HIGH-ORDER MASS- AND ENERGY-CONSERVING METHODS FOR THE NONLINEAR SCHRÖDINGER EQUATION

GENMING BAI*, JIASHUN HU*, BUYANG LI*

Abstract. A class of high-order mass- and energy-conserving methods is proposed for the nonlinear Schrödinger equation based on Gauss collocation in time and finite element discretization in space, by introducing a mass- and energy-correction post-process at every time level. The existence, uniqueness and high-order convergence of the numerical solutions are proved. In particular, the error of the numerical solution is $O(\tau^{k+1}+h^p)$ in the $L^\infty(0,T;H^1)$ norm after incorporating the accumulation errors arising from the post-processing correction procedure at all time levels, where k and p denote the degrees of finite elements in time and space, respectively, which can be arbitrarily large. Several numerical examples are provided to illustrate the performance of the proposed new method, including the conservation of mass and energy, and the high-order convergence in simulating solitons and bi-solitons.

Key words. nonlinear Schrödinger equation, mass conservation, energy conservation, high-order time discretization, post-processing correction, Gauss collocation, finite element method

AMS subject classifications. 65M12, 35K55

1. Introduction.

This paper is concerned with the numerical solution of the nonlinear Schrödinger (NLS) equation in a bounded domain $\Omega \subset \mathbb{R}^d$ with $d \le 3$ under the homogeneous Dirichlet boundary condition, i.e.,

$$\mathrm{i}\partial_t u + \Delta u + f(|u|^2)u = 0 \qquad \text{in } \Omega \times (0,T] \tag{1.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega \times (0,T] \tag{1.1b}$$

$$u|_{t=0} = u^0 \qquad \text{in } \Omega, \tag{1.1c}$$

where $\mathrm{i} = \sqrt{-1}$ is the imaginary unit, $f : \mathbb{R}_+ \to \mathbb{R}$ is a real-valued function as the derivative of some potential function $F : \mathbb{R}_+ \to \mathbb{R}$, and u^0 is a given initial value of the complex-valued solution. Typical examples for the nonlinearity are $f(|u|^2)u = \mu|u|^{q-1}u$ with $\mu \in \mathbb{R}$ and $q > 1$. The case $\mu > 0$ is often referred to as the self-focusing model, for which the solution will blow up in finite time when the initial energy is negative (for example, see [28]), and $\mu < 0$ is referred to as the self-defocusing model. As an important physical model in science and engineering, the NLS equation (1.1) is capable of describing the nonlinear dispersive waves in the modeling of the Bose-Einstein condensate [3, 12, 19], the nonlinear optics [11, 28], the deep-water modulation [22, 32], and other applications. Accordingly, the numerical computation of the NLS equation has been extensively studied with different methods, including finite difference methods [5–8], splitting methods [14, 21, 29], and finite element methods (FEMs) [10, 16–18, 20, 30, 31, 33].

It is well known that the solution of the NLS equation has conserved mass and energy, i.e., $M[u(t)] = M[u^0]$ and $E[u(t)] = E[u^0]$ for all $t \in (0,T]$, where

$$M[u] = \frac{1}{2}\int_\Omega |u|^2 \mathrm{d}x$$

and

$$E[u] = \frac{1}{2}\int_\Omega \Big(|\nabla u|^2 - F(|u|^2)\Big)\mathrm{d}x$$

*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong.
E-mail address: genming.bai@connect.polyu.hk, jiashun.hu@polyu.edu.hk, buyang.li@polyu.edu.hk

denote the mass and energy of the solution, respectively. Correspondingly, it is desirable to develop effective mass- and energy-conserving computational methods for the NLS equation for the long-time simulation or capturing the blow-up phenomena. The most widely used mass- and energy-conserving method for the NLS equation is the modified Crank-Nicolson method studied by Sanz-Serna [23], Akrivis & Dougalis [2], Wang [31], Henning & Peterseim [15], and so on. The method was initially constructed for the NLS equation with power nonlinearities and then generalized to the NLS equation with external potentials in [4, 15]. It was also extended to linearly implicit methods in [8, 9], known as the relaxation schemes. All these methods were based on the Crank–Nicolson scheme and therefore have second-order convergence in time. Recently, the scalar auxiliary variable (SAV) approach was introduced in [26, 27] for constructing energy-decaying methods for dissipative equations. Based on the scalar auxiliary variable (SAV) formulation of the NLS equation, a family of higher-order schemes was developed in [13] to conserve the mass and the SAV energy. However, there are still no high-order methods which conserve the original mass and energy simultaneously.

The objective of this paper is to construct a class of high-order mass- and energy-conserving methods for the NLS equation based on a post-processing correction procedure at every time level. For illustration, we present the algorithm in combination with a Gauss collocation FEM, but remark that this post-processing correction procedure may also be combined with other numerical methods to yield high-order mass- and energy-conserving discretizations of the NLS equation.

More specifically, for a given numerical solution u_h^{n-1} in a finite element space S_h, we compute $u_h(t) = \sum_{j=0}^{k} \phi_j t^j$ for $t \in [t_{n-1}, t_n]$ by a fully discrete FEM with Gauss collocation in time, with $\phi_j \in S_h$ being unknown functions to be determined in the algorithm, and then define the numerical solution at $t = t_n$ as

$$u_h^n = \alpha u_h(t_n) + \beta e^{\mathrm{i}\gamma} u_{h,\perp}(t_n), \tag{1.2}$$

where

$$u_{h,\perp}(t_n) := \hat{u}_h(t_n) - \frac{(\hat{u}_h(t_n), u_h(t_n))}{((-\Delta_h)^{-1} u_h(t_n), u_h(t_n))} (-\Delta_h)^{-1} u_h(t_n)$$

is a function orthogonal to $u_h(t_n)$ with respect to the L^2 inner product, defined in terms of the function

$$\hat{u}_h(t_n) := \Delta_h^{-1} P_h[\Delta_h u_h(t_n) + f(|u_h(t_n)|^2) u_h(t_n)],$$

which represents the direction in which the energy changes fastest (with Δ_h and P_h denoting the discrete Laplacian and L^2-orthogonal projection onto the finite element space S_h). The constants $\alpha, \beta \in \mathbb{R}$ will be chosen to guarantee that the numerical solution u_h^n defined by (1.2) conserves the mass and energy simultaneously. The parameter $\gamma \in [0, 2\pi]$ will be chosen to guarantee that the algebraic system governing (α, β) is uniquely solvable in a neighborhood of $(1, 0)$.

The existence, uniqueness and optimal-order convergence of the numerical solutions are established by incorporating the accumulation errors arising from the post-processing correction procedure at all time levels. Since the algorithm consists of two stages on each subinterval $I_n = [t_{n-1}, t_n]$, the error analysis is split into two parts to address the Gauss collocation FEM and the post-processing correction procedure, respectively. The error estimates for the Gauss collocation FEM basically follow from

the results established in [13] and are therefore stated without proof. The construction and analysis of the post-processing correction procedure are the main contributions of this paper. In particular, when the initial state u^0 is not an eigenfunction of the NLS operator $\mathcal{S} : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega)$ defined by

$$\mathcal{S}v = \Delta v + f(|v|^2)v, \tag{1.3}$$

the algebraic system governing $(\alpha, \beta) \in \mathbb{R}^2$ has a unique root in a neighborhood of $(1, 0)$, and the numerical solutions defined by (1.2) converge to the exact solution in $L^\infty(0, T; H^1(\Omega))$ with the optimal order. When the initial state u^0 is an eigenfunction of the NLS operator, i.e., $\mathcal{S}u^0 = \lambda u^0$, the solution of the NLS equation can be expressed analytically as $u(x, t) = e^{\mathrm{i}\lambda t}u^0(x)$, which is a standing wave stationary up to phase changes. This trivial case is excluded from our consideration.

The rest of this article is organized as follows. In Section 2, we present the two-stage algorithm for the NLS equation, which consists of a Gauss collocation FEM and a post-processing correction procedure at every time level. Then we present the main results concerning the solvability and the error of the proposed algorithm. The ideas which we use to construct the post-processing algorithm are revealed through the analysis of the algorithm in Section 3. Several numerical examples are presented in Section 4 to illustrate the performance of the proposed algorithm in conserving the mass and energy, as well as its high-order convergence in simulating solitons and bi-solitons.

2. The algorithm and main results. In this section, we present the notation to be used in this article and the new algorithm for the NLS equation, and then state the main results on the solvability and error estimates of the proposed algorithm. To avoid considering the approximation of a curved boundary by isoparametric mesh in the FEM, we assume that Ω is a convex polyhedral domain in $\mathbb{R}^d$ with $d \in \{1, 2, 3\}$.

2.1. Finite element spaces and discrete operators. We denote by $(\cdot, \cdot)$ and $\|\cdot\|$ the sesquilinear inner product and norm of the complex-valued Hilbert space $L^2(\Omega)$, i.e.,

$$(u, v) := \int_\Omega u\,\overline{v}\,\mathrm{d}x \quad \text{and} \quad \|u\| = \sqrt{\int_\Omega |u|^2\,\mathrm{d}x}.$$

For $s \geq 0$, we denote by $H^s(\Omega)$ the conventional Sobolev space of functions on Ω, and denote by $H_0^1(\Omega)$ the subspace of $H^1(\Omega)$ consisting of functions with zero traces on the boundary. The norms of $H^s(\Omega)$ and $L^2(\Omega)$ will be denoted by $\|\cdot\|_{H^s}$ and $\|\cdot\|_{L^2}$ by relaxing their dependence on Ω.

For finite element discretization in space, we introduce a shape-regular and quasi-uniform triangulation $\mathcal{T}_h$ of Ω with mesh size $h \in (0, 1]$, and denote by $S_h \subset H_0^1(\Omega)$ the complex-valued Lagrange finite element space of degree $p \geq 1$ subject to the triangulation $\mathcal{T}_h$.

For time discretization, we divide the time interval $[0, T]$ into subintervals $I_n = [t_{n-1}, t_n]$, $n = 1, \ldots, N$, with $t_n = n\tau$ and stepsize $\tau = T/N \in (0, 1]$. For $k \geq 1$ we denote by $\mathbb{P}^k$ the space of polynomials of degree $\leq k$ in time. For any function space $X \subset L^2(\Omega)$, the tensor product space $\mathbb{P}^k \otimes X$ is defined as

$$\mathbb{P}^k \otimes X = \mathrm{span}\Big\{ \sum_{j=0}^{k} t^j v_j : v_j \in X \Big\}.$$

The discrete Laplacian operator on the finite element space S_h is defined as the unique linear operator $\Delta_h : S_h \to S_h$ satisfying the following relation:

$$(\Delta_h v_h, w_h) = -(\nabla v_h, \nabla w_h), \quad \forall w_h \in S_h. \tag{2.1}$$

From the definition one sees that $-\Delta_h$ is a symmetric positive definite operator on the conformal finite element space S_h, and therefore its inverse operator $(-\Delta_h)^{-1} : S_h \to S_h$ exists (also being symmetric and positive definite). The discrete NLS operator $\mathcal{S}_h : S_h \to S_h$ is defined as

$$\mathcal{S}_h v_h := P_h[\Delta_h v_h + f(|v_h|^2)v_h] \quad \forall\, v_h \in S_h, \tag{2.2}$$

where P_h denotes the L^2-orthogonal projection from $L^2(\Omega)$ onto its finite element subspace S_h, i.e.,

$$(w - P_h w, v_h) = 0 \quad \forall\, v_h \in S_h,\ w \in L^2(\Omega). \tag{2.3}$$

2.2. Gauss collocation with post-processing correction. Let c_j and w_j, $j = 1, \ldots, k$, be the nodes and weights of the k-point Gauss quadrature rule on the standard interval $[-1, 1]$ (see [25, Table 3.1]), and let $t_{nj} = t_{n-1} + (1 + c_j)\tau/2$ be the Gauss points on the transformed interval $I_n = [t_{n-1}, t_n]$.

Let $u_h^0 := I_h u_0$, where I_h denotes the Lagrange interpolation operator onto S_h. For any given $u_h^{n-1} \in S_h$, we compute $u_h|_{I_n} \in \mathbb{P}^k \otimes S_h$ and $u_h^n \in S_h$ as follows.

1. Find $u_h|_{I_n} \in \mathbb{P}^k \otimes S_h$ satisfying the following equations for all test functions $v_{hj} \in S_h$ and $j = 1, \ldots, k$:

$$(\mathrm{i}\partial_t u_h(t_{nj}), v_{hj}) - (\nabla u_h(t_{nj}), \nabla v_{hj}) + \big(f(|u_h(t_{nj})|^2)u_h(t_{nj}), v_{hj}\big) = 0, \tag{2.4a}$$

$$u_h(t_{n-1}) = u_h^{n-1}. \tag{2.4b}$$

Remark 2.1. The equations are obtained by collocating the NLS equation at Gauss points $t_{nj} \in (t_{n-1}, t_n)$, $j = 1, \ldots, k$, with an initial condition at t_{n-1}.

2. Compute

$$\hat{u}_h(t_n) = \Delta_h^{-1}\mathcal{S}_h u_h(t_n), \tag{2.5a}$$

$$u_{h,\perp}(t_n) = \hat{u}_h(t_n) - \frac{(\hat{u}_h(t_n), u_h(t_n))}{((-\Delta_h)^{-1}u_h(t_n), u_h(t_n))}(-\Delta_h)^{-1}u_h(t_n), \tag{2.5b}$$

$$\gamma := \arg\big((\nabla u_h(t_n), \nabla u_{h,\perp}(t_n)) - (f(|u_h(t_n)|^2)u_h(t_n), u_{h,\perp}(t_n))\big) \in [0, 2\pi), \tag{2.5c}$$

and then determine α and β through the following two algebraic equations:

$$\frac{1}{2}\|u_h(t_n)\|_{L^2}^2\alpha^2 + \frac{1}{2}\|u_{h,\perp}(t_n)\|_{L^2}^2\beta^2 = M[u_h^{n-1}], \tag{2.6a}$$

$$\frac{1}{2}\|\nabla u_h(t_n)\|_{L^2}^2\alpha^2 + \mathrm{Re}(\nabla u_h(t_n), e^{\mathrm{i}\gamma}\nabla u_{h,\perp}(t_n))\alpha\beta + \frac{1}{2}\|\nabla u_{h,\perp}(t_n)\|_{L^2}^2\beta^2 - \frac{1}{2}\int_\Omega F(|\alpha u_h(t_n) + \beta e^{\mathrm{i}\gamma}u_{h,\perp}(t_n)|^2)\mathrm{d}x = E[u_h^{n-1}]. \tag{2.6b}$$

Remark 2.2. The left-hand sides of the two algebraic equations in (2.6) are the mass and energy of the function $\alpha u_h(t_n) + \beta e^{\mathrm{i}\gamma} u_{h,\perp}(t_n)$, respectively. The function $u_{h,\perp}(t_n)$ is defined to be L^2-orthogonal to $u_h(t_n)$ so that the mass of $\alpha u_h(t_n) + \beta e^{\mathrm{i}\gamma} u_{h,\perp}(t_n)$ has a simplified form (without the cross-product term), as shown on the left-hand side of (2.6a). The definition of $\hat{u}_h(t_n)$ is to regularize the post-processing correction. The solvability of (2.6) can be proved by using the inverse function theorem, as discussed in Section 3.2.

3. Set

$$u_h^n = \alpha u_h(t_n) + \beta e^{\mathrm{i}\gamma} u_{h,\perp}(t_n). \tag{2.7}$$

Remark 2.3. This guarantees the mass and energy conservations of the numerical solution, i.e., $M[u_h^n] = M[u_h^{n-1}]$ and $E[u_h^n] = E[u_h^{n-1}]$, as a result of (2.6).

2.3. Solvability and convergence of the algorithm. The solvability and convergence of the proposed algorithm in (2.4)–(2.7) are guaranteed by the following theorem.

Theorem 2.1. *We assume that $f(|u|^2)u$ is sufficiently smooth with respect to* $\mathrm{Re}(u)$ *and* $\mathrm{Im}(u)$*, and the exact solution of the NLS equation* (1.1) *is sufficiently smooth, with initial value u^0 not being an eigenfunction of the NLS operator $\mathcal{S}$. Then there exist $h_0 > 0$ and $\tau_0 > 0$ such that for all $h \le h_0$, $\tau \le \tau_0$ and $n = 1, \cdots, N$, the following results hold:*

(1) *The nonlinear systems* (2.4) *and* (2.6) *are uniquely solvable (in a neighborhood of the exact solution, see the discussions in Section 3).*

(2) *The mass and energy are conserved, i.e.,*

$$M(u_h^n) = M(u_h^0) \quad \textit{and} \quad E(u_h^n) = E(u_h^0). \tag{2.8}$$

(3) *The following error estimate holds:*

$$\max_{t\in[0,T]} \|u_h(t) - u(t)\|_{H^1} \le C(h^p + \tau^{k+1}). \tag{2.9}$$

The ideas which we use to construct the post-processing algorithm are revealed in the proof of Theorem 2.1, including advantages of the current choice of γ, $\hat{u}_{h,\perp}(t_n)$ and $u_{h,\perp}(t_n)$.

When the initial state u^0 is an eigenfunction of the NLS operator, i.e., $\mathcal{S}u^0 = \lambda u^0$, the solution of the NLS equation can be expressed analytically as $u(x,t) = e^{\mathrm{i}\lambda t} u^0(x)$, and therefore this case is trivial and excluded from our consideration.

The proof of Theorem 2.1 is presented in the next section. Throughout the proof, we denote by $a \lesssim b$ the statement "$a \le Cb$ for some constant C which is independent of τ, h and n, m".

3. Proof of Theorem 2.1. In view of [13, Eq. (3.19)-(3.21)], we define the temporal Ritz projection on the time interval $[t_{n-1}, t_n]$ as follows

$$R_\tau^n u(t) = u(t_{n-1}) + \int_{t_{n-1}}^{t} P_\tau^n \partial_s u(s) \mathrm{d}s, \tag{3.1}$$

where P_τ^n denotes the L^2 projection operator of $L^2(I_n; L^2(\Omega))$ onto its closed subspace $\mathbb{P}^{k-1} \otimes L^2(\Omega)$. Denote the spatial Ritz projection by R_h and then we define $u_h^*(t) = R_\tau^n R_h u(t)$ for $t \in [t_{n-1}, t_n]$ and $n = 1, ..., N$.

We consider the following decomposition

$$u_h(t) - u(t) = e_h(t) + u_h^*(t) - u(t),$$

with $e_h(t) = u_h(t) - u_h^*(t) = u_h(t) - R_\tau^n R_h u(t)$. Since $u_h(t)$ is discontinuous at t_n (due to the post-processing correction procedure), we denote $e_h^n = \lim_{t \downarrow t_n} e_h(t)$.

The proof is based on mathematical induction on n, by assuming that the following results hold for $n = 0, \ldots, m-1$:

$$M[u_h^n] = M[u_h^0] \quad \text{and} \quad E[u_h^n] = E[u_h^0], \tag{3.2}$$

$$\|e_h^n\|_{H^1} \le \tau^{k+\frac{1}{2}} + h^{p-\frac{1}{4}} \quad \text{and} \quad \|e_h^n\|_{L^\infty} \le 1. \tag{3.3}$$

We are going to prove that the numerical solution $u_h(t)$, $t \in [t_{m-1}, t_m)$, and u_h^m are well defined, satisfying (3.2)–(3.3) for $n = m$. This would complete the mathematical induction on m.

The subsequent proof consists of three parts, i.e., the analyses for the Gauss collocation FEM, the solvability of the post-processing algebraic system, and the error estimates for post-processing correction procedure.

3.1. Solvability and error of the Gauss collocation FEM. Given $v_h \in \mathbb{P}^k \otimes S_h$, testing (2.4a) by $v_{hj} = \frac{\tau}{2} v_h(t_{nj}) w_j$ and summing up the results for $j = 1, \ldots, k$, using the property of Gauss quadrature as in [13], we obtain the following integral identity:

$$\int_{I_n} \big(\mathrm{i}\partial_t u_h, v_h\big)\mathrm{d}t - \int_{I_n} (\nabla u_h, \nabla P_\tau^n v_h)\mathrm{d}t \tag{3.4}$$

$$+ \frac{\tau}{2} \sum_{j=1}^{k} w_j (f(|u_h(t_{nj})|^2) u_h(t_{nj}), v_h(t_{nj})) = 0 \qquad \forall\, v_h \in \mathbb{P}^k \otimes S_h. \tag{3.5}$$

This integral formulation is crucial for the analysis and construction of mass and energy conserving methods for the NLS equations. In particular, by choosing $v_h = u_h$ in (3.4) and considering the imaginary part of the result, we obtain the mass conservation for $u_h(t)|_{I_n}$:

$$\frac{1}{2}\|u_h(t_n)\|_{L^2}^2 = \frac{1}{2}\|u_h^{n-1}\|_{L^2}^2. \tag{3.6}$$

However, the energy conservation may be lost and has to be recovered by the post-processing correction procedure in (2.5)–(2.6).

The existence, uniqueness and error estimates for the numerical solutions of the standard Gauss collocation FEM in (2.4) on the subinterval $I_n = [t_{n-1}, t_n]$ can be proved by using the Schaefer's fixed point theorem similarly as the analysis in [13]. Since the analysis is almost the same as [13], we present the results in the following lemma and omit the proof.

Lemma 3.1. *Under the assumptions of Theorem 2.1, if* (3.3) *holds for* $n = 1, \cdots, m-1$, *then there exist positive constants* τ_1 *and* h_1, *such that for* $\tau \le \tau_1$ *and* $h \le h_1$, *the nonlinear system in* (2.4) *has a unique solution* $u_h|_{I_n} \in \mathbb{P}^k \otimes S_h$ *satisfying*

$$\max_{1 \le j \le k} (\|e_h(t_{mj})\|_{L^\infty} + \|e_h(t_{mj})\|_{H^1}) \le 1. \tag{3.7}$$

Moreover, the following estimates hold for $n = 1, \ldots, m$:

$$\|e_h\|^2_{L^\infty(I_n;H^1)} \lesssim \|e_h^{n-1}\|^2_{H^1} + \tau^2(\tau^{k+1} + h^p)^2, \tag{3.8}$$

$$\max_{1\le j\le k} \|e_h(t_{nj})\|_{L^\infty} \lesssim \min\{h^{-1/2}, \tau^{-1/2}\} \left(\|e_h^{n-1}\|_{H^1} + \tau^{k+1} + h^p\right), \tag{3.9}$$

$$\|e_h(t_n)\|^2_{H^1} - \|e_h^{n-1}\|^2_{H^1} \lesssim \tau\|e_h^{n-1}\|^2_{H^1} + \tau(\tau^{k+1} + h^p)^2, \tag{3.10}$$

$$\|\partial_t e_h\|^2_{L^2(I_n;H^{-1})} \lesssim \tau\|e_h\|^2_{L^\infty(I_n;H^1)} + \tau(\tau^{k+1} + h^p)^2. \tag{3.11}$$

The analysis of the solvability and error of the post-processing correction procedure requires the numerical solution to be bounded at t_n. This is not included in the estimates above and therefore should be estimated separately. To this end, we note that the Gauss points t_{nj}, $j = 1, \ldots, k$, are symmetrically distributed in the subinterval $I_n = [t_{n-1}, t_n]$. As a result, $e_h(t_m)$ can be represented as a linear combination of e_h^{m-1} and $e_h(t_{mj})$ in the following form:

$$e_h(t_m) = \pm e_h^{m-1} + \sum_{j=1}^{k} \beta_j e_h(t_{mj}),$$

where the coefficient of e_h^{m-1} has amplitude 1 due to the symmetry of the Gauss points, and the constants β_j depend only on k. This expression of $e_h(t_m)$ implies that

$$\|e_h(t_m)\|_{L^\infty} \le \|e_h^{m-1}\|_{L^\infty} + C \max_{1\le j\le k} \|e_h(t_{mj})\|_{L^\infty}. \tag{3.12}$$

Substituting (3.3) and (3.9) into (3.12) yields, for sufficiently small τ, h and $p, k \ge 1$,

$$\|e_h(t_m)\|_{L^\infty} \le 1 + C(\tau + h^{\frac{1}{4}}) \le 2. \tag{3.13}$$

Similarly, substituting the induction hypotheses (3.3) into (3.8) and (3.11) yields

$$\|e_h\|_{L^\infty(I_m;H^1)} + \|\partial_t e_h\|_{L^\infty(I_m;H^{-1})} \lesssim 1. \tag{3.14}$$

This means that $u_h(t_m)$ is bounded in $L^\infty(\Omega)$ and $H^1(\Omega)$ (uniformly with respect to τ and h) before it is modified by the post-processing procedure.

3.2. Solvability of the post-processing nonlinear algebraic system. In this subsection, we discuss the solvability of the nonlinear algebraic system in (2.6) for $n = m$ under induction hypotheses (3.2)–(3.3).

Let $\Theta = (\alpha, \beta)$ and consider the following function, which represents the loss of mass and energy of u_h^n defined by (2.7):

$$\begin{aligned} &G(\Theta, u_h(t_n)) \\ &= \begin{pmatrix} \dfrac{1}{2}\|u_h(t_n)\|^2_{L^2}\alpha^2 + \dfrac{1}{2}\|u_{h,\perp}(t_n)\|^2_{L^2}\beta^2 - M[u_h^{n-1}] \\ \dfrac{1}{2}\|\nabla u_h(t_n)\|^2_{L^2}\alpha^2 + \mathrm{Re}(\nabla u_h(t_n), e^{\mathrm{i}\gamma}\nabla u_{h,\perp}(t_n))\alpha\beta + \dfrac{1}{2}\|\nabla u_{h,\perp}(t_n)\|^2_{L^2}\beta^2 \\ \qquad\qquad - \dfrac{1}{2}\displaystyle\int_\Omega F(|\alpha u_h(t_n) + \beta e^{\mathrm{i}\gamma} u_{h,\perp}(t_n)|^2)\mathrm{d}x - E[u_h^{n-1}] \end{pmatrix}, \end{aligned}$$

where $u_{h,\perp}(t_n)$ and γ are uniquely determined by $u_h(t_n)$ through (2.5). We shall apply the inverse function theorem to show that the algebraic equation

$$G(\Theta, u_h(t_n)) = 0 \tag{3.15}$$

has a unique root Θ in a neighborhood of $\Theta_0 = (1, 0)$.

To this end, we consider the gradient $A = \nabla_\Theta G(\Theta, u_h(t_n))\big|_{\Theta=(1,0)}$, which is a lower triangular matrix in $\mathbb{R}^{2\times 2}$ with the following entries:

$$A_{11} = \|u_h(t_n)\|_{L^2}^2, \tag{3.16a}$$

$$A_{12} = 0, \tag{3.16b}$$

$$A_{21} = \|\nabla u_h(t_n)\|_{L^2}^2 - (f(|u_h(t_n)|^2)u_h(t_n), u_h(t_n)), \tag{3.16c}$$

$$\begin{aligned} A_{22} &= \mathrm{Re}\left(e^{-\mathrm{i}\gamma}\left[(\nabla u_h(t_n), \nabla u_{h,\perp}(t_n)) - (f(|u_h(t_n)|^2)u_h(t_n), u_{h,\perp}(t_n))\right]\right) \\ &= |(\mathcal{S}_h u_h(t_n), u_{h,\perp}(t_n))|, \end{aligned} \tag{3.16d}$$

where the definition of γ and Δ_h and $u_{h,\perp} \in S_h$ are used in (3.16d). The determinant of matrix A is given by

$$\det(\nabla_\Theta G(\Theta, u_h(t_n)))\big|_{\Theta=(1,0)} = \|u_h(t_n)\|_{L^2}^2 \big|\big(\mathcal{S}_h u_h(t_n), u_{h,\perp}(t_n)\big)\big|. \tag{3.17}$$

Due to the induction hypothesis in (3.2) and the mass conservation property of the Gauss collocation FEM, we have

$$\|u_h(t_n)\|_{L^2} = \|u_h^0\|_{L^2} \quad \text{for } n = 1, \ldots, m. \tag{3.18}$$

Therefore, $\det(\nabla_\Theta G(\Theta, u_h(t_n)))\big|_{\Theta=(1,0)} \neq 0$ if

$$\big|\big(\mathcal{S}_h u_h(t_n), u_{h,\perp}(t_n)\big)\big| \neq 0. \tag{3.19}$$

This is proved in the following lemma for the case that u^0 is not an eigenfunction of the NLS operator, as assumed in Theorem 2.1.

Lemma 3.2. *If $u \in C([0, T], H^2(\Omega))$ is the solution of the NLS equation* (1.1) *with initial value $u^0 \notin \mathcal{E}_\mathcal{S}$, where $\mathcal{E}_\mathcal{S}$ denotes the set of eigenfunctions of the NLS operator defined in* (1.3)*, then there exists $\kappa > 0$ such that*

$$\inf_{t\in[0,T]} \inf_{v\in\mathcal{E}_\mathcal{S}} \|u(t) - v\|_{H^2} \geq \kappa. \tag{3.20}$$

There exist positive constants h_2, ϵ and δ such that for $h \leq h_2$ and $u_h \in S_h$ satisfying

$$\inf_{t\in[0,T]} \|u_h - u(t)\|_{H^1} \leq \epsilon, \quad \|u_h\|_{L^\infty} \leq \|u\|_{L^\infty([0,T],L^\infty)} + 2, \quad \|u_h\|_{L^2} = \|u_h^0\|_{L^2}, \tag{3.21}$$

the following inequalities hold, with $u_{h,\perp}$ defined in the same way in (2.5a)–(2.5b)*,*

$$|(\mathcal{S}_h u_h, u_{h,\perp})| \geq \delta, \tag{3.22}$$

$$|\det(\nabla_\Theta G(\Theta, u_h))|_{\Theta=(1,0)}| \geq \delta \|u_h^0\|_{L^2}^2. \tag{3.23}$$

Proof. Inequality (3.20) can be proved by the method of contradiction. In fact, if there exist sequences $t_i \in [0, T]$ and $v_i \in \mathcal{E}_\mathcal{S}$ such that $\|u(t_i) - v_i\|_{H^2} \to 0$, then by the compactness of $[0, T]$ and by passing to a subsequence if necessary, there exists $t^* \in [0, T]$ such that $t_i \to t^*$. Then $\|u(t_i) - u(t^*)\|_{H^2} \to 0$ and therefore $\|u(t^*) - v_i\|_{H^2} \to 0$ as $i \to \infty$. This implies that $\mathcal{S}v_i \to \mathcal{S}u(t^*)$ in $L^2(\Omega)$.

Since $v_i \in \mathcal{E}_\mathcal{S}$, there exists $\lambda_i \in \mathbb{C}$ such that $\mathcal{S}v_i = \lambda_i v_i$. If λ_i has an accumulation point $\lambda^* \in \mathbb{C}$ then, by passing to a subsequence if necessary, we obtain $\mathcal{S}u(t^*) = \lambda^* u(t^*)$. In this case, the unique solution of (1.1) can be analytically expressed by

$u(t) = e^{\mathrm{i}\lambda^*(t-t^*)}u(t^*)$, which implies that $u(0) = e^{-\mathrm{i}\lambda^* t^*}u(t^*)$ is an eigenfunction of the NLS operator. This leads to a contradiction.

This means that λ_i cannot have an accumulation point in $\mathbb{C}$, and therefore $|\lambda_i| \to +\infty$. By passing to limit in $|\lambda_i|^{-1}\|\mathcal{S}v_i\|_{L^2} = \|v_i\|_{L^2}$ and the L^2 boundedness of $\mathcal{S}v_i$, we obtain that $\|v_i\|_{L^2} \to 0$ and consequently $u(t^*) = 0$. Then the solution over $t \in [0, T]$ is zero, which also leads to a contradiction. Hence (3.20) is proved.

Inequality (3.22) can also be proved by the method of contradiction. If there does not exist $h_2, \epsilon, \delta > 0$ such that (3.22) is valid for all $u_h \in S_h$ satisfying $h \le h_2$ and (3.21). Then there exist $h_i \to 0$, $u_{h_i} \in S_{h_i}$ and $t_i \in [0, T]$ such that the following relations hold:

$$\|u_{h_i} - u(t_i)\|_{H^1} \to 0 \quad \text{and} \quad |(\mathcal{S}_{h_i} u_{h_i}, u_{h_i,\perp})| \to 0. \tag{3.24}$$

By passing to a subsequence if necessary, we may assume that $t_i \to t^*$. Since $u \in C([0, T], H^2(\Omega))$, it follows that

$$\|u_{h_i} - u(t^*)\|_{H^1} \to 0. \tag{3.25}$$

Since $u(t^*) \in H^2(\Omega) \hookrightarrow L^\infty(\Omega)$, it follows that $f(|u(t^*)|^2)u(t^*) \in L^2(\Omega)$ and therefore we can define $\hat{u}(t^*) \in H_0^1(\Omega) \cap H^2(\Omega)$ to be the solution of the following elliptic equation (in the weak formulation):

$$-(\nabla \hat{u}(t^*), \nabla w) = -(\nabla u(t^*), \nabla w) + (f(|u(t^*)|^2)u(t^*), w), \quad \forall w \in H_0^1(\Omega). \tag{3.26}$$

At the discrete level, we define $\hat{u}_{h_i} \in S_h$ to be the solution of

$$-(\nabla \hat{u}_{h_i}, \nabla w_{h_i}) = -(\nabla u_{h_i}, \nabla w_{h_i}) + (f(|u_{h_i}|^2)u_{h_i}, w_{h_i}), \quad \forall w_{h_i} \in S_{h_i}. \tag{3.27}$$

For any $w \in H_0^1(\Omega)$, w_{h_i} can be chosen as its interpolation in S_{h_i}. Since $\{u_{h_i}\}$ is bounded in L^∞, by passing to the limit $h_i \to 0$, we obtain

$$-(\nabla u_{h_i}, \nabla w_{h_i}) + (f(|u_{h_i}|^2)u_{h_i}, w_{h_i}) \to -(\nabla u(t^*), \nabla w) + (f(|u(t^*)|^2)u(t^*), w),$$

which implies that $\|\hat{u}_{h_i} - \hat{u}(t^*)\|_{H^1} \to 0$. The direct computation of (2.5a)–(2.5b) yields

$$(\mathcal{S}_{h_i} u_{h_i}, u_{h_i,\perp}) = (\nabla \hat{u}_{h_i}, \nabla \hat{u}_{h_i}) - \frac{|(\hat{u}_{h_i}, u_{h_i})|^2}{((-\Delta_{h_i})^{-1} u_{h_i}, u_{h_i})}.$$

By passing to limit in the above equation and using (3.24), we obtain

$$0 = \|(-\Delta)^{1/2}\hat{u}(t^*)\|_{L^2}^2 - \frac{|(\hat{u}(t^*), u(t^*))|^2}{||(-\Delta)^{-1/2}u(t^*)||_{L^2}^2}. \tag{3.28}$$

Rewriting $|(\hat{u}(t^*), u(t^*))|$ as $|((-\Delta)^{1/2}\hat{u}(t^*), (-\Delta)^{-1/2}u(t^*))|$, (3.28) can be regarded as the case that the Cauchy–Schwarz inequality reduces to an equality. In this case, there exists a constant λ such that $-\Delta\hat{u}(t^*) = \lambda u(t^*)$. In view of (3.26), it is equivalent to

$$\mathcal{S}u(t^*) = -\lambda u(t^*),$$

which means that $u(t^*) \in \mathcal{E}_\mathcal{S}$ and therefore contradicts (3.20) and ends the proof of (3.22). Inequality (3.23) is an immediate consequence of (3.17) and (3.22). □

Remark 3.1. *Since* (3.21) *can be guaranteed by* (3.3)–(3.9), (3.13) *and* (3.18), *there exist positive constants* τ_2, h_2 *and* δ *depending only on* T *and* u *such that for* $\tau \le \tau_2$ *and* $h \le h_2$, (3.22)–(3.23) *is satisfied by the numerical solution* $u_h(t_n)$ *for* $n = 1, \cdots, m$.

Lemma 3.3. *Under the assumptions of Theorem 2.1, there exist positive constants* τ_3, h_3 *and* $r_0 \in (0,1)$ *(independent of* τ *and* h*) such that for* $\tau \le \tau_3$ *and* $h \le h_3$, *the nonlinear algebraic system in* (2.6) *has a unique root* (α, β) *satisfying* $\sqrt{|\alpha - 1|^2 + |\beta|^2} < r_0$.

Proof. This is equivalent to prove that equation $G(\Theta, u_h(t_n)) = 0$ has a unique root $\Theta \in B_{r_0}(\Theta_0)$. Since $A = \nabla_\Theta G(\Theta, u_h(t_n))\big|_{\Theta=(1,0)}$ is a lower triangle matrix with entries in (3.16), it follows that A^{-1} has the following decomposition:

$$A^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & A_{22}^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -A_{21} & 1 \end{pmatrix} \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & 1 \end{pmatrix}.$$

Based on the decomposition above, the matrix 2-norm of A^{-1} can be estimated by

$$\|A^{-1}\| \le \max\{1, A_{22}^{-1}\} \max\{1, A_{11}^{-1}\} |A_{21} + 2|. \tag{3.29}$$

The mass conservation ensures that A_{11} is a constant, while the lower bound of the determinant in (3.23) guarantees the boundedness of A_{22}^{-1} according to Remark 3.1. The boundedness of A_{21} can be obtained by the L^∞ and H^1 boundedness of $u_h(t_n)$, as shown in (3.3), (3.10) and (3.13). This leads to the following estimates:

$$|A_{21}| \lesssim 1, \quad |A_{11}^{-1}| \lesssim 1, \quad |A_{22}^{-1}| \lesssim 1.$$

Substituting these estimates into (3.29) yields

$$\|A^{-1}\| \lesssim 1.$$

From the boundedness of $\|u_h(t_n)\|_{L^\infty}$ we can also conclude that $\|\nabla^2 G(\Theta, u_h(t_n))\| \lesssim 1$ for Θ in a fixed bounded neighborhood of Θ_0. In particular, there exists a constant $r_0 \in (0,1)$ such that

$$\|A^{-1}\| \sup_{\Theta \in B_{r_0}(\Theta_0)} \|\nabla G(\Theta, u_h(t_n)) - \nabla G(\Theta_0, u_h(t_n))\| \le \frac{1}{2}. \tag{3.30}$$

By the implicit function theorem, there exists a constant $c > 0$ (independent of τ and h) such that

$$\begin{aligned} &\text{equation } G(\Theta, u_h(t_n)) = y \text{ has a unique root } \Theta \in B_{r_0}(\Theta_0) \\ &\qquad\qquad \text{for } \|y - G(\Theta_0, u_h(t_n))\| < c. \end{aligned} \tag{3.31}$$

Note that

$$G(\Theta_0, u_h(t_m)) = \begin{pmatrix} M[u_h(t_m)] - M[u_h^{m-1}] \\ E[u_h(t_m)] - E[u_h^{m-1}] \end{pmatrix} = \begin{pmatrix} 0 \\ E[u_h(t_m)] - E[u_h^{m-1}] \end{pmatrix},$$

where we have used (3.6). Since u_h^0 is the Lagrange interpolation of u^0 and

$$\|u_h(t_m) - u(t_m)\|_{H^1} \le \|e_h\|_{L^\infty(I_m; H^1)} \lesssim \|e_h^{m-1}\|_{H^1} + \tau(\tau^{k+1} + h^p) \lesssim \tau^{k+\frac{1}{2}} + h^{p-\frac{1}{4}},$$

as shown in (3.3) and (3.8), by using the L^∞ boundedness of $u_h(t_m)$ proved in (3.13), we obtain

$$E[u_h(t_m)] - E[u(t_m)] = O(\tau^{k+\frac{1}{2}} + h^{p-\frac{1}{4}}), \quad E[u^0] - E[u_h^0] = O(h^p),$$

and therefore

$$G(\Theta_0, u_h(t_m)) = \begin{pmatrix} 0 \\ O(\tau^{k+\frac{1}{2}} + h^{p-\frac{1}{4}}) \end{pmatrix}. \tag{3.32}$$

As a result, $\|0 - G(\Theta_0, u_h(t_n))\| < c$ for sufficiently small τ and h (say $\tau \le \tau_3$ and $h \le h_3$). In this case, the implicit function theorem implies that equation $G(\Theta, u_h(t_n)) = 0$ has a unique root $\Theta \in B_{r_0}(\Theta_0)$, as discussed in (3.31). □

3.3. Error from the post-processing correction procedure.

Lemma 3.4. *Under the assumptions of Theorem 2.1, for $\tau \le \tau_3$ and $h \le h_3$, the parameters α and β from the post-processing correction procedure satisfy the following estimate:*

$$|1-\alpha| + |\beta| \lesssim \tau \|e_h\|_{L^\infty(I_m;H^1)} + \tau(\tau^{k+1} + h^p).$$

Proof. We recall that in Section 3 we define $u_h^*(t) = R_\tau^n R_h u(t)$ for $t \in [t_{n-1}, t_n]$. By choosing $v_h = \partial_t u_h$ in (3.4) and considering the real part of the result, we find that

$$\begin{aligned} \frac{1}{2}\|\nabla u_h(t_m)\|_{L^2}^2 - \frac{1}{2}\|\nabla u_h^{m-1}\|_{L^2}^2 &= \frac{\tau}{2}\sum_{j=1}^{k} w_j \mathrm{Re}(f(|u_h(t_{mj})|^2)u_h(t_{mj}), \partial_t u_h(t_{mj})) \\ &=: \int_{I_m} \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \frac{1}{2}F(|u_h(t)|^2)\mathrm{d}x\mathrm{d}t + \mathrm{Re}(R_m), \end{aligned} \tag{3.33}$$

where

$$\begin{aligned} R_m &:= \frac{\tau}{2}\sum_{j=1}^{k} w_j (f(|u_h(t_{mj})|^2)u_h(t_{mj}), \partial_t u_h(t_{mj})) - \int_{I_m} (f(|u_h|^2)u_h, \partial_t u_h)\mathrm{d}t \\ &= -\int_{I_m} (f(|u_h|^2)u_h, \partial_t e_h)\mathrm{d}t \\ &\quad - \int_{I_m} (f(|u_h|^2)u_h - f(|u_h^*|^2)u_h^*, \partial_t u_h^*)\mathrm{d}t \\ &\quad + \frac{\tau}{2}\sum_{j=1}^{k} w_j (f(|u_h^*(t_{mj})|^2)u_h^*(t_{mj}), \partial_t u_h^*(t_{mj})) - \int_{I_m} (f(|u_h^*|^2)u_h^*, \partial_t u_h^*)\mathrm{d}t \\ &\quad + \frac{\tau}{2}\sum_{j=1}^{k} w_j [(f(|u_h(t_{mj})|^2)u_h(t_{mj}), \partial_t u_h(t_{mj})) - (f(|u_h^*(t_{mj})|^2)u_h^*(t_{mj}), \partial_t u_h^*(t_{mj}))] \\ &=: R_{m1} + R_{m2} + R_{m3} + R_{m4}. \end{aligned} \tag{3.34}$$

By comparing (3.33) with expression $E[u_h(t_m)] = \frac{1}{2}\int_\Omega \big(|\nabla u_h(t_m)|^2 - F(|u_h(t_m)|^2)\big)\mathrm{d}x$, we find that

$$E(u_h(t_m)) - E(u_h^{m-1}) = \mathrm{Re}(R_m).$$

From the boundedness of $\|u_h\|_{L^\infty(I_m;L^\infty)} + \|u_h\|_{L^\infty(I_m;H^1)}$ and the smoothness of $f(|u|^2)u$ with respect to $\mathrm{Re}(u)$ and $\mathrm{Im}(u)$ (as assumed in Theorem 2.1), we can estimate the difference of the nonlinear term as follows

$$|f(|u_h|^2)u_h - f(|u_h^*|^2)u_h^*| \lesssim |u_h - u_h^*| \lesssim |e_h|.$$

Using the boundedness and estimate above together with (3.8)–(3.11), the following estimates of R_{mj}, $j = 1, 2, 3, 4$, can be derived:

$$\begin{aligned}
|R_{m1}| &= \left| \int_{I_m} (f(|u_h|^2)u_h, \partial_t e_h) \mathrm{d}t \right| \leq \|f(|u_h|^2)u_h\|_{L^2(I_m;H^1)} \|\partial_t e_h\|_{L^2(I_m;H^{-1})} \\
&\lesssim \tau \left(\|e_h\|_{L^\infty(I_m;H^1)} + \tau^{k+1} + h^p \right), \qquad (3.35) \\
|R_{m2}| &= \left| \int_{I_m} (f(|u_h|^2)u_h - f(|u_h^*|^2)u_h^*, \partial_t u_h^*) \mathrm{d}t \right| \lesssim \|e_h\|_{L^2(I_m;L^2)} \|\partial_t u_h^*\|_{L^2(I_m;L^2)} \\
&\lesssim \tau \|e_h\|_{L^\infty(I_m;L^2)}, \qquad (3.36) \\
|R_{m3}| &= \left| \int_{I_m} [(f(|u_h^*|^2)u_h^*, \partial_t u_h^*) - (I_\tau^m(f(|u_h^*|^2)u_h^*), \partial_t u_h^*)] \mathrm{d}t \right| \\
&\lesssim \tau \|f(|u_h^*|^2)u_h^* - I_\tau^m(f(|u_h^*|^2)u_h^*)\|_{L^\infty(I_m,L^2)} \lesssim \tau^{k+2}, \qquad (3.37)
\end{aligned}$$

where the local temporal interpolation operator I_τ^m is defined as $I_\tau^m : C(I_m; L^2(\Omega)) \to \mathbb{P}^k \otimes L^2(\Omega), u(t) \mapsto \sum_{j=0}^k u(t_j)\phi_j(t)$ with $\phi_j(t)$ being the jth Lagrange basis of degree k associated to the node $t = t_{nj}$, and the expression of R_{m3} is obtained by using the property of Gauss quadrature (i.e., it is exact for polynomials of degree $2k-1$ in time). By decomposing R_{m4} into two parts we can estimate it as follows, again using the result in (3.11):

$$\begin{aligned}
|R_{m4}| \leq & \left| \frac{\tau}{2} \sum_{j=1}^k w_j (f(|u_h(t_{mj})|^2)u_h(t_{mj}) - f(|u_h^*(t_{mj})|^2)u_h^*(t_{mj}), \partial_t u_h^*(t_{mj})) \right| \\
& + \left| \frac{\tau}{2} \sum_{j=1}^k w_j (f(|u_h(t_{mj})|^2)u_h(t_{mj}), \partial_t e_h(t_{mj})) \right| \\
\lesssim & \tau^{1/2} \|e_h\|_{L^2(I_m;L^2)} + \tau^{1/2} \|\partial_t e_h\|_{L^2(I_m;H^{-1})} \\
\lesssim & \tau \left(\|e_h\|_{L^\infty(I_m;H^1)} + \tau^{k+1} + h^p \right). \qquad (3.38)
\end{aligned}$$

By substituting (3.35)–(3.38) into expression $E(u_h(t_m)) - E(u_h^{m-1}) = \mathrm{Re}(R_m)$, we obtain

$$|E(u_h(t_m)) - E(u_h^{m-1})| \lesssim \tau \|e_h\|_{L^\infty(I_m;H^1)} + \tau(\tau^{k+1} + h^p). \qquad (3.39)$$

Since $G(\Theta, u_h(t_m)) = (0,0)^T$ and $G(\Theta_0, u_h(t_m)) = (0, E[u_h(t_m)] - E[u_h^{m-1}])^T$, the implicit function theorem implies that

$$\begin{aligned}
\sqrt{|\alpha - 1|^2 + |\beta|^2} = \|\Theta - \Theta_0\| &\lesssim \|G(\Theta, u_h(t_m)) - G(\Theta_0, u_h(t_m))\| \\
&= \|G(\Theta_0, u_h(t_m))\| \\
&= |E[u_h(t_m)] - E[u_h^{m-1}]| \\
&\lesssim \tau \|e_h\|_{L^\infty(I_m;H^1)} + \tau(\tau^{k+1} + h^p).
\end{aligned}$$

This proves the result of Lemma 3.4. ∎

Lemma 3.5. *Under the assumptions of Theorem 2.1, there exist positive constants τ_4 and h_4 such that for $\tau \le \tau_4$ and $h \le h_4$ the error $e_h^n = u_h^n - u_h^*(t_n)$ satisfies the following estimates for $n = 1, \dots, m$:*

$$\|e_h^n\|_{H^1}^2 \le (1+\tau)\|e_h(t_n)\|_{H^1}^2 + C\tau\|e_h\|_{L^\infty(I_m;H^1)}^2 + C\tau(\tau^{k+1}+h^p)^2, \tag{3.40}$$

$$\|e_h^n\|_{L^\infty} \le \|e_h(t_n)\|_{L^\infty} + C\tau\|e_h\|_{L^\infty(I_n;H^1)} + C\tau(\tau^{k+1}+h^p). \tag{3.41}$$

Proof. After the post-processing procedure in (2.7), the error $e_h^n = u_h^n - u_h^*(t_n)$ can be related to $e_h(t_n) = u_h(t_n) - u_h^*(t_n)$ through

$$e_h^n = e_h(t_n) + (\alpha - 1)u_h(t_n) + \beta e^{\mathrm{i}\gamma} u_{h,\perp}(t_n).$$

The boundedness of $\|u_{h,\perp}(t_n)\|_{H^1}$ and $\|u_h(t_n)\|_{H^1}$, as shown in (3.13)–(3.14), implies that

$$\begin{aligned}\|e_h^n\|_{H^1} &\le \|e_h(t_n)\|_{H^1} + C|\alpha - 1| + C|\beta| \\ &\le \|e_h(t_n)\|_{H^1} + C\tau\|e_h\|_{L^\infty(I_n;H^1)} + C\tau(\tau^{k+1}+h^p),\end{aligned}$$

where the last inequality has been proved in Lemma 3.4. By taking square of the inequality and using Young's inequality for the cross-product term, we obtain (3.40).

The boundedness of $\|u_h(t_n)\|_{L^\infty}$, as shown in (3.13), also implies the boundedness of $\|u_{h,\perp}(t_n)\|_{L^\infty}$ defined in (2.5a)–(2.5b). As a result, the following inequality holds:

$$\begin{aligned}\|e_h^n\|_{L^\infty} &\le \|e_h(t_n)\|_{L^\infty} + C|\alpha - 1| + C|\beta| \\ &\le \|e_h(t_n)\|_{L^\infty} + C\tau\|e_h\|_{L^\infty(I_n;H^1)} + C\tau(\tau^{k+1}+h^p),\end{aligned}$$

where the last inequality has been proved in Lemma 3.4. This proves (3.41). ∎

3.4. Completion of the proof. By substituting (3.8)–(3.10) and (3.12) into (3.40)–(3.41), we obtain the following estimates for $n = 1, \dots, m$:

$$\|e_h^n\|_{H^1}^2 \le (1+C\tau)\|e_h^{n-1}\|_{H^1}^2 + C\tau(\tau^{k+1}+h^p)^2, \tag{3.42}$$

$$\begin{aligned}\|e_h^n\|_{L^\infty} &\le \|e_h^{n-1}\|_{L^\infty} + C\max_{1\le j\le k}\|e_h(t_{nj})\|_{L^\infty} + C\tau\|e_h^{n-1}\|_{H^1} + C\tau(\tau^{k+1}+h^p) \\ &\le \|e_h^{n-1}\|_{L^\infty} + C\min\{h^{-1/2}, \tau^{-1/2}\}\left(\|e_h^{n-1}\|_{H^1} + \tau^{k+1} + h^p\right) \\ &\quad + C\tau\|e_h^{n-1}\|_{H^1} + C\tau(\tau^{k+1}+h^p).\end{aligned} \tag{3.43}$$

By applying the discrete Gronwall's inequality to (3.42), we obtain the following result for $n = 1, \dots, m$:

$$\|e_h^n\|_{H^1} \le C(\tau^{k+1}+h^p). \tag{3.44}$$

From (3.8) we also obtain

$$\|e_h\|_{L^\infty(I_n;H^1)}^2 \le C(\tau^{k+1}+h^p). \tag{3.45}$$

Then, using the discrete Sobolev inequality and inverse inequality of the finite element space and (3.44), we have

$$\|e_h^n\|_{L^\infty} \le Ch^{-\frac{1}{2}}\|e_h^n\|_{H^1} \le 1 \quad \text{if } \tau^{k+1} \le h^{\frac{2}{3}} \text{ and } h \text{ is sufficiently small.} \tag{3.46}$$

If $\tau^{k+1} \geq h^{\frac{2}{3}}$ then (3.43) implies that

$$\begin{aligned}\|e_h^n\|_{L^\infty} &\leq \|e_h^{n-1}\|_{L^\infty} + C(\tau^{k+\frac{1}{2}} + h^{p-\frac{1}{2}}) \\ &\leq \|e_h^{n-1}\|_{L^\infty} + C(\tau^{\frac{3}{2}} + h^{\frac{1}{2}}) \qquad \text{(as a result of } k \geq 1 \text{ and } p \geq 1) \\ &\leq \|e_h^{n-1}\|_{L^\infty} + C\tau^{\frac{3}{2}} \qquad \text{(as a result of } \tau^{k+1} \geq h^{\frac{2}{3}} \text{ and } k \geq 1).\end{aligned}$$

By iterating this inequality for $n = 1, \ldots, m$, we obtain

$$\|e_h^m\|_{L^\infty} \leq \|e_h^0\|_{L^\infty} + C\tau^{\frac{1}{2}} \leq Ch^p + C\tau^{\frac{1}{2}} \leq 1 \quad \text{(for sufficiently small } \tau \text{ and } h).$$

For sufficiently small τ and h, (3.44) reduces to

$$\|e_h^m\|_{H^1} \leq \tau^{k+\frac{1}{2}} + h^{p-\frac{1}{4}}.$$

This proves (3.3) for $n = m$ and therefore completes the mathematical induction on m. Therefore, the nonlinear systems in (2.4) and (2.6) are uniquely solvable (in a neighborhood of the exact solution) with error estimate (3.45) for $n = 1, \ldots, N$. ∎

4. Numerical results. In this section, we present several numerical examples to illustrate the performance of the proposed method for the NLS equation, including the optimal convergence order and the conservation of mass and energy. The numerical experiments are performed by using the open-source high-performance finite element software NGSolve; see [24].

Example 4.1 (A one-dimensional soliton). We consider the one-dimensional focusing NLS equation

$$\begin{aligned}&\mathrm{i}\partial_t u + \partial_{xx} u + 2|u|^2 u = 0 && \text{in } (-L, L) \times (0, T], && (4.1)\\ &u = 0 && \text{at } \pm L, && (4.2)\\ &u|_{t=0} = u_0 && \text{in } (-L, L), && (4.3)\end{aligned}$$

which is an approximation of the NLS equation on $\mathbb{R}$. Since it is difficult to construct an analytical expression for the solution of the NLS equation in a bounded domain, we consider an analytical expression of the solution to the NLS equation in $\mathbb{R}^d$ with exponential decay at infinity and choose a moderately large domain so that the solution is practically zero on the boundary of the domain up to round-off errors. This analytical expression of the solution is used to test the error of the numerical computations and the convergence orders of the proposed method. In particular, we consider the bright soliton solution in [1] with the following analytical expression:

$$u(x, t) = \operatorname{sech}(x + 4t) \exp(-\mathrm{i}(2x + 3t)). \tag{4.4}$$

Since $|u(x, t)|$ decays rapidly to zero for $t \in [0, T]$, choosing $T = 1$ and $L = 40$ is sufficient to make the error of domain truncation negligible (up to round-off errors) compared to the temporal and spatial discretization errors in the convergence test.

The Newton iteration is used to solve the nonlinear algebraic system in (2.6) with a tolerance error of 10^{-10} in the H^1 norm. We compute the modified u_h^n (2.7) by solving (α, β) from the algebraic system (2.6) by the Newton iteration with initial guess $(\alpha, \beta) = (1, 0)$ and tolerance 10^{-10}. The $L^\infty(0, T, H^1)$ error between the numerical solution and (4.4) is measured by

$$H^1 \text{ error} = \max_{0 \leq n \leq N} \max_{1 \leq j \leq k} \|u_h(x, t_{nj}) - P_{p+2} u(x, t_{nj})\|_{H^1}, \tag{4.5}$$

where P_{p+2} denotes the L^2 projection to finite element space of degree $p+2$ (two degrees higher than S_h).

The errors from the spatial discretizations are presented in Figure 4.1, where we choose $k = 3$ and $\tau = 1/1000$ so that the time discretization errors are negligible in observing the spatial convergence orders. The numerical results in Figure 4.1 show that the errors from the spatial discretizations are $\mathcal{O}(h^p)$ in the $L^\infty(0, T, H^1)$ norm, which is consistent with the result proved in Theorem 2.1.

The errors from the time discretizations are presented in Figure 4.2, where we choose $p = 3$ and $h = 2L/4000$ so that the spatial discretization errors are negligible in observing the temporal convergence orders. The numerical results in Figure 4.2 show that the errors from the time discretizations are $\mathcal{O}(\tau^{k+1})$ in the $L^\infty(0, T, H^1)$ norm, which is consistent with the result proved in Theorem 2.1.

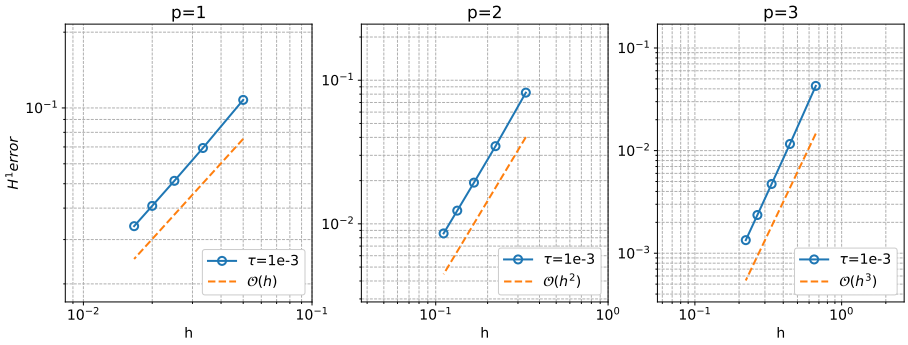


Fig. 4.1. *Space discretization errors in the $L^\infty(0, T, H^1)$ norm (Example 4.1).*

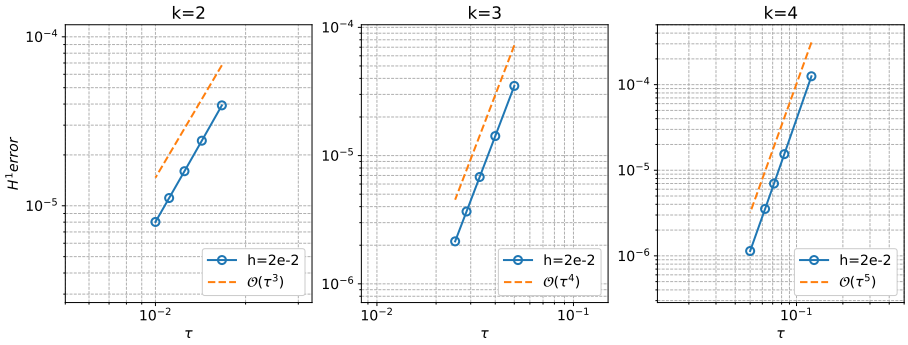
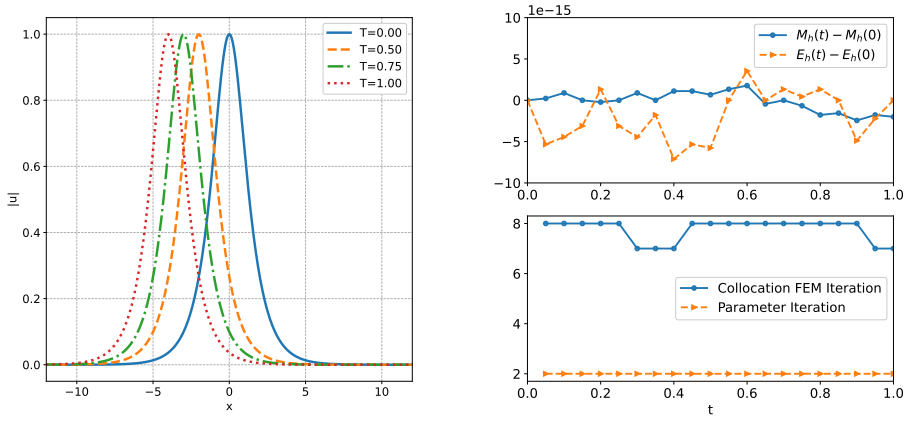


Fig. 4.2. *Time discretization errors in the $L^\infty(0, T, H^1)$ norm (Example 4.1).*

The evolution of the amplitude of the numerical solution and the conservation of mass and energy are examined in Figure 4.3 with $L = 40$, $k = p = 3$, $T = 1$, $\tau = 1/20$ and $h = 1/5$. Figure 4.3 (left) shows that no visible mass loss during the evolution, and the mass and energy are conserved up to $O(10^{-15})$, which is much lower than the errors from the temporal and spatial discretizations. This shows the conservation of mass and energy of the proposed method.

The numbers of Newton iterations for solving the Gauss collocation FEM and the nonlinear algebraic system for (α, β) are presented in Figure 4.3 (right below).

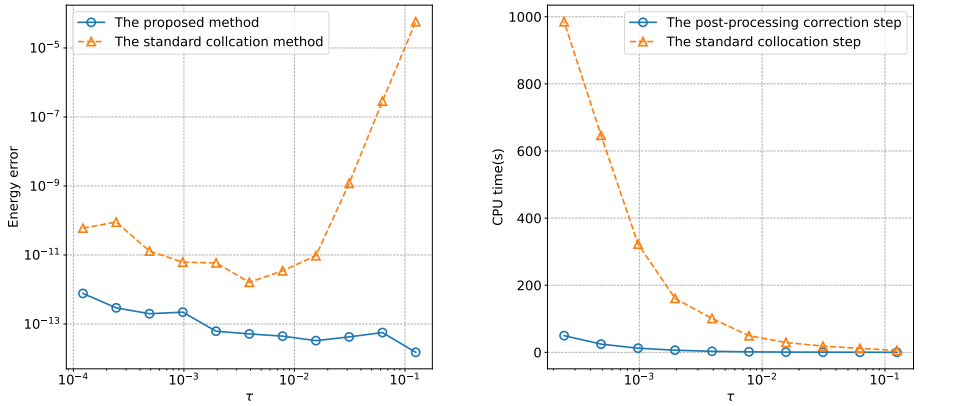
This shows that the number of iterations are not large and acceptable in exchange of the conservation of mass and energy in the numerical solutions. To quantify the computational efficiency of the proposed post-processing correction method for energy conservation, we compare it to the standard Gauss collocation finite element method (FEM) in terms of energy loss and CPU time for various time steps. The results are illustrated in Figure 4.4, which shows that the proposed method significantly decreases the energy loss without essentially increasing the computational cost.



(a) Evolution of the amplitude

(b) Mass and energy conservations, and the number of Newton iterations

Fig. 4.3. *Evolution of the amplitude, conservations, and number of iterations (Example 4.1)*



(a) Energy error for the proposed method, and the standard collocation method

(b) CPU time for the post-processing correction, and the standard collocation steps

Fig. 4.4. *The energy loss and the CPU times for different τ (Example 4.1)*

Example 4.2 (A one-dimensional bi-soliton). We consider the one-dimensional focusing NLS equation on the real line $\mathbb{R}$ with the following solution:

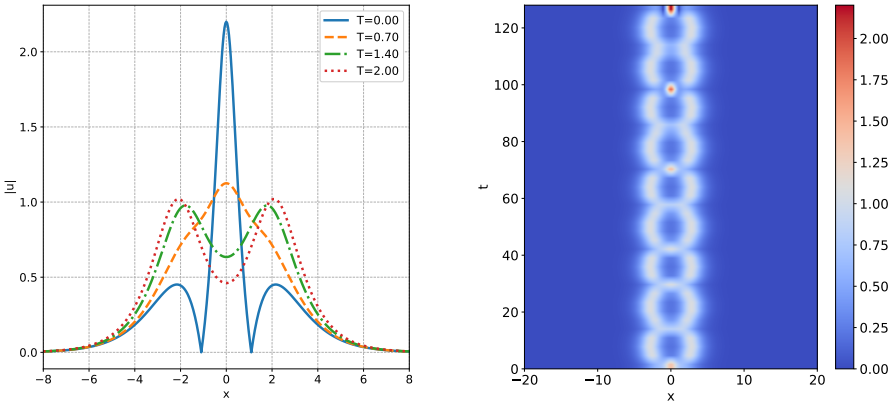
$$u(t,x) = \frac{e^{\mathrm{i}M^2 t} M \operatorname{sech} Mx - e^{\mathrm{i}N^2 t} N \operatorname{sech} Nx}{\cosh J - \sinh J \left(\tanh Mx \tanh Nx + \cos S \operatorname{sech} Mx \operatorname{sech} Nx\right)}, \tag{4.6}$$

with

$$S = \left(M^2 - N^2\right) t, \quad \tanh J = K^2 / \left(1 + L^2\right) = 2MN / \left(M^2 + N^2\right).$$

The solution in (4.6) represents the interaction between two individual solitons; see [22]. In this example, we choose $M = 1.2$, $N = 1$. At $t = 0$, the strong interaction between two solitons results in a striking peak of $|u|$ at the origin, as shown in Figure 4.5 (a). This peak indicates larger L^∞ and H^1 norms of the initial function. The most distinctive characteristic of $|u|$ is its periodicity, with a period of $2\pi/(M^2 - N^2)$. Consequently, the periodic appearance of the initial peak (see Figure 4.5 (b)) presents challenges for numerical methods in terms of energy conservation and accuracy, making this case ideal for examining the long-time performance of the proposed method.

In this example, the parameters for numerical discretization are chosen as $\tau = 2^{-5}$, $h = 2^{-4}$, $k = 2$, $p = 3$, and the end time is set to $T = 128$. Given that (4.6) decays exponentially as $|x| \to \infty$, the chosen parameter settings ensure that the solution before $T = 128$ has negligible amplitude (up to round-off errors) at the boundary of the truncated domain $[-20, 20]$. The performance of the proposed method is compared with the standard Gauss collocation method (using the same parameters) in terms of energy loss and H^1 error. As illustrated in Figure 4.6, the proposed method maintains energy conservation up to round-off errors and significantly outperforms the standard Gauss collocation method. Moreover, the proposed method's advantage in energy conservation also greatly reduces the H^1 error of the numerical solutions.



(a) Initial function and short time evolution

(b) Long time evolution of the numerical solution of the proposed method

Fig. 4.5. *Evolution of the amplitude of the numerical solution (Example 4.2)*

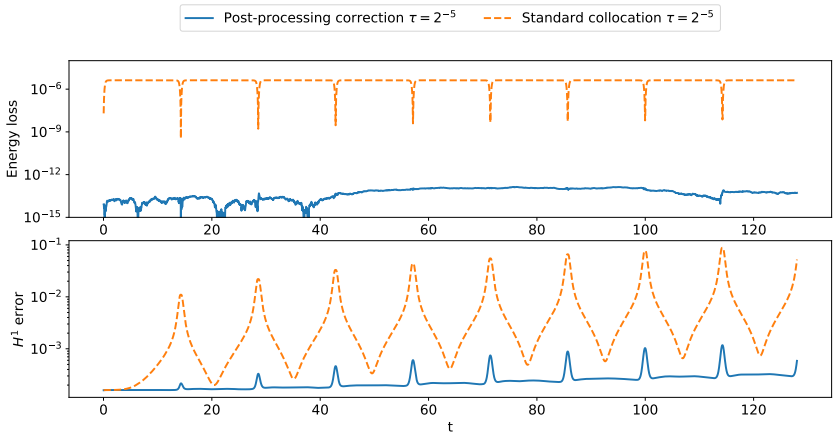


Fig. 4.6. *Energy loss and H^1 error in a long-time evolution, up to $T = 128$ (Example 4.2).*

Example 4.3 (A two-dimensional soliton). We consider the cubic focusing NLS equation in two dimensions with the following exact solution:

$$u(t, x, y) = \sqrt{2}\,\mathrm{sech}(x + y - 2\sqrt{2}t)e^{\mathrm{i}\left(\frac{\sqrt{2}}{2}(x+y)+t\right)}. \tag{4.7}$$

Since the solution u decays exponentially as $|x+y|$ tends to ∞ and is a constant along $x + y = C$, we choose a rectangular domain

$$\Omega = \{(x, y) \in \mathbb{R}^2 : |x| + |y| \le \sqrt{2}L\},$$

with $L = 10$. For $t \in [0, T]$ with $T = 1$, the restriction of the solution to Ω approximately satisfies the periodic boundary condition up to round-off errors. Therefore, we consider the cubic focusing NLS equation in the rectangular domain Ω with the periodic boundary condition, and solve the equation by the proposed method.

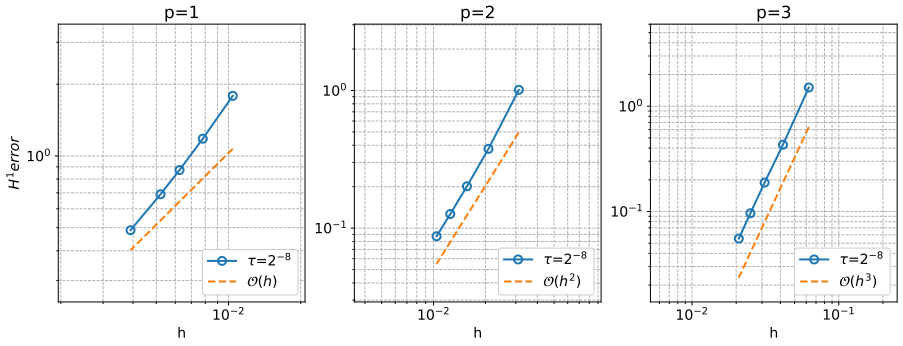


Fig. 4.7. *Space discretization errors in the $L^\infty(0, T, H^1)$ norm (Example 4.3).*

In Figure 4.7 we can see that the errors from the spatial discretizations are $\mathcal{O}(h^p)$, where the time stepsize is chosen to be $\tau = 2^{-8}$ with $k = 3$, which guarantees that the errors from the time discretizations are negligible in observing the spatial convergence

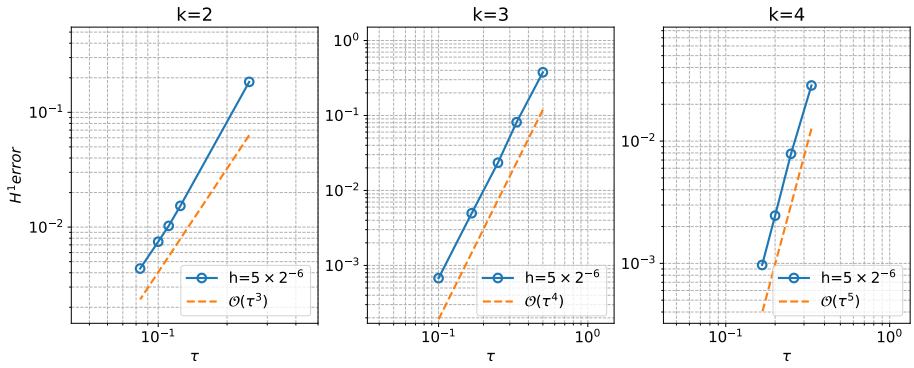
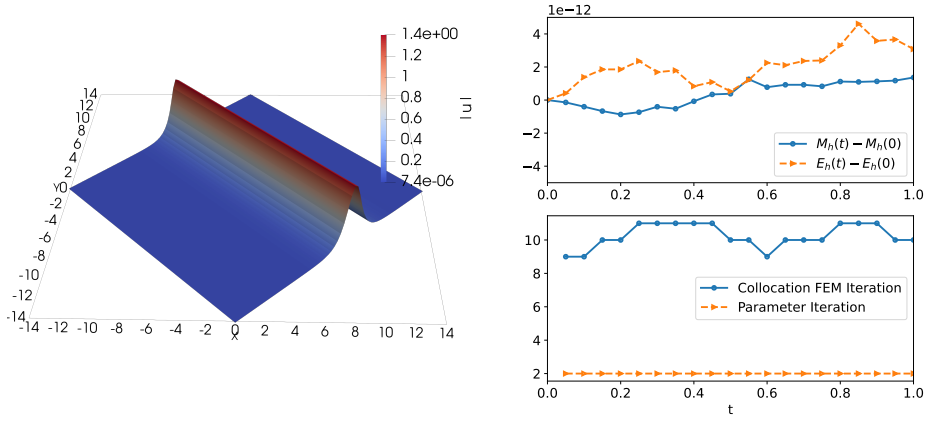


Fig. 4.8. *Time discretization errors in the* $L^\infty(0,T,H^1)$ *norm (Example 4.3).*



(a) Amplitude of numerical solution at $T = 1$ (b) Mass and energy conservations, and the number of Newton iterations

Fig. 4.9. *Amplitude, conservations, and number of iterations (Example 4.3)*

orders. In Figure 4.8 we can see that the errors from the time discretizations are $\mathcal{O}(\tau^{k+1})$, where the spatial mesh size is chosen to be $h = 5 \times 2^{-6}$ with $p = 3$, which guarantees that the errors from the spatial discretizations are negligible in observing the temporal convergence orders. These numerical results are consistent with the theoretical results proved in Theorem 2.1.

The amplitudes of the numerical solution at $T = 1$ and errors in conserving the mass and energy are presented in Figure 4.9 by choosing $L = 10$, $k = 2$, $p = 3$, $T = 1$, $\tau = 0.05$, $h = 5 \times 2^{-4}$. We see that the mass and energy are conserved up to $O(10^{-12})$, which is due to the tolerance errors for the Newton iterations and negligibly small compared with the errors from the time and space discretizations. In the computations, about 10 Newton iterations are needed for the Gauss collocation FEM on every time level, and only 2 iterations are needed for computing the parameters α and β. In particular, the computational cost for the post-processing procedure is negligible compared with that for the Gauss collocation FEM. Therefore, the additional com-

putational cost in the post-processing correction procedure is cheap and acceptable in exchange of the conservation of mass and energy in the numerical solutions.

5. Conclusion. We have proposed a new post-processing correction procedure which, in combination with the Gauss collocation time-stepping method, yields a class of high-order methods for the NLS equation with the desired mass and energy conservation properties for solutions which are not standing waves (i.e., with initial data not being eigenfunctions of the NLS operator). We have shown that the error of the numerical solution is $O(\tau^{k+1} + h^p)$ in the $L^\infty(0, T; H^1)$ norm, where k and p are the degrees of finite elements in time and space, respectively, which can be arbitrarily large. In the numerical examples we have illustrated the performance of the proposed new method in conserving the mass and energy, as well as its high-order convergence in simulating solitons and bi-solitons.

Acknowledgements. The work of the authors was partially supported by the CAS AMSS-PolyU Joint Laboratory of Applied Mathematics and a grant from the Research Grants Council of Hong Kong (GRF Project No. PolyU15301321).

REFERENCES

[1] M. J. Ablowitz, *Nonlinear Dispersive Waves: Asymptotic Analysis and Solitons*, Cambridge University Press, 1 ed., 2011.
[2] G. D. Akrivis and V. A. Dougalis, *On fully discrete Galerkin methods of second-order temporal accuracy for the nonlinear Schrödinger equation.*, Numer. Math., 59 (1991), pp. 31–54.
[3] W. Bao and Y. Cai, *Mathematical theory and numerical methods for Bose-Einstein condensation*, Kinet. Relat. Models, 6 (2012), pp. 1–135.
[4] W. Bao and Y. Cai, *Optimal error estimates of finite difference methods for the Gross-Pitaevskii equation with angular momentum rotation*, Math. Comput., 82 (2012), pp. 99–128.
[5] W. Bao and Q. Du, *Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow*, SIAM J. Sci. Comput., 25 (2003), pp. 1674–1697.
[6] W. Bao, Q. Tang, and Z. Xu, *Numerical methods and comparison for computing dark and bright solitons in the nonlinear Schrödinger equation*, J. Comput. Phys., 235 (2013), pp. 423–445.
[7] W. Bao and W. Tang, *Ground-state solution of Bose–Einstein condensate by directly minimizing the energy functional*, J. Comput. Phys., 187 (2003), pp. 230–254.
[8] C. Besse, *A relaxation scheme for the nonlinear Schrödinger equation*, SIAM J. Numer. Anal., 42 (2004), pp. 934–952.
[9] C. Besse, S. Descombes, G. Dujardin, and I. Lacroix-Violet, *Energy-preserving methods for nonlinear Schrödinger equations*, IMA J. Numer. Anal., 41 (2021), pp. 618–653.
[10] W. Cai, J. Li, and Z. Chen, *Unconditional convergence and optimal error estimates of the Euler semi-implicit scheme for a generalized nonlinear Schrödinger equation*, Adv. Comput. Math., 42 (2016).
[11] C.-L. Chen, *Foundations for Guided-Wave Optics*, Wiley-Interscience, 1st ed., 2006.
[12] L. Erdős, B. Schlein, and H.-T. Yau, *Derivation of the Gross-Pitaevskii equation for the dynamics of Bose-Einstein condensate*, Ann. Math., 172 (2010), pp. 291–370.
[13] X. Feng, B. Li, and S. Ma, *High-order mass- and energy-conserving SAV-Gauss collocation finite element methods for the nonlinear Schrödinger equation*, SIAM J. Numer. Anal., 59 (2021), pp. 1566–1591.
[14] L. Gauckler and C. Lubich, *Splitting integrators for nonlinear Schrödinger equations over long times*, Found. Comput. Math., 10 (2010), pp. 275–302.
[15] P. Henning and D. Peterseim, *Crank–Nicolson Galerkin approximations to nonlinear Schrödinger equations with rough potentials*, Math. Models Methods Appl. Sci., 27 (2017), pp. 2147–2184.
[16] P. Henning and J. Wärnegard, *Superconvergence of time invariants for the Gross–Pitaevskii equation*, Math. Comp., 91 (2022), pp. 509–555.
[17] O. Karakashian and C. Makridakis, *A space-time finite element method for the nonlinear Schrödinger equation: the discontinuous Galerkin method*, Math. Comp., 67 (1998), pp. 479–499.

[18] O. Karakashian and C. Makridakis, *A space-time finite element method for the nonlinear Schrödinger equation: the continuous Galerkin method*, SIAM J. Numer. Anal., 36 (1999), pp. 1779–1807.
[19] E. H. Lieb, R. Seiringer, and J. Yngvason, *A rigorous derivation of the Gross–Pitaevskii energy functional for a two-dimensional bose gas*, Commun. Math. Phys., 224 (2001), pp. 17–31.
[20] H. Liu, Y. Huang, W. Lu, and N. Yi, *On accuracy of the mass-preserving DG method to multi-dimensional Schrödinger equations*, IMA J. Numer. Anal., 39 (2019), pp. 760–791.
[21] C. Lubich, *On splitting methods for Schrödinger-Poisson and cubic nonlinear Schrödinger equations*, Math. Comp., 77 (2008), pp. 2141–2153.
[22] D. H. Peregrine, *Water waves, nonlinear Schrödinger equations and their solutions*, ANZIAM J., 25 (1983), pp. 16–43.
[23] J. M. Sanz-Serna, *Methods for the numerical solution of the nonlinear Schrödinger equation*, Math. Comput., 43 (1984), pp. 21–27.
[24] J. Schöberl, *C++11 implementation of finite elements in NGSolve*, Technical Report ASC Report 30, Institute for Analysis and Scientific Computing, 2014.
[25] J. Shen, T. Tang, and L. Wang, *Spectral Methods: Algorithms, Analysis and Applications*, vol. 41 of Springer Series in Computational Mathematics, Springer, 2011.
[26] J. Shen, J. Xu, and J. Yang, *The scalar auxiliary variable (SAV) approach for gradient flows*, J. Comput. Phys., 353 (2018), pp. 407–416.
[27] J. Shen, J. Xu, and J. Yang, *A new class of efficient and robust energy stable schemes for gradient flows*, SIAM Rev., 61 (2019), pp. 474–506.
[28] C. Sulem and P.-L. Sulem, *The Nonlinear Schrödinger Equation: Self-Focusing and Wave Collapse*, Applied Mathematical Sciences 139, Springer-Verlag New York, 1 ed., 1999.
[29] M. Thalhammer, *Convergence analysis of high-order time-splitting pseudospectral methods for nonlinear Schrödinger equations*, SIAM J. Numer. Anal., 50 (2012), pp. 3231–3258.
[30] Y. Tourigny, *Optimal H1 estimates for two time-discrete Galerkin approximations of a nonlinear Schrödinger equation*, IMA J. Numer. Anal., 11 (1991), pp. 509–523.
[31] J. Wang, *A new error analysis of Crank–Nicolson Galerkin FEMs for a generalized nonlinear Schrödinger equation*, J. Sci. Comput., 60 (2014), pp. 390–407.
[32] H. C. Yuen and B. M. Lake, *Instabilities of waves on deep water*, Annu. Rev. Fluid Mech., 12 (1980), pp. 303–334.
[33] G. Zouraris, *On the convergence of a linear two-step finite element method for the nonlinear Schrödinger equation*, Math. Model. Numer. Anal., 35 (2001), pp. 389–405.