tau T_2^n[r_2](\mathscr{L}\mathbf{u}^{*,n+\frac{1}{2}},G\boldsymbol{\omega}^{n+\frac{1}{2}})+T_2^n[r_2](\mathscr{L}\mathbf{u}^{*,n+\frac{1}{2}},T_1^n[\mathbf{u}]).\end{aligned}$$

We need to bound the terms on the righthand sides of the above.

To start with estimation, we first present and estimate T_1^n, T_2^n, T_3^n, T_4^n as follows.

$$\begin{aligned}T_1^n[\mathbf{u}]&=\mathbf{u}(x,t^{n+1})-\mathbf{u}(x,t^n)-\tau\partial_t\mathbf{u}(x,t^{n+\frac{1}{2}})\\ &=\frac{1}{2}\int_{t^n}^{t^{n+1}}\Big(\frac{1}{2}\tau-|t^{n+\frac{1}{2}}-s|\Big)^2\partial_{ttt}\mathbf{u}(x,s)\mathrm{d}s.\end{aligned}\tag{30}$$

Hence,

$$\begin{aligned}\|T_1^n[\mathbf{u}]\|^2&=\frac{1}{4}\int_{\mathbb{R}}\bigg(\int_{t^n}^{t^{n+1}}\Big(\frac{1}{2}\tau-|t^{n+\frac{1}{2}}-s|\Big)^2\partial_{ttt}\mathbf{u}(x,s)\mathrm{d}s\bigg)^2\mathrm{d}x\\ &\leqslant\frac{1}{4}\int_{\mathbb{R}}\bigg(\int_{t^n}^{t^{n+1}}\Big(\frac{1}{2}\tau-|t^{n+\frac{1}{2}}-s|\Big)^4\mathrm{d}s\cdot\int_{t^n}^{t^{n+1}}|\partial_{ttt}\mathbf{u}(x,s)|^2\mathrm{d}s\bigg)\mathrm{d}x\\ &=\frac{1}{320}\tau^5\int_{t^n}^{t^{n+1}}\|\partial_{ttt}\mathbf{u}(s)\|^2\mathrm{d}s,\end{aligned}\tag{31}$$

and

$$\begin{aligned}\big\|\mathscr{L}T_1^n[\mathbf{u}]\big\|^2&\leqslant 2\big(\big\|T_1^n[\mathbf{u}]\big\|^2+\big\|\nabla T_1^n[\mathbf{u}]\big\|^2\big)\\ &\leqslant\frac{1}{160}\tau^5\int_{t^n}^{t^{n+1}}\big(\big\|\partial_{ttt}\mathbf{u}(s)\big\|^2+\big\|\partial_{ttt}\nabla\mathbf{u}(s)\big\|^2\big)\mathrm{d}s.\end{aligned}\tag{32}$$

Similarly, we can show that

$$T_2^n[\mathbf{u}]=\frac{1}{2}\mathbf{u}(t^{n+1})-\mathbf{u}(t^{n+\frac{1}{2}})+\frac{1}{2}\mathbf{u}(t^n)=\frac{1}{2}\int_{t^n}^{t^{n+1}}(\tau-|s-t^{n+\frac{1}{2}}|)\partial_{tt}\mathbf{u}(s)\mathrm{d}s,$$

$$\begin{aligned}\left\|T_2^n[\mathbf{u}]\right\|^2 &= \frac{1}{4}\int_{\mathbb{R}}\left(\int_{t^n}^{t^{n+1}}(\tau-|s-t^{n+\frac{1}{2}}|)\partial_{tt}\mathbf{u}(x,s)\mathrm{d}s\right)^2\mathrm{d}x\\ &\leqslant \frac{1}{4}\int_{\mathbb{R}}\left(\int_{t^n}^{t^{n+1}}(\tau-|s-t^{n+\frac{1}{2}}|)^2\mathrm{d}s\cdot\int_{t^n}^{t^{n+1}}|\partial_{tt}\mathbf{u}(x,s)|^2\mathrm{d}s\right)\mathrm{d}x\\ &= \frac{7}{48}\tau^3\int_{t^n}^{t^{n+1}}\left\|\partial_{tt}\mathbf{u}(s)\right\|^2\mathrm{d}s.\end{aligned}$$

We split the term T_3^n as follows:

$$T_3^n = \frac{\mathbf{U}[\mathbf{u}^{*,n+\frac{1}{2}}]}{\sqrt{\mathcal{E}_1[\mathbf{u}^{*,n+\frac{1}{2}}]}} - \frac{\mathbf{U}[\mathbf{u}(t^{n+\frac{1}{2}})]}{\sqrt{\mathcal{E}_1[\mathbf{u}(t^{n+\frac{1}{2}})]}} := B_1 + B_2,$$

with

$$\begin{aligned}B_1 &= \frac{\mathbf{U}[\mathbf{u}^{*,n+\frac{1}{2}}]-\mathbf{U}[\mathbf{u}(t^{n+\frac{1}{2}})]}{\sqrt{\mathcal{E}_1[\mathbf{u}^{*,n+\frac{1}{2}}]}},\\ B_2 &= \frac{\mathbf{U}[\mathbf{u}(t^{n+\frac{1}{2}})](\mathcal{E}_1[\mathbf{u}(t^{n+\frac{1}{2}})]-\mathcal{E}_1[\mathbf{u}^{*,n+\frac{1}{2}}])}{\sqrt{\mathcal{E}_1[\mathbf{u}(t^{n+\frac{1}{2}})]\mathcal{E}_1[\mathbf{u}^{*,n+\frac{1}{2}}]}(\sqrt{\mathcal{E}_1[\mathbf{u}(t^{n+\frac{1}{2}})]}+\sqrt{\mathcal{E}_1[\mathbf{u}^{*,n+\frac{1}{2}}]})},\end{aligned}$$

which can be bounded by

$$\left\|B_1\right\|^2 \leqslant \left\|\frac{\delta\mathbf{U}}{\delta\mathbf{u}}\right\|^2_{L^\infty(\mathbb{R}^d)}\left\|T_4^n\right\|^2,\ \left\|B_2\right\|^2 \leqslant \left\|\mathbf{U}\right\|^4_{L^\infty(\mathbb{R}^d)}\left\|T_4^n\right\|^2,$$

which lead

$$\left\|T_3^n\right\|^2 \lesssim \left\|T_4^n\right\|^2,\ \left\|\mathscr{L}T_3^n\right\|^2 \lesssim \left\|\mathscr{L}T_4^n\right\|^2.$$

For the term T_4^n, we have

$$\begin{aligned}T_4^n &= \mathbf{u}^{*,n+\frac{1}{2}} - \mathbf{u}(t^{n+\frac{1}{2}})\\ &= \frac{3}{2}\mathbf{e}^n - \frac{1}{2}\mathbf{e}^{n-1} + \frac{1}{2}\int_{t^n}^{t^{n+1}}(\tau-|s-t^{n+\frac{1}{2}}|)\partial_{tt}\mathbf{u}(x,s)\mathrm{d}s\\ &\quad - \frac{1}{2}\int_{t^{n-1}}^{t^{n+1}}(\tau-|s-t^n|)\partial_{tt}\mathbf{u}(x,s)\mathrm{d}s,\end{aligned}$$

which leads to

$$\left\|T_4^n\right\|^2 \lesssim \|\mathbf{e}^n\|^2 + \|\mathbf{e}^{n-1}\|^2 + \tau^3\int_{t^{n-1}}^{t^{n+1}}\left\|\partial_{tt}\mathbf{u}(s)\right\|^2\mathrm{d}s.$$

Similarly, we can derive

$$\left\|\mathscr{L}T_4^n\right\|^2 \lesssim \left\|T_4^n\right\|^2 + \|\nabla\mathbf{e}^n\|^2 + \|\nabla\mathbf{e}^{n-1}\|^2 + \tau^3\int_{t^{n-1}}^{t^{n+1}}\left\|\partial_{tt}\nabla\mathbf{u}(s)\right\|^2\mathrm{d}s.$$

Then, we bound the given functions under conditions. Note that $r_2(t)\equiv 0$ in the continuous case for $t\in(0,T)$, we derive from (26a) and (26b) that

$$\begin{aligned}\left\|\boldsymbol{\omega}^{n+\frac{1}{2}}\right\|^2 \lesssim{}& \left\|\nabla\mathbf{e}^{n+1}\right\|^2 + \left\|\nabla\mathbf{e}^n\right\|^2 + \left\|\mathbf{e}^{n+1}\right\|^2 + \left\|\mathbf{e}^n\right\|^2 + \left\|\mathscr{L}T_2^n[\mathbf{u}]\right\|^2\\ &+ (s_1^{n+\frac{1}{2}})^2\left\|\mathbf{p}^n\right\|^2 + (T_2^n[r_1])^2\left\|\mathbf{p}^n\right\|^2 + \left\|T_3^n\right\|^2\\ &+ \frac{1}{\varepsilon^2}\left[(s_2^{n+\frac{1}{2}})^2\left\|\mathbf{u}^{*,n+\frac{1}{2}}\right\|^2 + (T_2^n[r_2])^2\left\|\mathbf{u}^{*,n+\frac{1}{2}}\right\|^2\right].\end{aligned}\tag{33}$$

On the other hand, the functions $\mathbf{p}^n, \mathscr{L}\mathbf{p}^n, \mathbf{u}^{*,n+\frac{1}{2}}$ could be estimated by

$$\begin{aligned}
&\|\mathbf{p}^n\| \leqslant \|\mathbf{p}[\mathbf{u}(t^{n+\frac{1}{2}})]\| + \|T_3^n\|, \quad \|\mathscr{L}\mathbf{p}^n\| \leqslant \|\mathscr{L}\mathbf{p}[\mathbf{u}(t^{n+\frac{1}{2}})]\| + \|\mathscr{L}T_3^n\|,\\
&\|\mathbf{u}^{*,n+\frac{1}{2}}\| \leqslant \|\mathbf{u}(t^{n+\frac{1}{2}})\| + \|T_4^n\|, \ \|\mathscr{L}\mathbf{u}^{*,n+\frac{1}{2}}\| \leqslant \|\mathscr{L}\mathbf{u}(t^{n+\frac{1}{2}})\| + \|\mathscr{L}T_4^n\|,\\
&\|\mathbf{p}^n - \mathscr{L}\mathbf{p}^n\| \leqslant \|\mathbf{p}^n\| + \|\mathscr{L}\mathbf{p}^n\|, \|\mathbf{u}^{*,n+\frac{1}{2}} - \mathscr{L}\mathbf{u}^{*,n+\frac{1}{2}}\| \leqslant \|\mathbf{u}^{*,n+\frac{1}{2}}\| + \|\mathscr{L}\mathbf{u}^{*,n+\frac{1}{2}}\|.
\end{aligned}$$

Next, we propose the estimation of the right hands consist by $I_1, I_2, \cdots, I_7$. For the sake of convenience, we put τ^{-1} in front of the terms T_1^ns, and take τ before the terms $\boldsymbol{\omega}^{n+\frac{1}{2}}, s_1^{n+\frac{1}{2}}, s_2^{n+\frac{1}{2}}$, T_2^ns, T_3^ns and T_4^ns by Cauchy's inequality. Then there hold

$$\begin{aligned}
I_1 &\lesssim \tau\|\boldsymbol{\omega}^{n+\frac{1}{2}}\|^2 + \frac{1}{\tau}\|\mathscr{L}T_1^n[\mathbf{u}]\|^2 + \tau\|\mathscr{L}T_2^n[\mathbf{u}]\|^2 + \tau\|\mathscr{L}^2T_2^n[\mathbf{u}]\|^2,\\
I_2 &\lesssim \tau(s_1^{n+\frac{1}{2}})^2 + \tau\|\boldsymbol{\omega}^{n+\frac{1}{2}}\|^2,\\
I_3 &\lesssim \tau(s_1^{n+\frac{1}{2}})^2 + \frac{1}{\tau}\|T_1^n[\mathbf{u}]\|^2 + \tau\|T_3^n\|^2 + \frac{1}{\tau}|T_1^n[r_1]|^2,\\
I_4 &\lesssim \tau\|\boldsymbol{\omega}^{n+\frac{1}{2}}\|^2 + \tau|T_2^n[r_1]|^2 + \frac{1}{\tau}\|T_1^n[\mathbf{u}]\|^2 + \tau\|\mathscr{L}T_3^n\|^2,\\
I_5 &\lesssim \frac{1}{\varepsilon}\tau(s_2^{n+\frac{1}{2}})^2 + \varepsilon\tau\|\boldsymbol{\omega}^{n+\frac{1}{2}}\|^2,\\
I_6 &\lesssim \frac{1}{\varepsilon}\tau(s_2^{n+\frac{1}{2}})^2 + \frac{\varepsilon}{\tau}\|T_1^n[\mathbf{u}]\|^2 + \frac{\varepsilon}{\tau}|T_1^n[r_2]|^2 + \varepsilon\tau\|T_4^n\|^2,\\
I_7 &\lesssim \varepsilon\tau\|\boldsymbol{\omega}^{n+\frac{1}{2}}\|^2 + \frac{\tau}{\varepsilon}|T_2^n[r_2]|^2 + \frac{\varepsilon}{\tau}\|T_1^n[\mathbf{u}]\|^2.
\end{aligned}$$

Finally, combining the above estimates into (29), we obtain

$$\begin{aligned}
&\frac{1}{2}\big(\|\nabla\mathbf{e}^{n+1}\|^2 + \gamma^2\|\mathbf{e}^{n+1}\|^2 - \|\nabla\mathbf{e}^n\|^2 - \gamma^2\|\mathbf{e}^n\|^2\big)\\
&\qquad + (s_1^{n+1})^2 + \frac{1}{4\varepsilon}(s_2^{n+1})^2 - (s_1^n)^2 - \frac{1}{4\varepsilon}(s_2^n)^2\\
&\lesssim \tau\|\boldsymbol{\omega}^{n+\frac{1}{2}}\|^2 + \tau(s_1^{n+\frac{1}{2}})^2 + \frac{1}{\tau}\|T_1^n[\mathbf{u}]\|^2 + \frac{1}{\tau}\|\mathscr{L}T_1^n[\mathbf{u}]\|^2 + \frac{1}{\tau}|T_1^n[r_1]|^2\\
&\qquad + \frac{1}{\tau}|T_1^n[r_2]|^2 + \tau\|\mathscr{L}T_2^n[\mathbf{u}]\|^2 + \tau\|\mathscr{L}^2T_2^n[\mathbf{u}]\|^2 + \tau|T_2^n[r_1]|^2\\
&\qquad + \tau\|T_3^n\|^2 + \tau\|\mathscr{L}T_3^n\|^2 + \tau\|T_4^n\|^2 + \frac{1}{\varepsilon^2}\Big[\tau(s_2^{n+\frac{1}{2}})^2 + \tau|T_2^n[r_2]|^2\Big],\\
&\lesssim \tau\Big[\|\nabla\mathbf{e}^{n+1}\|^2 + \|\mathbf{e}^{n+1}\|^2 + \|\nabla\mathbf{e}^n\|^2 + \|\mathbf{e}^n\|^2\\
&\qquad + (s_1^{n+1})^2 + \frac{1}{\varepsilon^2}(s_2^{n+1})^2 + (s_1^n)^2 + \frac{1}{\varepsilon^2}(s_2^n)^2\Big]\\
&\qquad + \tau\Big[\|\mathscr{L}T_2^n[\mathbf{u}]\|^2 + \|\mathscr{L}^2T_2^n[\mathbf{u}]\|^2 + \|T_3^n\|^2 + \|\mathscr{L}T_3^n\|^2\\
&\qquad + \|T_4^n\|^2 + |T_2^n[r_1]|^2 + \frac{1}{\varepsilon^2}|T_2^n[r_2]|^2\Big]\\
&\qquad + \frac{1}{\tau}\Big[\|T_1^n[\mathbf{u}]\|^2 + \|\mathscr{L}T_1^n[\mathbf{u}]\|^2 + |T_1^n[r_1]|^2 + |T_1^n[r_2]|^2\Big]\\
&\lesssim \tau\Big[\|\nabla\mathbf{e}^{n+1}\|^2 + \|\mathbf{e}^{n+1}\|^2 + \|\nabla\mathbf{e}^n\|^2 + \|\mathbf{e}^n\|^2 + \|\nabla\mathbf{e}^{n-1}\|^2 + \|\mathbf{e}^{n-1}\|^2\\
&\qquad + (s_1^{n+1})^2 + \frac{1}{\varepsilon^2}(s_2^{n+1})^2 + (s_1^n)^2 + \frac{1}{\varepsilon^2}(s_2^n)^2\Big]
\end{aligned}$$

$$+\tau^4\int_{t^n}^{t^{n+1}}\Big(\big\|\partial_{ttt}\mathbf{u}(s)\big\|^2+\big\|\partial_{ttt}\nabla\mathbf{u}(s)\big\|^2+\big\|\partial_{tt}\Delta\mathbf{u}(s)\big\|^2+\big\|\partial_{tt}\mathbf{u}(s)\big\|^2$$
$$+\big\|r_1'''(s)\big\|^2+\frac{1}{\varepsilon^2}\big\|r_2'''(s)\big\|^2+\big\|r_1''(s)\big\|^2+\frac{1}{\varepsilon^2}\big\|r_2''(s)\big\|^2\Big)\mathrm{d}s$$
$$+\tau^4\int_{t^{n-1}}^{t^{n+1}}\Big(\big\|\partial_{tt}\nabla\mathbf{u}(s)\big\|^2+\big\|\partial_{tt}\mathbf{u}(s)\big\|^2\Big)\mathrm{d}s.$$

Applying a discrete Grönwall inequality (e.g, Lemma B.10 in [15])to the above, we derive the desired results. □

5. **Spatial discretization and fast implementation.** The SAV schemes presented in the last two sections can be used with any consistent Galerkin type spatial discretization. Since the domain is $\mathbb{R}^d$, we shall use the Hermite–Galerkin method with the Hermite function as basis functions (cf. Section 7.4.2 in [15]) for spatial discretization. It is easy to verify that the fully discrete version of (19) with the Hermite–Galerkin method in space has the same stability results in Theorem 3.2 for (19). We recall that the Hermite functions are defined by

$$\widehat{H}_n(x)=\frac{1}{\pi^{\frac{1}{4}}\sqrt{2^n n!}}e^{-\frac{x^2}{2}}H_n(x),\quad n\geqslant 0,\quad x\in\mathbb{R},\tag{34}$$

where $H_n(x)$ represents Hermite polynomial of order n *(see (7.65) on Page 255 in* [15]*)*. Thanks to the orthogonality of the Hermite functions *(see (7.71)–(7.72) on Page 256 in* [15]*)*, we have

$$\int_{-\infty}^{+\infty}\widehat{H}_m(x)\widehat{H}_n(x)\mathrm{d}x=\delta_{mn},\tag{35}$$

and *(see(7.75) on Page 257 in* [15]*)*

$$\widehat{H}_n'(x)=\sqrt{\frac{n}{2}}\widehat{H}_{n-1}(x)-\sqrt{\frac{n+1}{2}}\widehat{H}_{n+1}(x).\tag{36}$$

We now describe how to solve the decoupled systems in (14) and and (23) which consist a sequence of the following equations:

$$\Big(I-\frac{1}{2}\tau\sum_{j=1}^{d}GD_j\partial_j-\frac{1}{2}\tau\gamma GA\Big)\mathbf{v}=\mathbf{c},\tag{37}$$

where $\mathbf{c}$ is a given function depending on solutions at previous time steps.

In 1–D case, we define our approximation space as

$$\hat{P}_M=\mathrm{span}\{\hat{H}_0(x),\ \hat{H}_1(x),\ \cdots,\ \hat{H}_M(x)\}.$$

In 2–D case, we set $\hat{\mathbb{P}}_M:=\hat{P}_M^2=\hat{P}_M\times\hat{P}_M$.

Then, the Hermite spectral method for (37) is to find $\mathbf{v}_M\in X_M:=(\hat{P}_M^d)^4, d=1,2$ such that

$$\Big((I-\frac{1}{2}\tau\sum_{j=1}^{d}GD_j\partial_j-\frac{1}{2}\tau\gamma GA)\mathbf{v}_M,w_M\Big)=(\mathbf{c},w_M),\quad\forall w_M\in X_M.\tag{38}$$

We now describe how to implement (38) efficiently.

We start with the 1–D case. Setting $\mathbf{v}_M = (v_{1,M}, v_{2,M}, v_{3,M}, v_{4,M})$, $\mathbf{c} = (c_1, c_2, c_3, c_4)$ and $D_1 = (d_{kl}^{(1)})_{k,l=1,2,3,4}$, we can rewrite (38) component–wise as

$$\Big(v_{j,M} - \frac{1}{2}\tau \sum_{k=0}^{4}\sum_{l=0}^{4} g_{jk} d_{kl}^{(1)} \partial_x v_{l,M} - \frac{1}{2}\gamma\tau \sum_{k=0}^{4}\sum_{l=0}^{4} g_{jk} a_{kl} v_{l,M}, w_M\Big) = (c_j, w_M), \\ \forall w_M \in \hat{P}_M,\ j = 1,2,3,4. \tag{39}$$

Setting

$$v_{j,M}(x) = \sum_{m=0}^{M} \tilde{v}_{jm} \widehat{H}_m(x), \quad q_{mm'} = (\partial_x H_{m'}, H_m), \quad \tilde{c}_{jm}(x) = (c_j, \widehat{H}_m),$$

it is easy to derive from (35) and (36) that

$$q_{mm'} = \begin{cases} \sqrt{\dfrac{m}{2}}, & m' = m-1, \\ -\sqrt{\dfrac{m+1}{2}}, & m' = m+1, \\ 0, & \text{otherwise.} \end{cases}$$

Taking $w_M = H_m, m = 0, 1, \cdots, M$ separately in (39), we obtain

$$\tilde{v}_{jm} - \frac{1}{2}\tau \sum_{k=0}^{4}\sum_{l=0}^{4}\sum_{m'=0}^{M} g_{jk} d_{kl}^{(1)} \tilde{v}_{lm} q_{mm'} - \frac{1}{2}\gamma\tau \sum_{k=0}^{4}\sum_{l=0}^{4} g_{jk} a_{kl} \tilde{v}_{lm} = \tilde{c}_{jm},$$

which can be written in matrix form as

$$\Big(I - \frac{1}{2}\gamma\tau GA\Big)\tilde{\mathbf{V}}_M - \frac{1}{2}\tau G D_1 \tilde{\mathbf{V}}_M Q = \tilde{\mathbf{C}}_M,$$

with

$$\tilde{\mathbf{V}}_M = (v_{jm})_{1\leqslant j\leqslant 4; 0\leqslant m\leqslant M}, \quad \tilde{\mathbf{C}} = (c_{jm})_{1\leqslant j\leqslant 4; 0\leqslant m\leqslant M}, \quad Q = (q_{mm'})_{0\leqslant m, m'\leqslant M}.$$

Since Q is a tridiagonal antisymmetric matrix, the above system can be solved in $O(M)$ operations.

In the 2–D case, we write

$$v_{j,M}(x,y) = \sum_{l=0}^{M}\sum_{m=0}^{M} \check{v}_{lm}^{(j)} \widehat{H}_l(x) \widehat{H}_m(y), \quad \check{V}^{(j)} = \Big(\check{v}_{lm}^{(j)}\Big)_{(M+1)\times(M+1)}, \quad j = 1,2,3,4,$$

and denote

$$\check{\mathbf{V}} = \big(\check{V}^{(1)}, \quad \check{V}^{(2)}, \quad \check{V}^{(3)}, \quad \check{V}^{(4)}\big)^T.$$

Similarly, we define the matrices $\check{\mathbf{C}}$. Then, by taking each component of w_M in (38) to be $\widehat{H}_l(x)\widehat{H}_m(y), l, m = 0, 1, \cdots, M$, then (38) becomes:

$$\check{V}^{(1)} + \frac{1}{2}\tau Q^T \check{V}^{(3)} + \frac{1}{2}\tau \check{V}^{(4)} Q - \frac{1}{2}\tau\gamma \check{V}^{(2)} = \check{C}^{(1)}, \tag{40a}$$

$$\check{V}^{(2)} + \frac{1}{2}\tau Q^T \check{V}^{(4)} + \frac{1}{2}\tau \check{V}^{(3)} Q - \frac{1}{2}\tau\gamma \check{V}^{(1)} = \check{C}^{(2)}, \tag{40b}$$

$$\check{V}^{(3)} + \frac{1}{2}\tau Q^T \check{V}^{(1)} + \frac{1}{2}\tau \check{V}^{(2)} Q - \frac{1}{2}\tau\gamma \check{V}^{(4)} = \check{C}^{(3)}, \tag{40c}$$

$$\check{V}^{(4)} + \frac{1}{2}\tau Q^T \check{V}^{(2)} + \frac{1}{2}\tau \check{V}^{(1)} Q - \frac{1}{2}\tau\gamma \check{V}^{(3)} = \check{C}^{(4)}. \tag{40d}$$

We now describe how to solve (40) by using the matrix decomposition method. We first compute the eigenpairs (E,Λ) of Q, i.e., $EQ=\Lambda E$, where Λ is a diagonal matrix with entries being the eigenvalues of matrix Q, and each column of E is an eigenvector. Since $Q^T=-Q$, the eigenvalues of Q are 0 or pure imaginary. Let $\bar{E}$ be the conjugate matrix of E, and set $\check{V}^{(k)}=\bar{E}^T\check{W}^{(k)}E$, $(k=1,2,3,4)$. Then (40) becomes

$$\begin{aligned}
&\check{W}^{(1)}+\frac{1}{2}\tau\bar{\Lambda}\check{W}^{(3)}+\frac{1}{2}\tau\check{W}^{(4)}\Lambda-\frac{1}{2}\tau\gamma\check{W}^{(2)}=E(\check{C}^{(1)})\bar{E}^T,\\
&\check{W}^{(2)}+\frac{1}{2}\tau\bar{\Lambda}\check{W}^{(4)}+\frac{1}{2}\tau\check{W}^{(3)}\Lambda-\frac{1}{2}\tau\gamma\check{W}^{(1)}=E(\check{C}^{(2)})\bar{E}^T,\\
&\check{W}^{(3)}+\frac{1}{2}\tau\bar{\Lambda}\check{W}^{(1)}+\frac{1}{2}\tau\check{W}^{(2)}\Lambda-\frac{1}{2}\tau\gamma\check{W}^{(4)}=E(\check{C}^{(3)})\bar{E}^T,\\
&\check{W}^{(4)}+\frac{1}{2}\tau\bar{\Lambda}\check{W}^{(2)}+\frac{1}{2}\tau\check{W}^{(1)}\Lambda-\frac{1}{2}\tau\gamma\check{W}^{(3)}=E(\check{C}^{(4)})\bar{E}^T.
\end{aligned}$$

Setting $\check{F}^{(k)}=E(\check{C}^{(k)})\bar{E}^T$ $(k=1,2,3,4)$, we arrive at

$$\Big(I-\frac{1}{2}\tau\bar{\lambda}_iGD_1-\frac{1}{2}\tau\lambda_jGD_2-\frac{1}{2}\tau\gamma GA\Big)W_{ij}=F_{ij},\quad 0\leqslant i,j\leqslant(M+1)^2,\tag{41}$$

where

$$W_{ij}=\begin{pmatrix}\check{W}_{ij}^{(1)}\\ \check{W}_{ij}^{(2)}\\ \check{W}_{ij}^{(3)}\\ \check{W}_{ij}^{(4)}\end{pmatrix},\quad F_{ij}=\begin{pmatrix}\check{F}_{ij}^{(1)}\\ \check{F}_{ij}^{(2)}\\ \check{F}_{ij}^{(3)}\\ \check{F}_{ij}^{(4)}\end{pmatrix}.$$

For each i,j, the scheme (41) is a 4×4 system that can be solved directly in $O(1)$ operations, so the cost of solving $\check{W}^{(k)}$ $(k=1,2,3,4)$ is $O(M^2)$. We can then obtain $\check{\mathbf{V}}$ from $\check{V}^{(k)}=\bar{E}^T\check{W}^{(k)}E$ $(k=1,2,3,4)$ which are just a few matrix multiplications with $O(M^3)$ operations.

In summary, the total cost of solving the linear system (38) at each time step is $O(M^d)+C_dM^{d+1}$ where C_d is a small constant for $d=1,2$.

6. Numerical results.

We present in this section some numerical experiments to validate the LagSAV/CN and MSAV/CN–Hermite method.

6.1. Energy and mass conservation.

We consider first the of 1–D Dirac equation which admits a standing wave solution, which its initial condition is defined by

$$\mathbf{u}^0(x)=\left(\frac{0.36\sqrt{10}\cosh(0.6x)}{1+0.8\cosh(0.6x)},0,0,\frac{0.12\sqrt{10}\sinh(0.6x)}{1+0.8\cosh(0.6x)}\right)^T.$$

The conservation of original energy and mass by LagSAV/CN scheme and MSAV/CN scheme are presented in Figure 1 and Figure 2.

6.2. Standing wave solution in 1–D.

We consider first the of 1–D Dirac equation which admits a standing wave solution defined *(see Eq.(4)–Eq.(6) in* [14]*)* by

$$\Psi^*(x,t)=\begin{pmatrix}\psi_1^*(x,t)\\ \psi_2^*(x,t)\end{pmatrix}=\begin{pmatrix}\alpha(x)\\ \mathrm{i}\beta(x)\end{pmatrix}\exp(-\mathrm{i}\gamma_0t),\quad 0<\gamma_0\leqslant\gamma,$$

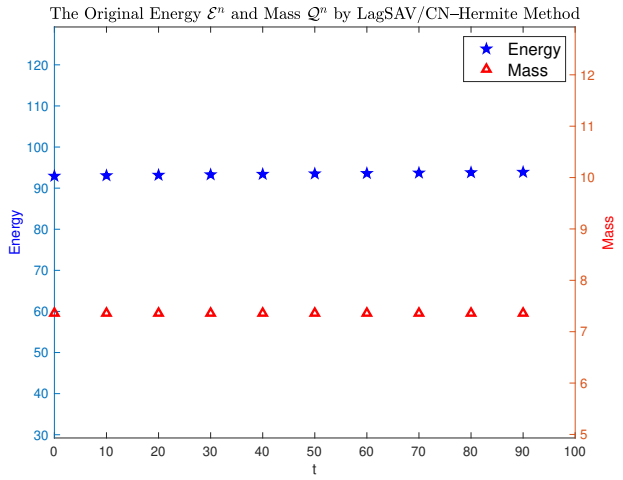


FIGURE 1. The original energy $\mathcal{E}^n$ and mass $\mathcal{Q}^n$ with $\mathbf{u}^0$ satisfies single soliton solution by LagSAV/CN–Hermite Method ($\tau = 10^{-4}$). Both original energy and mass preserved by the following LagSAV/CN–Hermite approach. It is obvious that the mass preserved because of (9d). To make sure energy preservation, it needs lots of computation to keep the Lagrange factors close to the continuous case.

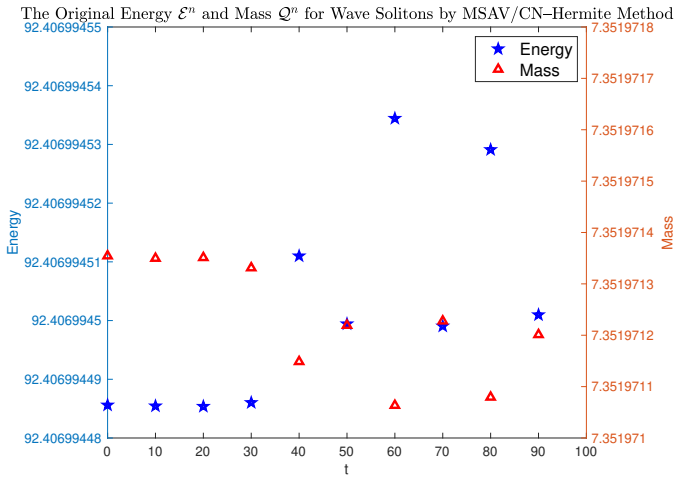


FIGURE 2. The original energy $\mathcal{E}^n$ and mass $\mathcal{Q}^n$ by MSAV/CN–Hermite method with $\mathbf{u}^0$ satisfies single soliton solution ($\tau = 10^{-4}, \varepsilon = 10^{-6}$). The modified discrete energy and mass are both calculated by the MSAV/CN–Hermite method. We observe that the errors for the energy and mass conservation are of order 10^{-8} and 10^{-7}, respectively.

where

$$\alpha(x) = \frac{\sqrt{2(\gamma^2-\gamma_0^2)(\gamma+\gamma_0)}\cosh(\sqrt{\gamma^2-\gamma_0^2}x)}{\gamma+\gamma_0\cosh(\sqrt{\gamma^2-\gamma_0^2}x)},$$
$$\beta(x) = \frac{\sqrt{2(\gamma^2-\gamma_0^2)(\gamma-\gamma_0)}\sinh(\sqrt{\gamma^2-\gamma_0^2}x)}{\gamma+\gamma_0\cosh(\sqrt{\gamma^2-\gamma_0^2}x)}.$$

We use the MSAV/CN–Hermite method with $\varepsilon = 10^{-4}$. We first fix $\tau = 10^{-4}$ so that the time discretization error is negligible compared with the spatial discretization error, and plot in the left of Figure 3 the L^2–error vs. M. Then, we fix $M = 128$ so that the the spatial discretization error is negligible compared with the time discretization error, and plot in the right of Figure 3 the L^2–error vs. τ. We observe an exponential convergence in space and second–order convergence in time.

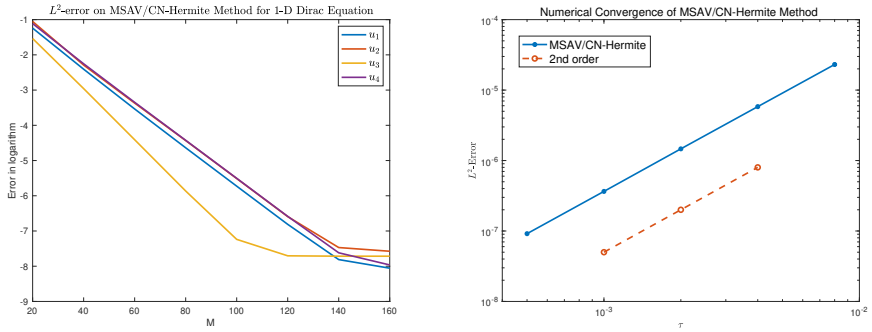


FIGURE 3. Left: L^2–Error vs. M with $\tau = 10^{-4}$ at $t = 1$; Right: L^2–Error vs. τ with $M = 128$ at $t = 1$.

6.3. **Collision phenomena in 1–D.** Let us denote a solitary wave solution of the Dirac equation by

$$\Psi^c(x-x_0,t) = \left(\psi_1^c(x-x_0,t), \quad \psi_2^c(x-x_0,t)\right)^T,$$

with

$$\psi_1^c(x-x_0,t) = \sqrt{\frac{\mu+1}{2}}\psi_1^*(\tilde{x},\tilde{t}) + \text{sign}(v)\sqrt{\frac{\mu-1}{2}}\psi_2^*(\tilde{x},\tilde{t}),$$
$$\psi_2^c(x-x_0,t) = \sqrt{\frac{\mu+1}{2}}\psi_2^*(\tilde{x},\tilde{t}) + \text{sign}(v)\sqrt{\frac{\mu-1}{2}}\psi_1^*(\tilde{x},\tilde{t}),$$

here $\mu = (1-v^2)^{-1/2}$, $\tilde{x} = \mu(x-x_0-vt)$, $\tilde{t} = \mu(t-v(x-x_0))$ and $\text{sign}(x)$ is the sign function.

Next, we simulate the collision of two solitons with the initial condition

$$\Psi^b(x,0) = \Psi^c(x-x_l,0) + \Psi^c(x-x_r,0),$$

and of three solitons with the initial condition

$$\Psi^t(x,0) = \Psi^c(x-x_l,0) + \Psi^c(x-x_m,0) + \Psi^c(x-x_r,0).$$

Some examples of the binary and ternary collisions are shown in Figures 4 – 8. These results are consistent with published results in [14].

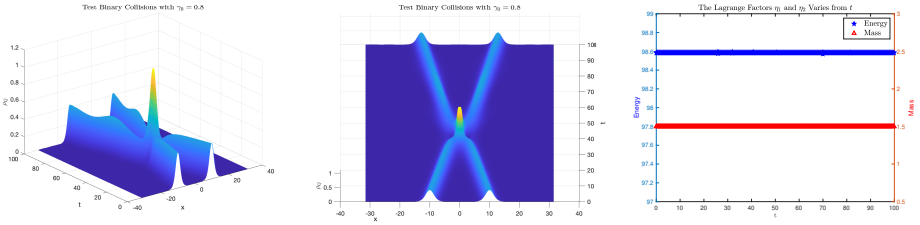


FIGURE 4. Binary Collision Phenomena with $\tau = 0.01$ by LagSAV/CN–Hermite Method.

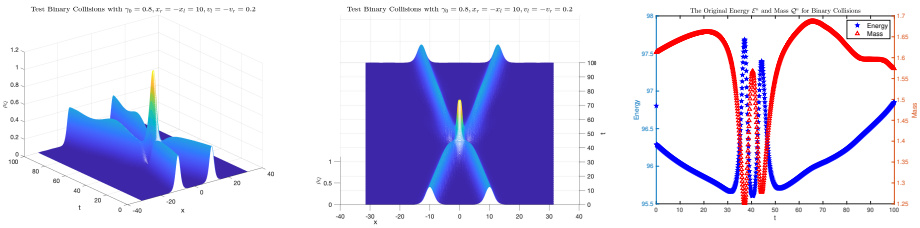


FIGURE 5. Binary Collision Phenomena with $\tau = 0.01$ by MSAV/CN–Hermite Method.

We observe from these simulations that for sufficiently small time steps, the LagSAV/CN–Hermite method works well and preserves mass and the original energy. However, if the time step is not sufficiently small, the Newton iteration for the nonlinear algebraic system may not converge and the scheme may become unstable. On the other hand, the MSAV/CN–Hermite method always admits a unique solution and preserves a modified energy, but at larger time steps, the original energy is not well preserved.

6.4. **2–D dirac equation.** As the last example, we simulate the collision phenomena by both LagSAV/CN– and MSAV/CN–Hermite methods for the 2–D Dirac equation. We set the initial condition to be

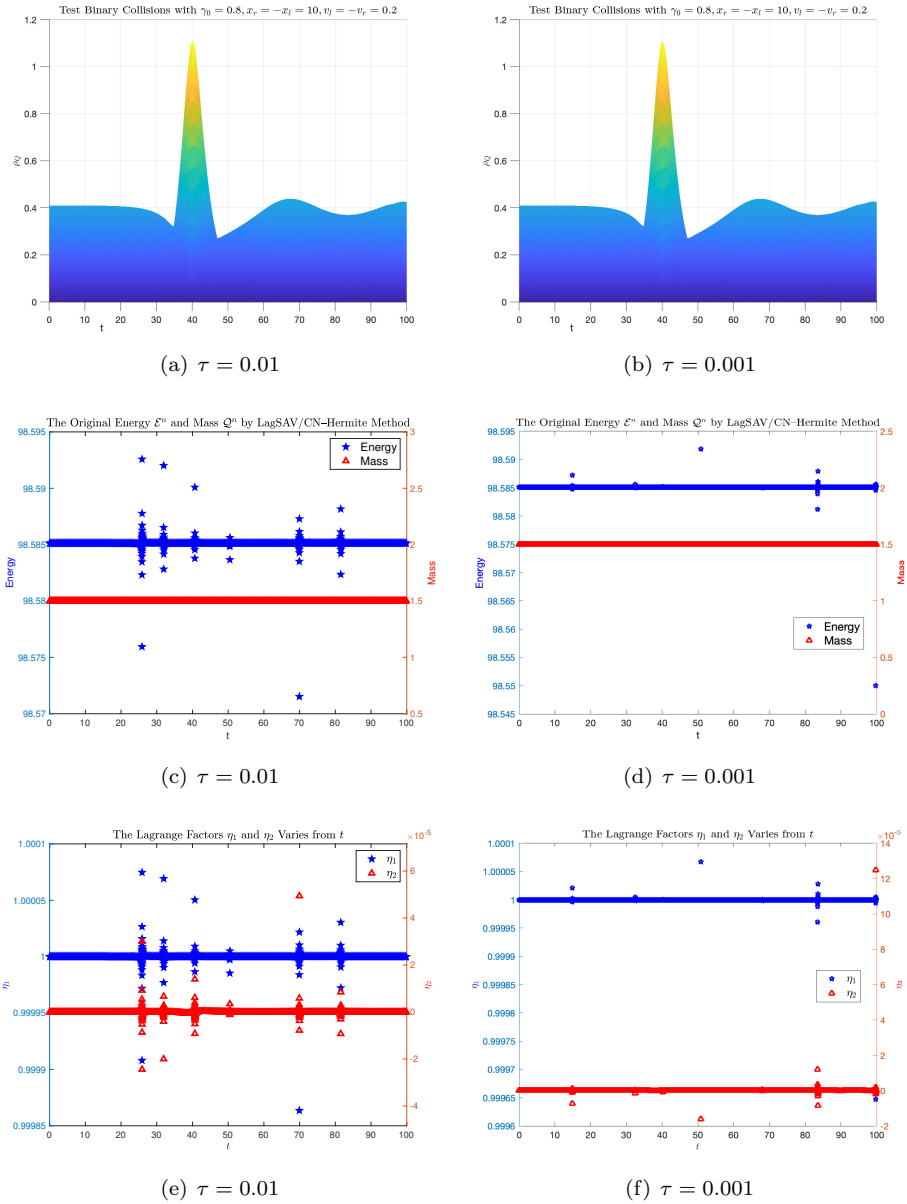
$$\mathbf{u}(x,y,0) = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} \exp(-(x^2+y^2)/4) \\ 0 \\ 0 \\ \exp(-((x-5)^2+(y-5)^2)/4) \end{pmatrix},$$

and plot snapshots of the simulation by both schemes in Figures 9 and 10, respectively. We observe that both schemes lead to essentially the same results.

7. **Summary.** We presented in this paper some energy preserving schemes for the one and two dimensional Dirac equation. Our schemes are based on the LagSAV/CN and MSAV/CN approaches coupled with Hermite approximation in space. The SAV (resp. MSAV) approach enables us to preserve the energy (resp. a modified energy), and the Hermite approximation leads to spectral convergence in space. In addition, we established the convergence for the MSAV/CN–Hermite scheme in both space and time. We also presented ample numerical experiments to verify the mass

and energy preservations and the convergence rates. Furthermore, our numerical simulations for the binary and ternary collision indicate that our schemes are robust and accurate.

Acknowledgements. The work of Z.Y. is partially supported by Guangdong Basic and Applied Basic Research Foundation No. 2020B1515310006 and the China



(a) $\tau = 0.01$ (b) $\tau = 0.001$

(c) $\tau = 0.01$ (d) $\tau = 0.001$

(e) $\tau = 0.01$ (f) $\tau = 0.001$

FIGURE 6. The Original Energy, Mass and Lagrange Factors on Binary Collision Phenomena by LagSAV/CN–Hermite Method.

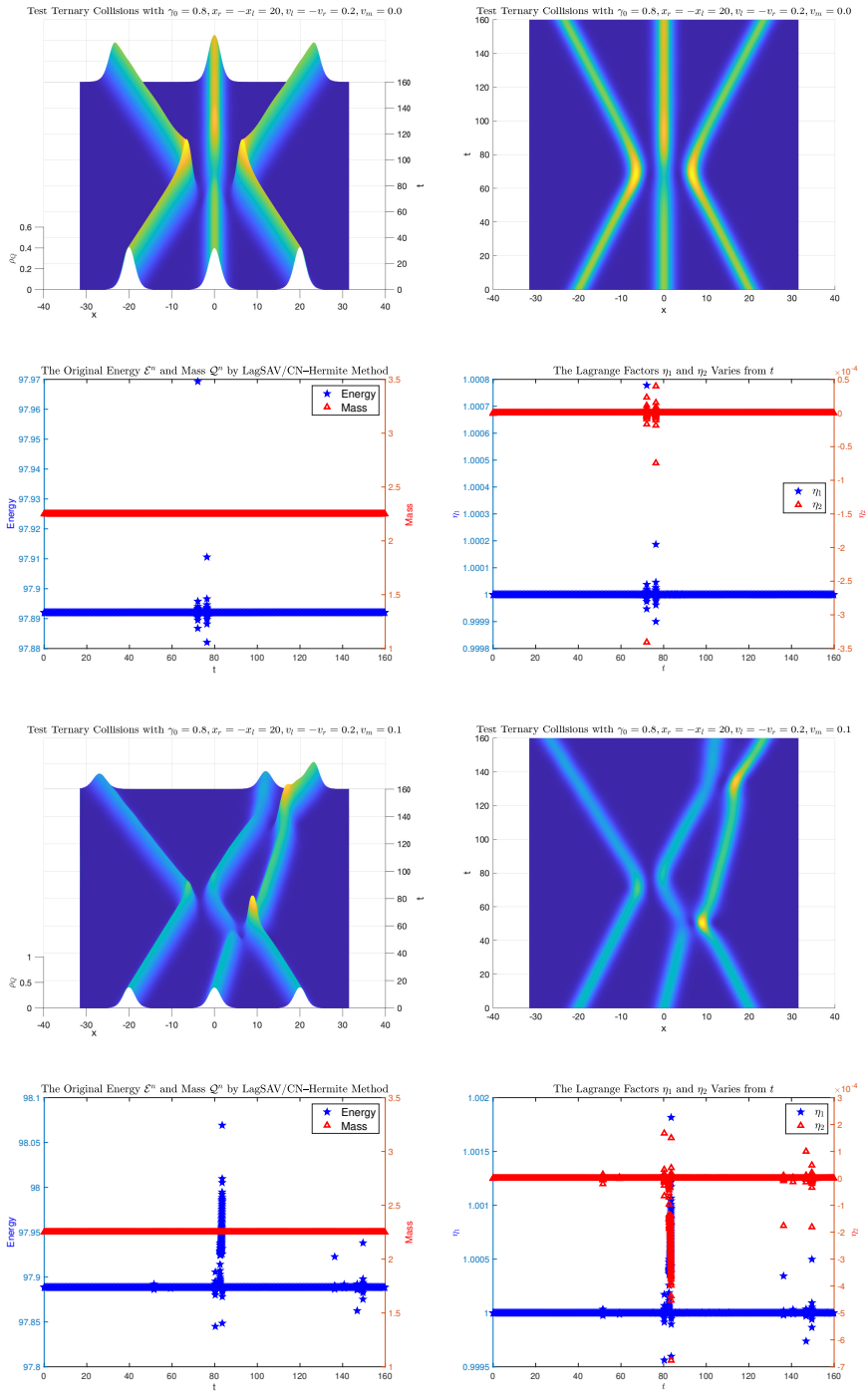


FIGURE 7. Ternary Collision Phenomena with $\tau = 0.01$ by LagSAV/CN–Hermite Method.

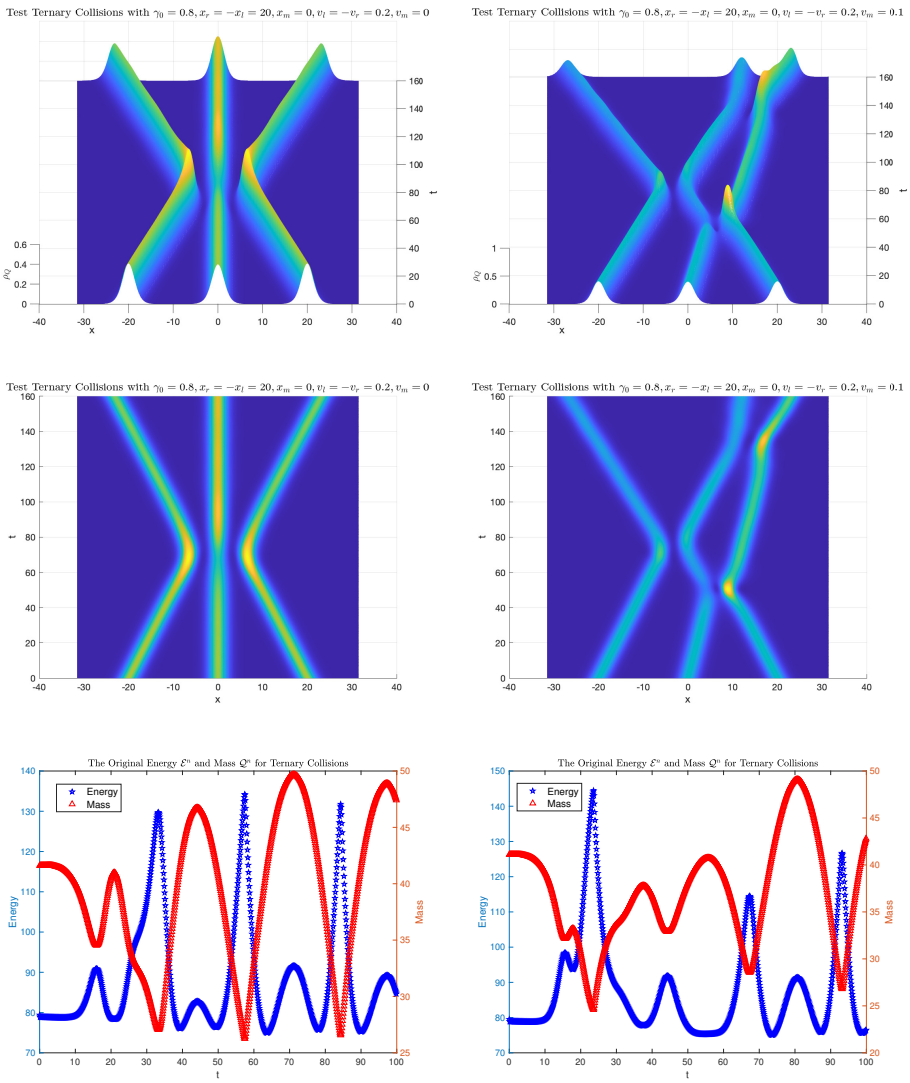
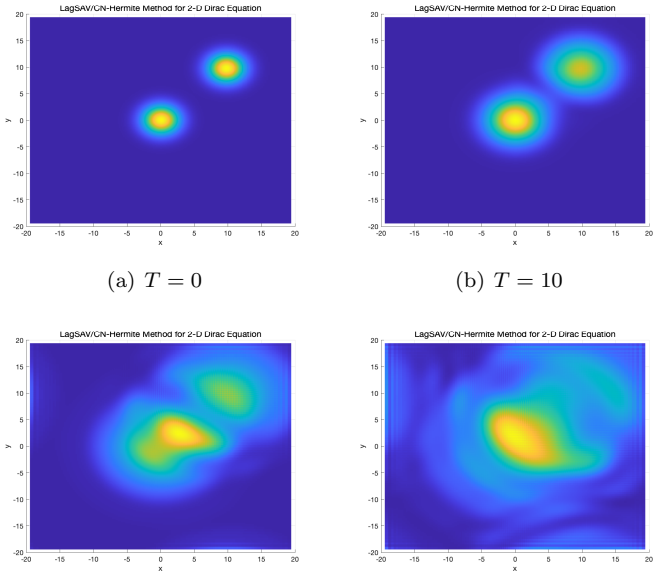


FIGURE 8. Ternary Collision Phenomena with $\tau = 0.01$ by MSAV/CN–Hermite Method.

Scholarship Council, while the work of J.S is partially supported by NSFC Grant 11971407.

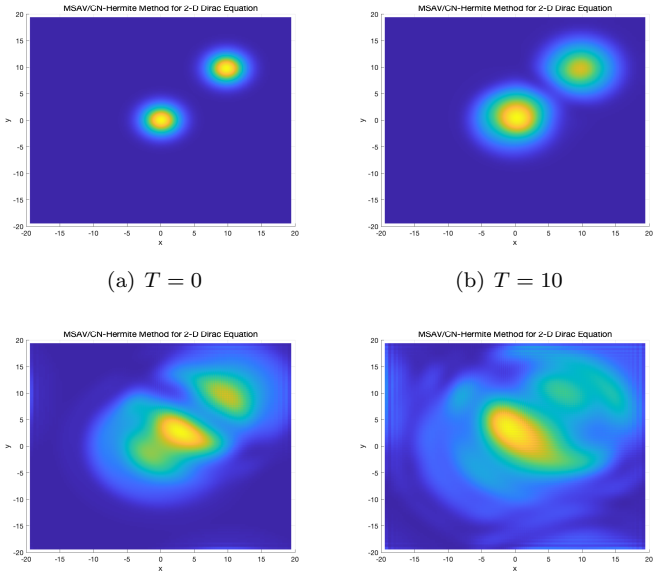
CRediT authorship contribution statement. Zhe Yu and Jie Shen all contributed to the conceptualization, methodology and writing. Zhe Yu carried out most of the error estimation and numerical simulations, and Jie Shen proposed the motivation and methodology.



(a) $T = 0$ (b) $T = 10$

(c) $T = 20$ (d) $T = 30$

FIGURE 9. Collision Phenomena by 2–D Dirac Equation with $\tau = 0.01, M = 200$ by LagSAV/CN–Hermite Method.



(a) $T = 0$ (b) $T = 10$

(c) $T = 20$ (d) $T = 30$

FIGURE 10. Collision Phenomena by 2–D Dirac Equation with $\tau = 0.01, M = 200, \varepsilon = 0.1$ by MSAV/CN–Hermite Method.

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.

REFERENCES

[1] W. Bao, Y. Cai, X. Jia and Q. Tang, Numerical methods and comparison for the Dirac equation in the nonrelativistic limit regime, *Journal of Scientific Computing*, **71** (2017), 1094–1134.

[2] Q. Cheng and J. Shen, Multiple scalar auxiliary variable (MSAV) approach and its application to the phase–field vesicle membrane model, *SIAM Journal on Scientific Computing*, **40** (2018), A3982–A4006.

[3] Q. Cheng and J. Shen, Global constraints preserving scalar auxiliary variable schemes for gradient flows, *SIAM Journal on Scientific Computing*, **42** (2020), A2489–A2513.

[4] B. Deng, J. Shen and Q. Zhuang, Second–order SAV schemes for the nonlinear Schrödinger equation and their error analysis, *J. Sci. Comput.*, **88** (2021), Paper No. 69, 24pp.

[5] C. M. Elliott and A. M. Stuart, The global dynamics of discrete semilinear parabolic equations, *SIAM J. Numer. Anal.*, **30** (1993), 1622–1663.

[6] F. Fillion-Gourdeau, E. Lorin and A. Bandrauk, Numerical solution of the time–dependent Dirac equation in coordinate space without fermion–doubling, *Computer Physics Communications*, **183** (2012), 1403–1415.

[7] F. Fillion-Gourdeau, S. MacLean and R. Laflamme, Algorithm for the solution of the Dirac equation on digital quantum computers, *Physical Review A*, **95** (2017), 042343.

[8] Y. Fu, D. Hu and Y. Wang, High–order structure–preserving algorithms for the multi–dimensional fractional nonlinear Schrödinger equation based on the SAV approach, *Math. Comput. Simulation*, **185** (2021), 238–255.

[9] W. Heisenberg, Quantum theory of fields and elementary particles, In *Scientific Review Papers, Talks, and Books Wissenschaftliche Übersichtsartikel, Vorträge und Bücher*, 1984, 552–561.

[10] D. Hou, M. Azaiez and C. Xu, A variant of scalar auxiliary variable approaches for gradient flows, *Journal of Computational Physics*, **395** (2019), 307–332.

[11] C. Jiang, J. Cui, X. Qian and S. Song, High–order linearly implicit structure–preserving exponential integrators for the nonlinear Schrödinger equation, *J. Sci. Comput.*, **90** (2022), Paper No. 66, 27pp.

[12] M. Jiang, Z. Zhang and J. Zhao, Improving the accuracy and consistency of the scalar auxiliary variable (SAV) method with relaxation, *J. Comput. Phys.*, **456** (2022), Paper No. 110954, 20pp.

[13] A. F. Ranada, Classical nonlinear Dirac field models of extended particles. In, quantum theory, groups, fields and particles, *Mathematical Physics Studies*, 1983.

[14] S. Shao and H. Tang, Interaction for the solitary waves of a nonlinear Dirac model, *Physics Letters A*, **345** (2005), 119–128.

[15] J. Shen, T. Tang and L.-L. Wang, *Spectral Methods: Algorithms, Analysis and Applications*, Springer Series in Computational Mathematics, 41. Springer, Heidelberg, 2011.

[16] J. Shen, J. Xu and J. Yang, The scalar auxiliary variable (SAV) approach for gradient flows, *Journal of Computational Physics*, **353** (2018), 407–416.

[17] J. Shen, J. Xu and J. Yang, A new class of efficient and robust energy stable schemes for gradient flows, *SIAM Review*, **61** (2019), 474–506.

[18] B. Thaller, *The Dirac Equation*, Texts and Monographs in Physics. Springer-Verlag, Berlin, 1992.

[19] H. Weyl, A remark on the coupling of gravitation and electron, *Physical Review*, **77** (1950), 699-701.

[20] J. Xu, S. Shao and H. Tang, Numerical methods for nonlinear Dirac equation, *Journal of Computational Physics*, **245** (2013), 131–149.

[21] Z. Xu, W. Cai, Y. Song and Y. Wang, Explicit high–order energy–preserving exponential time differencing method for nonlinear Hamiltonian PDEs, *Appl. Math. Comput.*, **404** (2021), Paper No. 126208, 15pp.

[22] X. Yang, Linear, first and second–order, unconditionally energy stable numerical schemes for the phase field model of homopolymer blends, *J. Comput. Phys.*, **327** (2016), 294–316.

[23] Y. Zhang and J. Shen, Efficient structure preserving schemes for the Klein–Gordon–Schrödinger equations, *J. Sci. Comput.*, **89** (2021), Paper No. 47, 26pp.
[24] Q. Zhuang and J. Shen, Efficient SAV approach for imaginary time gradient flows with applications to one– and multi–component Bose–Einstein condensates, *J. Comput. Phys.*, **396** (2019), 72–88.

Received June 2022; revised November 2022; early access November 2022.