ESAIM: M2AN 57 (2023) 1143–1170
https://doi.org/10.1051/m2an/2023007

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

A VARIABLE TIME-STEP IMEX-BDF2 SAV SCHEME AND ITS SHARP ERROR ESTIMATE FOR THE NAVIER–STOKES EQUATIONS

Yana Di[1,2,4], Yuheng Ma[3,4], Jie Shen[4,5,*] and Jiwei Zhang[4,6]

Abstract. We generalize the implicit-explicit (IMEX) second-order backward difference (BDF2) scalar auxiliary variable (SAV) scheme for Navier–Stokes equation with periodic boundary conditions (Huang and Shen, *SIAM J. Numer. Anal.* **59** (2021) 2926–2954) to a variable time-step IMEX-BDF2 SAV scheme, and carry out a rigorous stability and convergence analysis. The key ingredients of our analysis are a new modified discrete Grönwall inequality, exploration of the discrete orthogonal convolution (DOC) kernels, and the unconditional stability of the proposed scheme. We derive global and local optimal H^1 error estimates in 2D and 3D, respectively. Our analysis provides a theoretical support for solving Navier–Stokes equations using variable time-step IMEX-BDF2 SAV schemes. We also design an adaptive time-stepping strategy, and provide ample numerical examples to confirm the effectiveness and efficiency of our proposed methods.

Mathematics Subject Classification. 65M15, 76D05 and 65M70.

Received June 23, 2022. Accepted January 22, 2023.

1. Introduction

We here consider numerical approximation of the following incompressible Navier–Stokes (N-S) equations

$$\begin{aligned} &\partial_t u - \nu\Delta u + (u\cdot\nabla u) + \nabla p = 0, && \boldsymbol{x}\in\Omega, t\in(0,T],\\ &\nabla\cdot u = 0, && \boldsymbol{x}\in\Omega, t\in(0,T],\\ &u(\boldsymbol{x},0) = u_0(\boldsymbol{x}), && \boldsymbol{x}\in\Omega, \end{aligned} \tag{1.1}$$

with periodic boundary conditions, where $\Omega = (-\pi,\pi)^d\,(d=2,3)$ and $\nu>0$ represents the viscosity.

How to efficiently and accurately solve the N-S equations has been a research focus for many decades, see [4, 7–9, 21] and the references therein. While most of the work are concerned with non periodic boundary

Keywords and phrases. Navier–Stokes, variable time stepping, error analysis.

[1] Research Center for Mathematics, Beijing Normal University, Zhuhai 519087, P.R. China.
[2] Department of Mathematical Sciences, BNU-HKBU United International College, Zhuhai 519087, China.
[3] School of Mathematics and Statistics, Wuhan University, Wuhan 430072, P.R. China.
[4] School of Mathematical Science, Xiamen University, Xiamen 361005, China.
[5] Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA.
[6] School of Mathematics and Statistics, Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan 430072, P.R. China.
*Corresponding author: shen7@purdue.edu

conditions, the N-S equations with periodic boundary conditions, which retain the basic mathematical properties of N-S equations with non-periodic boundary conditions but can be more efficiently solved with a Fourier-spectral method, are also of important theoretical and practical interests, particularly in the study of well posedness [25] and of homogeneous turbulence [19,20,24]. Recently, a high-order IMEX SAV scheme was developed for solving the N-S equations [11], and its numerical solution is shown to be stable without any constraint on time-step size. The results in [11] are established for schemes with constant time-step size. However, in order to efficiently capture the dynamics at different stages of the problem, it is highly beneficial to use variable time-step schemes [6, 10, 12, 13]. In fact, the variable time-step BDF2 scheme plays an important role in constructing efficient and accurate algorithms for solving stiff problems, and has been used frequently for solving the parabolic-type equations [3, 14–16, 26, 30]. Thus, it is natural to ask if the stability and convergence properties proved in [11] still hold with variable time-step scheme.

The main contributions of this paper are the stability and convergence analysis for the variable time-step IMEX-BDF2 SAV scheme for N-S equations, more precisely, its unconditional stability and optimal convergence order in H^1-norm. Our theoretical results are achieved under the following mild condition on the adjacent time-step ratio

A1 : $0 < r_k \le r_{\max} - \delta$ for any small constant $0 < \delta < r_{\max} \approx 4.8645$ and $2 \le k \le N$, where $r_{\max}$ is a root of $x^3 = 1 + 2x$.

We point out that the main difficulties of analyzing the variable time-step IMEX-BDF2 SAV scheme are two-fold. On the one hand, since a first-order scheme is used in the first time step, a direct use of the standard discrete Grönwall inequality for the error estimate in the H^1-norm would lead to order reduction. This order reduction of $\mathcal{O}(\tau^{1.5})$ has been observed for linear parabolic equations [27, 31] and N-S equations [28]. Recently, Ma *et al.* proved unconditional optimal $\mathcal{O}(\tau^2)$ convergence in H^1-norm by constructing a modified discrete Grönwall inequality [18]. Inspired by the idea in [18], which is used to obtain the optimal convergence order in H^1-norm in [18], we further generalize the discrete Grönwall inequality to make it applicable to the theoretical analysis of IMEX-BDF2 SAV scheme for N-S equations. On the other hand, the SAV and IMEX approaches in the variable time-step IMEX-BDF2 SAV scheme makes the analysis more difficult than a typical semi-implicit numerical scheme. The error estimate of SAV requires the use of the stability of numerical solutions in the H^2-norm, which has not been studied in the framework of the discrete orthogonal convolution (DOC) kernels, and the existing theories require that the time-step ratio satisfies a more strict assumption than $r_k \le 4.8645$ [28]. Therefore, in order to establish a rigorous theories under $r_k \le 4.8645$, we need to discover new properties of DOC kernels to circumvent the difficulties arisen from applying DOC kernels to IMEX-BDF2 SAV scheme. We point out that for the Newton linearized variable time-step BDF2 scheme, Zhao *et al.* studied the unconditionally optimal convergence in L^2-norm for general semi-linear equations [31].

The remainder of this paper is organized as follows. In Section 2, we present some preliminary theories that will be used in the paper, including important properties of the N-S equations and of IMEX-BDF2 SAV schemes 2.2. In Section 2.3 we present some important lemmas for the DOC kernels. In Section 3, we present our main results: the global optimal second-order H^1-error estimate of the IMEX-BDF2 SAV scheme in 2D, and a local optimal second-order H^1-error estimate in 3D, and defer their proofs to Section 5. In Section 4, we propose an adaptive time-stepping strategy, and present numerical results to validate our theoretical findings.

2. Preliminaries

2.1. Some basic functional settings and the trilinear form

The N-S equations considered here satisfy a periodic boundary condition and will be solved by Fourier spectral methods. We now introduce some relevant functional spaces related to the periodic boundary conditions. If we assume $\int_\Omega u_0 \,\mathrm{d}x = 0$, then it is easy to see that the solution u of (1.1) also satisfies $\int_\Omega u \,\mathrm{d}x = 0$. Hence, in this paper, we assume $\int_\Omega u_0 \,\mathrm{d}x = 0$ so that $\int_\Omega u \,\mathrm{d}x = 0$ at all times.

We set

$$H_p^k(\Omega)=\left\{u:u=\sum_{j\in Z^d}c_je^{2ij\cdot x},\ \bar{c}_j=c_{-j},\ \sum_{j\in Z^d}|c_j|^2|j|^{2k}<\infty\right\},\tag{2.1}$$

with norm denoted by $\|\cdot\|_k$ (for simplicity, $\|\cdot\|:=\|\cdot\|_0$), and

$$\dot{H}_p^k(\Omega)=\left\{u\in H_p^k(\Omega):\int_\Omega u\,\mathrm{d}x=0\right\}.$$

We define below two spaces which are particularly useful for the N-S equations:

$$H=\big\{v\in L^2(\Omega)|\nabla\cdot v=0\big\},\quad V=\big\{v\in H_p^1(\Omega)|\nabla\cdot v=0\big\}.$$

For periodic problems, the operators $\nabla,\nabla\cdot$ and Δ^{-1} can be defined in Fourier space. For example, when $\Omega=(-\pi,\pi)^2$, for any $u\in L^2(\Omega)$, the spectrum of $\Delta^{-1}u\in L_0^2(\Omega):=\{u\in L^2(\Omega):\int_\Omega u\,\mathrm{d}x=0\}$ is defined by:

$$\widehat{\Delta^{-1}u}(\xi):=-\frac{\widehat{u}(\xi)}{\xi_1^2+\xi_2^2},\quad\forall\xi\in\mathbb{Z}^2-\{0\}\quad\text{and}\quad\widehat{\Delta^{-1}u}(0):=0.$$

In fact, the operators commute with each other in the following way:

$$\nabla\times\nabla\times w=-\Delta w+\nabla(\nabla\cdot w),\qquad\forall w\in H_p^2(\Omega),\tag{2.2}$$

and satisfy:

$$\|\nabla\times\nabla\times w\|^2=\|\Delta w\|^2-\|\nabla(\nabla\cdot w)\|^2,\qquad\forall w\in H_p^2(\Omega).$$

According to (2.2), we can define a linear operator that will be used to simplify (1.1) later:

$$A(v):=\nabla\times\nabla\times\Delta^{-1}v,\qquad\forall v\in L_0^2(\Omega).$$

One can prove that the property $\nabla A(v)=A(\nabla v)$ for $v\in H_p^1(\Omega)$ directly using Fourier transform.

We define the following trilinear form $b(\cdot,\cdot,\cdot)$ and $b_A(\cdot,\cdot,\cdot)$ by

$$b(u,v,w)=\int_\Omega(u\cdot\nabla)v\cdot w\,\mathrm{d}x,\qquad b_A(u,v,w)=\int_\Omega A((u\cdot\nabla)v)\cdot w\,\mathrm{d}x.$$

In particular, we have

$$b(u,v,w)=-b(u,w,v),\qquad\forall u\in H,\ v,w\in H_p^1(\Omega),$$

which implies

$$b(u,v,v)=0,\qquad\forall u\in H,\ v\in H_p^1(\Omega).$$

In the two-dimensional periodic case, $b(\cdot,\cdot,\cdot)$ further satisfies:

$$b(u,u,\Delta u)=0,\quad\forall u\in H_p^2(\Omega).$$

According to [25], the following inequalities hold, which will play an important role in our proof later:

$$\begin{aligned}
&b(u,v,w),\ b_A(u,v,w)\le C\|u\|_1^{1/2}\|u\|^{1/2}\|v\|_2^{1/2}\|v\|_1^{1/2}\|w\|, && d=2;\\
&b(u,v,w),\ b_A(u,v,w)\le C\|u\|_1\|\nabla v\|_{1/2}\|w\|, && d\le3;\\
&b(u,v,w),\ b_A(u,v,w)\le\begin{cases}C\|u\|_1\|v\|_1\|w\|_1,\\ C\|u\|_2\|v\|_0\|w\|_1,\\ C\|u\|_2\|v\|_1\|w\|_0,\\ C\|u\|_1\|v\|_2\|w\|_0,\\ C\|u\|_0\|v\|_2\|w\|_1,\end{cases} && d\le4.
\end{aligned}\tag{2.3}$$

By using these inequalities, it is straightforward to verify

$$\begin{aligned} b_A(u,v,\Delta w) &= -\int_\Omega \nabla A((u\cdot\nabla)v)\cdot\nabla w\,\mathrm{d}x = -\int_\Omega A\Big(\nabla\big((u\cdot\nabla)v\big)\Big)\cdot\nabla w\,\mathrm{d}x \\ &= -\sum_i \big(b_A(u,\partial_i v,\partial_i w) + b_A(\partial_i u, v, \partial_i w)\big), \end{aligned}$$

which implies

$$b_A(u,v,\Delta w) \le c\|u\|_2\|v\|_2\|w\|_1. \tag{2.4}$$

2.2. The SAV scheme

Following [11], the form of N-S equations (1.1) can be simplified to a system that only involves u. To do so, we take the divergence on both sides of (1.1) and derive

$$\Delta p + \nabla\cdot(u\cdot\nabla u) = 0.$$

By applying the equality (2.2), one has

$$\begin{aligned} \nabla p = \nabla\Delta^{-1}\Delta p &= -\nabla\Delta^{-1}\nabla\cdot(u\cdot\nabla u) = -\nabla\nabla\cdot\Delta^{-1}(u\cdot\nabla u) \\ &= -(\Delta + \nabla\times\nabla\times)\Delta^{-1}(u\cdot\nabla u) = -u\cdot\nabla u - A(u\cdot\nabla u). \end{aligned}$$

Thus, we equivalently reformulate the N-S equation into

$$u_t = \nu\Delta u + A(u\cdot\nabla u). \tag{2.5}$$

Applying the properties of trilinear form $b(\cdot,\cdot,\cdot)$, it is easy to check the solution to N-S equations (1.1) satisfies the following energy dissipation law:

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 = -\nu\|\nabla u\|^2, && d = 2,3, \\ &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla u\|^2 = -\nu\|\Delta u\|^2, && d = 2. \end{aligned}$$

Based on the energy law, an IMEX-BDFk SAV scheme is developed in [11] by introducing a SAV $\gamma(t) = \mathcal{E}(t) + 1$, and expand (1.1) as

$$u_t = \nu\Delta u + A(u\cdot\nabla u), \tag{2.6}$$

$$\frac{\mathrm{d}\gamma}{\mathrm{d}t} = \begin{cases} -\nu\dfrac{\gamma(t)}{\mathcal{E}(u(t))+1}\|\Delta u(t)\|^2, & d = 2, \\ -\nu\dfrac{\gamma(t)}{\mathcal{E}(u(t))+1}\|\nabla u(t)\|^2, & d = 3. \end{cases} \tag{2.7}$$

We construct below a variable time step IMEX-BDF2 scheme for the above system.

We partition the time interval $[0,T]$ into a general nonuniform time mesh, *i.e.*, $0 = t^0 < t^1 < \cdots < t^M = T$, with a given integer M. Denote by $\tau_k := t^k - t^{k-1}$ the kth time-step size, by $\tau := \max_{1\le k\le M}\tau_k$ the maximum time-step size, and by $r_k = \tau_k/\tau_{k-1}$ $(2 \le k \le M)$, the adjacent time-step ratio. Denote by $\bar{U}^n$ the approximation of the exact solution $u(t^n)$, and by $\nabla_\tau\bar{U}^n := \bar{U}^n - \bar{U}^{n-1}$ the difference operator. The BDF2 with variable time-step is defined as

$$\mathcal{D}_2\bar{U}^{n+1} := \frac{1+2r_{n+1}}{\tau_{n+1}(1+r_{n+1})}\nabla_\tau\bar{U}^{n+1} - \frac{r_{n+1}^2}{\tau_{n+1}(1+r_{n+1})}\nabla_\tau\bar{U}^n,$$

and two-step Adams–Bashforth extrapolation is defined as

$$B_2(\bar{U}^n) := (1+r_{n+1})\bar{U}^n - r_{n+1}\bar{U}^{n-1}.$$

We point out that the start of BDF2 scheme needs two steps' information. Differing from the initial setting in [11], we here use BDF1 (*i.e.*, set $r_1 = 0$), and one-step Adams–Bashforth extrapolation to compute first step value U^1. By setting $r_1 = 0$, $b_0^{(n)} = (1+2r_n)/(\tau_n(1+r_n))$, $b_1^{(n)} = -r_n^2/(\tau_n(1+r_n))$ and $b_j^{(n)} = 0$ for $2 \le j \le n-1$, the BDF1 and BDF2 can be reformulated into a unified convolution form of

$$\mathcal{D}_2\bar{U}^n = \sum_{k=1}^{n} b_{n-k}^{(n)} \nabla_\tau \bar{U}^k, \qquad n \ge 1. \tag{2.8}$$

The semi-discrete variable time-step IMEX-BDF2 SAV scheme for (2.6) and (2.7) can be written as:

$$\begin{aligned}
&(\mathcal{D}_2\bar{U}^{n+1}, v) + \nu(\nabla\bar{U}^{n+1}, \nabla v) - (A(B_2(U^n)\cdot\nabla B_2(U^n)), v) = 0, \quad \forall v \in L^2(\Omega),\\
&\frac{\gamma^{n+1}-\gamma^n}{\tau_{n+1}} = \begin{cases} -\nu\dfrac{\gamma^{n+1}}{\mathcal{E}(\bar{U}^{n+1})+1}\|\Delta\bar{U}^{n+1}\|^2, & d=2,\\ -\nu\dfrac{\gamma^{n+1}}{\mathcal{E}(\bar{U}^{n+1})+1}\|\nabla\bar{U}^{n+1}\|^2, & d=3, \end{cases}\\
&\xi^{n+1} = \frac{\gamma^{n+1}}{\mathcal{E}(\bar{U}^{n+1})+1},\\
&U^{n+1} = \eta^{n+1}\bar{U}^{n+1},
\end{aligned} \tag{2.9}$$

with $\eta^{n+1} = 1-(1-\xi^{n+1})^2$ and

$$\mathcal{E}(\bar{U}^{n+1}) = \begin{cases} \dfrac{1}{2}\|\nabla\bar{U}^{n+1}\|^2, & d=2,\\ \dfrac{1}{2}\|\bar{U}^{n+1}\|^2, & d=3. \end{cases}$$

Next, we construct a Fourier-spectral method for (2.9). For the sake of brevity, we only consider the three-dimensional case with $\Omega = (-s,s)^3$, the two-dimensional case can be dealt with similarly. We define the Fourier approximation space as

$$S_N = \operatorname{span}\{e^{i\xi_j x}e^{i\eta_k y}e^{i\zeta_l z} : -N_x \le j \le N_x, -N_y \le k \le N_y, -N_z \le l \le N_z\},$$

where $i = \sqrt{-1}$, $\xi_j = \pi j/s$, $\eta_k = \pi k/s$ and $\zeta_l = \pi l/s$. Then, any function $u(x,y,z) \in L^2(\Omega)$ can be approximated by:

$$u(x,y,z) \approx u_N(x,y,z) = \sum_{j=-N_x}^{N_x}\sum_{k=-N_y}^{N_y}\sum_{l=-N_z}^{N_z} \hat{u}_{j,k,l}e^{i\xi_j x}e^{i\eta_k y}e^{i\zeta_l z},$$

with the Fourier coefficients defined as

$$\hat{u}_{j,k,l} = \frac{1}{|\Omega|}\int_\Omega u\cdot e^{-i(\xi_j x+\eta_k y+\zeta_l z)}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z.$$

In remainder of this paper, we fix $N_x = N_y = N_z = N$ for simplicity. It is easy to notice that S_N is a subspace that contains low frequency functions.

Given $\bar{U}_N^n, \bar{U}_N^{n-1}, U_N^n, U_N^{n-1} \in S_N$ and γ^n, we compute $\bar{U}_N^{n+1}, U_N^{n+1} \in S_N$ and γ^{n+1} by

$$\begin{aligned}
&\big(\mathcal{D}_2\bar{U}_N^{n+1}, v\big) + \nu\big(\nabla\bar{U}_N^{n+1}, \nabla v\big) - \big(A(B_2(U_N^n)\cdot\nabla B_2(U_N^n)), v\big) = 0, \qquad \forall v \in S_N \subset L^2(\Omega),\\
&\frac{\gamma^{n+1}-\gamma^n}{\tau_{n+1}} = \begin{cases} -\nu\dfrac{\gamma^{n+1}}{\mathcal{E}\big(\bar{U}_N^{n+1}\big)+1}\big\|\Delta\bar{U}_N^{n+1}\big\|^2, & d=2,\\ -\nu\dfrac{\gamma^{n+1}}{\mathcal{E}\big(\bar{U}_N^{n+1}\big)+1}\big\|\nabla\bar{U}_N^{n+1}\big\|^2, & d=3, \end{cases}\\
&\xi^{n+1} = \frac{\gamma^{n+1}}{\mathcal{E}\big(\bar{U}_N^{n+1}\big)+1}, \qquad U_N^{n+1} = \eta^{n+1}\bar{U}_N^{n+1} \text{ with } \eta^{n+1} = 1-\big(1-\xi^{n+1}\big)^2
\end{aligned} \tag{2.10}$$

where $\mathcal{E}$ is defined as above. To get the initial value in our scheme, we define the L^2-orthogonal projection operator $\Pi_N : L^2(\Omega) \to S_N$ as

$$(\Pi_N u - u, \omega) = 0, \quad \forall\, \omega \in S_N, \quad u \in L^2(\Omega).$$

Setting $\bar{U}_N^0 = U_N^0 := \Pi_N u_0$, $\gamma^0 := E(U_N^0) + 1$, we have the following lemma [2].

Lemma 2.1. *For any $0 \le k \le m$, $\exists\, C$, s.t.*

$$\|\Pi_N u - u\|_k \le C\|u\|_m N^{k-m}, \quad \forall u \in H_p^m(\Omega). \tag{2.11}$$

It is easy to see that the following properties hold:

- Given the initial condition $u_0 \in V$, then the IMEX-BDF2 SAV scheme (2.9) (resp. (2.10)) admits a unique solution satisfies $U^q V$ (resp. $U_N^q \in S_N \cup V$).
- Whenever the pressure is needed in (2.10), it can be computed from

$$p_N^{n+1} = -\Delta^{-1}\Pi_N\nabla\cdot\big(U_N^{n+1}\cdot\nabla U_N^{n+1}\big). \tag{2.12}$$

Since the mathematical properties of semi-discrete scheme (2.9) and fully discrete scheme (2.10) are similar, so in this paper we only focus on the fully discrete scheme (2.10).

The variable time-step IMEX-BDF2 SAV scheme (2.10) satisfies the following stability result:

Theorem 2.2. *Let $\gamma_0 = \mathcal{E}(U_N^0) + 1 \ge 0$, $u_0 \in V \cap H_p^2$ if $d = 2$, and $u_0 \in V$ if $d = 3$. Then, it holds*

$$\gamma^{n+1} - \gamma^n = \begin{cases} -\tau_{n+1}\nu\xi_{n+1}\|\Delta\bar{U}_N^{n+1}\|^2, & d=2,\\ -\tau_{n+1}\nu\xi_{n+1}\|\nabla\bar{U}_N^{n+1}\|^2, & d=3. \end{cases} \tag{2.13}$$

Furthermore, there exists a $F > 0$ such that $\|\nabla U_N^n\| \le F$ for $d = 2$, and $\|U_N^n\| \le F$ for $d = 3$, where F is a constant determined by γ^0 only.

The proof of this theorem is exactly the same to the one in [11], and we omit it here. Note that since $\bar{U}_N^q \in L_0^2(\Omega)$, it also holds $\|U_N^n\|_1 \le F$ for $d = 2$ thanks to the Poincaré inequality.

2.3. Some useful lemmas and their proofs

We now present several Lemmas used in our proof later. The technique of DOC kernels plays a key role in the proof of our main results, and their definition and properties will be presented in this subsection.

Set $\delta_{nk} = 1$ if $n = k$ and $\delta_{nk} = 0$ if $n \ne k$. The DOC kernels $\theta_{n-j}^{(n)}$ are defined by

$$\sum_{j=k}^{n}\theta_{n-j}^{(n)}b_{j-k}^{(j)} = \delta_{nk}, \qquad \forall 1 \le k \le n, \tag{2.14}$$

which produces the property of

$$\sum_{j=1}^{n}\theta_{n-j}^{(n)}\mathcal{D}_2 w^j=\sum_{l=1}^{n}\nabla_\tau w^l\sum_{j=l}^{n}\theta_{n-j}^{(n)}b_{j-l}^{(j)}=w^n-w^{n-1},\qquad 1\le n\le N. \tag{2.15}$$

Lemma 2.3 ([30]). *Assume the adjacent time-step ratio r_k satisfies* **A1**. *Then for any real sequence $\{w_k\}_{k=1}^n$ and any given small constant $0<\delta<r_{\max}\approx 4.8645$ (see* **A1**), *it holds*

$$2w_k\sum_{j=1}^{k}b_{k-j}^{(k)}w_j\ge\frac{r_{k+1}\sqrt{r_{\max}}}{(1+r_{k+1})}\frac{w_k^2}{\tau_k}-\frac{r_k\sqrt{r_{\max}}}{(1+r_k)}\frac{w_{k-1}^2}{\tau_{k-1}}+\frac{\delta\sqrt{r_{\max}}}{(1+r_{\max})^2}\frac{w_k^2}{\tau_k},\qquad k\ge 2, \tag{2.16}$$

$$2\sum_{k=1}^{n}w_k\sum_{j=1}^{k}b_{k-j}^{(k)}w_j\ge\frac{\delta\sqrt{r_{\max}}}{(1+r_{\max})^2}\sum_{k=1}^{n}\frac{w_k^2}{\tau_k}\ge\frac{\delta}{20}\sum_{k=1}^{n}\frac{w_k^2}{\tau_k}\ \ge\ 0,\qquad \text{for}\ \ n\ge 1. \tag{2.17}$$

Corollary 2.4. *If the condition in Lemma 2.3 is satisfied, then it holds*

$$2\sum_{k=1}^{n}w_k\sum_{j=1}^{k}\theta_{k-j}^{(k)}w_j\ge\sum_{k=1}^{n}\frac{\delta}{20}\frac{\left(\sum_{s=k}^{n}\theta_{s-k}^{(s)}w_s\right)^2}{\tau_k}\ \ge\ 0,\quad \text{for}\ \ n\ge 1. \tag{2.18}$$

Lemma 2.5. *If r_k satisfies* **A1**, *$\{w_k\}_{k=1}^n$ is a complex sequence and $\lambda\in\mathbb{C}$, then it holds that*

$$\lambda\sum_{l=1}^{n}\sum_{j=1}^{l}\theta_{l-j}^{(l)}\left(w^j,\overline{w^l}\right)+\overline{\lambda}\sum_{l=1}^{n}\sum_{j=1}^{l}\theta_{l-j}^{(l)}\left(\overline{w^j},w^l\right)\ge\Re\{\lambda\}C_r\sum_{l=1}^{n}\tau_l|w^l|^2,\quad \text{for}\ \ n\ge 1, \tag{2.19}$$

where $C_r:=\delta/280$, and $(u,v):=u\cdot v$.

Proof. By defining matrix $\boldsymbol{B}$ and $\boldsymbol{\Theta}$ in the same way in [30] as

$$\boldsymbol{B}:=\begin{bmatrix} b_0^{(n)} & 0 & 0 & \cdots & 0 & 0 & 0\\ b_1^{(n)} & b_0^{(n-1)} & 0 & \cdots & 0 & 0 & 0\\ 0 & b_1^{(n-1)} & b_0^{(n-2)} & \cdots & 0 & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & b_1^{(3)} & b_0^{(2)} & 0\\ 0 & 0 & 0 & \cdots & 0 & b_1^{(2)} & b_0^{(1)} \end{bmatrix},$$

$$\boldsymbol{\Theta}:=\begin{bmatrix} \theta_0^{(n)} & 0 & 0 & \cdots & 0 & 0 & 0\\ \theta_1^{(n)} & \theta_0^{(n-1)} & 0 & \cdots & 0 & 0 & 0\\ \theta_2^{(n)} & \theta_1^{(n-1)} & \theta_0^{(n-2)} & \cdots & 0 & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\\ \theta_{n-2}^{(n)} & \theta_{n-3}^{(n-1)} & \theta_{n-4}^{(n-2)} & \cdots & \theta_1^{(3)} & \theta_0^{(2)} & 0\\ \theta_{n-1}^{(n)} & \theta_{n-2}^{(n-1)} & \theta_{n-3}^{(n-2)} & \cdots & \theta_2^{(3)} & \theta_1^{(2)} & \theta_0^{(1)} \end{bmatrix},$$

we have

$$\sum_{l=1}^{n} w^l \sum_{j=1}^{l} \theta_{l-j}^{(l)} \overline{w^j} = \boldsymbol{W}^H \boldsymbol{\Theta} \boldsymbol{W},$$

where $\boldsymbol{W} := [w^n, w^{n-1}, \cdots, w^1]^T$.

Let us first consider the following equation

$$\boldsymbol{W}^H \Big(\lambda \boldsymbol{\Theta} + \overline{\lambda} \boldsymbol{\Theta}^T\Big) \boldsymbol{W} = \boldsymbol{W}^H \boldsymbol{\Theta}^T \big(\lambda \boldsymbol{B}^T + \overline{\lambda} \boldsymbol{B}\big) \boldsymbol{\Theta} \boldsymbol{W} = (\boldsymbol{\Theta} \boldsymbol{W})^H \big(\lambda \boldsymbol{B}^T + \overline{\lambda} \boldsymbol{B}\big) (\boldsymbol{\Theta} \boldsymbol{W}). \tag{2.20}$$

We claim that $\lambda \boldsymbol{B}^T + \overline{\lambda} \boldsymbol{B}$ is a Hermitian matrix. In fact, according to Lemma 2.3, we see $\boldsymbol{B}$ is a real matrix which has the following estimate:

$$\boldsymbol{W}^T \big(\boldsymbol{B} + \boldsymbol{B}^T\big) \boldsymbol{W} \geq \frac{\delta}{20} \|\boldsymbol{T} \boldsymbol{W}\|^2, \quad \boldsymbol{T} := \operatorname{diag}\Big((\sqrt{\tau_n})^{-1}, \big(\sqrt{\tau_{n-1}}\big)^{-1}, \cdots, (\sqrt{\tau_1})^{-1} \Big), \qquad \forall \boldsymbol{W} \in \mathbb{R}^n.$$

So, by Hermitian matrix's properties, it is easy to check the following estimate holds in complex space:

$$\boldsymbol{W}^H \big(\overline{\lambda} \boldsymbol{B} + \lambda \boldsymbol{B}^T\big) \boldsymbol{W} \geq \frac{\Re\{\lambda\} \delta}{20} \|\boldsymbol{T} \boldsymbol{W}\|^2, \qquad \forall \boldsymbol{W} \in \mathbb{C}^n.$$

Applying the above inequality to (2.20), we have

$$\begin{aligned} \boldsymbol{W}^H \Big(\lambda \boldsymbol{\Theta} + \overline{\lambda} \boldsymbol{\Theta}^T\Big) \boldsymbol{W} &\geq \frac{\Re\{\lambda\} \delta}{20} \|\boldsymbol{T} \boldsymbol{\Theta} \boldsymbol{W}\|^2 \\ &= \frac{\Re\{\lambda\} \delta}{20} (\boldsymbol{W})^H \Big(\boldsymbol{\Theta}^T \boldsymbol{T}^2 \boldsymbol{\Theta}\Big) (\boldsymbol{W}) \\ &= \frac{\Re\{\lambda\} \delta}{20} \big(\boldsymbol{T}^{-1} \boldsymbol{W}\big)^H \Big(\boldsymbol{T} \boldsymbol{\Theta}^T \boldsymbol{T}^2 \boldsymbol{\Theta} \boldsymbol{T}\Big) \big(\boldsymbol{T}^{-1} \boldsymbol{W}\big). \end{aligned} \tag{2.21}$$

Here $\boldsymbol{T} \boldsymbol{\Theta}^T \boldsymbol{T}^2 \boldsymbol{\Theta} \boldsymbol{T}$ is a positive definite real matrix, and its inverse matrix is $\boldsymbol{T}^{-1} \boldsymbol{B} \boldsymbol{T}^{-2} \boldsymbol{B}^T \boldsymbol{T}^{-1}$, which is a positive definite matrix that could be written explicitly as:

$$\widetilde{\boldsymbol{B}} := \boldsymbol{T}^{-1} \boldsymbol{B} \boldsymbol{T}^{-2} \boldsymbol{B}^T \boldsymbol{T}^{-1} = \begin{bmatrix} 2D_0^{(n)} & d_1^{(n)} & & & \\ d_1^{(n)} & 2D_0^{(n-1)} & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 2D_0^{(2)} & d_1^{(2)} \\ & & & d_1^{(2)} & 2D_0^{(1)} \end{bmatrix}, \tag{2.22}$$

where (for simplicity, we set $r_{n+1} = 0$ here)

$$2D_0^{(i)} := \frac{(1+2r_i)^2}{(1+r_i)^2} + \frac{r_{i+1}^3}{(1+r_{i+1})^2}, \quad \text{and} \quad d_1^{(i)} := -\frac{r_i \sqrt{r_i}}{(1+r_i)} \left(\frac{1+2r_i}{(1+r_i)} + \frac{1+2r_{i-1}}{(1+r_i)} \right).$$

This lemma is naturally proved if we can find an uniform up bound for the largest eigenvalue of $\widetilde{\boldsymbol{B}}$, *i.e.*,

$$2 \sum_{l=1}^{n} \left(\sum_{j=1}^{l} d_{l-j}^{(l)} w^j, w^l \right) = \boldsymbol{W}^T \widetilde{\boldsymbol{B}} \boldsymbol{W} \leq \Lambda \|\boldsymbol{W}\|^2, \quad \forall \boldsymbol{W} \in \mathbb{R}^n. \tag{2.23}$$

In fact, it is easy to find a proper Λ which is only decided by $r_{\max}$:

$$\begin{aligned}\sum_{l=1}^{n}\left(\sum_{j=1}^{l} d_{l-j}^{(l)} w^j, w^l\right) &\le \sum_{l=1}^{n-1}\Big(d_0^{(l)}|w^l|^2 + d_1^{(l+1)}\big(w^l, w^{l+1}\big)\Big) + d_0^{(n)}|w^n|^2\\ &\le \sum_{l=1}^{n-1}\left(d_0^{(l)} + \frac{1}{2}\Big|d_1^{(l+1)}\Big| + \frac{1}{2}\Big|d_1^{(1)}\Big|\right)\big|w^l\big|^2 + \left(d_0^{(n)} + \frac{1}{2}\Big|d_1^{(n)}\Big|\right)|w^n|^2\\ &\le \Lambda\sum_{l=1}^{n}|w^l|^2,\end{aligned}$$

where $\Lambda := (1+2r_{\max}+\sqrt{r_{\max}^3})^2/(1+r_{\max})^2$ (*i.e.*, $\Lambda \approx 13.39$). By the property of positive definite matrix, it is obvious that $\boldsymbol{T\Theta}^T\boldsymbol{T}^2\boldsymbol{\Theta T}$ has a smallest eigenvalue, and it is bigger than Λ^{-1}, which together with (2.21) derive that

$$\boldsymbol{W}^H\Big(\lambda\boldsymbol{\Theta} + \overline{\lambda}\boldsymbol{\Theta}^T\Big)\boldsymbol{W} \ge \frac{\Re\{\lambda\}\delta(1+r_{\max})^2}{20\Big(1+2r_{\max}+\sqrt{r_{\max}^3}\Big)^2}\big\|\boldsymbol{T^{-1}W}\big\|^2 \ge \frac{\Re\{\lambda\}\delta}{280}\big\|\boldsymbol{T^{-1}W}\big\|^2.$$

Taking the definition of $\boldsymbol{\Theta}, \boldsymbol{T}$ into the above inequality, the proof is completed. □

Lemma 2.6 (Discrete Grönwall inequality). *Assume $\lambda > 0$ and the sequences $\{v_j\}_{j=1}^N$ and $\{\eta_j\}_{j=0}^N$ are nonnegative. If*

$$v_n \le \lambda\sum_{j=1}^{n-1}\tau_j v_j + \sum_{j=0}^{n}\eta_j, \quad \text{for} \quad 1 \le n \le N,$$

then it holds

$$v_n \le \exp\big(\lambda t^{n-1}\big)\sum_{j=0}^{n}\eta^j, \quad \text{for} \quad 1 \le n \le N.$$

Lemma 2.6 can be proved by the standard induction hypothesis and is omitted here. To prove the optimal H^1-error estimate, we further introduce a modified Grönwall's inequality, see also [18].

Lemma 2.7 (A modified Grönwall's inequality). *Assume a constant C and sequences $\{x_j\}_{j=1}^N, \{y_j\}_{j=1}^N, \{a_j\}_{j=1}^N$ and $\{b_j\}_{j=1}^N$ are nonnegative. If*

$$x_n^2 + y_n \le C + \sum_{j=1}^{n-1}\big(a_j x_j^2 + 2b_j x_j\big) + 2b_n x_n, \quad \text{for} \quad 1 \le n \le N,$$

then it holds

$$\begin{aligned} x_n &\le \exp\left(\frac{1}{2}\sum_{j=1}^{n-1} a_j\right)\left(\sqrt{C} + \sum_{j=1}^{n} b_j\right), \qquad \text{for} \quad 1 \le n \le N,\\ y_n &\le \exp\left(\sum_{j=1}^{n-1} a_j\right)\left(\sqrt{C} + \sum_{j=1}^{n} b_j\right)^2, \qquad \text{for} \quad 1 \le n \le N.\end{aligned} \tag{2.24}$$

Proof. Denote by $u_n := C + \sum_{j=1}^{n}(a_j x_j^2 + 2b_j x_j) + 2b_{n+1}x_{n+1}$ for all $1 \le n \le N-1$, and $u_0 := C + 2b_1 x_1$. Clearly we have $x_n \le \sqrt{u_{n-1}}$. Using this property, we can get the estimate of u_0 immediately:

$$u_0 \le C + 2b_1\sqrt{u_0} \qquad \Rightarrow \qquad u_0 \le \sqrt{C} + 2b_1.$$

For the rest of the u_n, it holds

$$u_n - u_{n-1} = a_n x_n^2 + 2b_{n+1}x_{n+1} \le a_n u_{n-1} + 2b_{n+1}\sqrt{u_n}, \qquad \text{for} \quad 1 \le n \le N-1, \tag{2.25}$$

which implies

$$\sqrt{u_n} \le 2b_{n+1} + \sqrt{1+a_n}\sqrt{u_{n-1}}, \qquad \text{for} \quad 1 \le n \le N-1. \tag{2.26}$$

Noting $a_n \ge 0$, it holds $1 \le \sqrt{1+a_n} \le \exp(a_n/2)$, which together with the inequality above, has

$$\begin{aligned}
\sqrt{u_n} &\le \exp(a_n/2)\big(2b_{n+1} + \sqrt{u_{n-1}}\big) \\
&\le \exp((a_n + a_{n-1})/2)\big(2b_{n+1} + 2b_n + \sqrt{u_{n-2}}\big) \\
&\le \exp\left(\frac{1}{2}\sum_{j=1}^{n} a_j\right)\left(2\sum_{j=2}^{n+1} b_j + \sqrt{u_0}\right) \\
&= \exp\left(\frac{1}{2}\sum_{j=1}^{n} a_j\right)\left(\sqrt{C} + 2\sum_{j=1}^{n+1} b_j\right), \qquad \text{for} \quad 1 \le n \le N-1.
\end{aligned} \tag{2.27}$$

This lemma is directly proved by noticing $x_n^2 + y_n \le u_{n-1}$. □

Lemma 2.8 ([30]). *The truncation error $R_2^{j-1} = \mathcal{D}_2 u(t^j) - \partial_t u(t^j)$ $(1 \le j \le N)$ can be expressed by*

$$R_2^{j-1} = \sum_{l=1}^{j} b_{j-l}^{(j)} G^l + P^j, \quad 1 \le j \le n, \tag{2.28}$$

where

$$\begin{aligned}
&G^l = -\frac{1}{2}\int_{t^{l-1}}^{t^l} \big(t - t^{l-1}\big)^2 \partial_{ttt} u \,\mathrm{d}t, \quad 1 \le l \le N, \quad P^1 = \frac{1}{2\tau_1}\int_0^{t^1} t^2 \partial_{ttt} u \,\mathrm{d}t - \frac{1}{\tau_1}\int_0^{t^1} t u_{tt}(t)\,\mathrm{d}t, \\
&P^j = -\frac{1}{2} b_1^{(j)} \tau_{j-1} \int_{t^{j-1}}^{t^j} \big(2\big(t - t^{j-1}\big) + \tau_{j-1}\big)\partial_{ttt} u \,\mathrm{d}t, \quad 2 \le j \le N.
\end{aligned} \tag{2.29}$$

One has the following estimate

$$\sum_{l=1}^{n+1}\sum_{j=1}^{l} \theta_{l-j}^{(l)} \|R_2^{j-1}\| \le \frac{\tau^2}{2}\int_0^T \|\partial_{ttt} u\|\,\mathrm{d}t + 3\tau^2 \int_0^T \|\partial_{ttt} u\|\,\mathrm{d}t + (\tau_1 + \tau)\int_0^{t^1} \|\partial_{tt} u\|\,\mathrm{d}t. \tag{2.30}$$

One can refer to Lemma 2.1 and Theorem 3.4 of [30] for the proof of Lemma 2.8 in details. We now deliver a lemma that will only be used in the analysis of local error estimates for 3D case. It is a variant of the lemma in [17], so we omit its proof here.

Lemma 2.9. *Let $\phi : (0,\infty) \to (0,\infty)$ be continuous and increasing, and let $\mathcal{F} > 0$. Given T_* such that $0 < T_* < \int_{\mathcal{F}}^{\infty} \mathrm{d}z/\phi(z)$. Suppose that sequences $\{z_j\}_{j=1}^N$, $\{w_j\}_{j=1}^N \ge 0$ satisfy*

$$z_{n+1} + \sum_{k=1}^{n} \tau_k w_k \le \mathcal{F} + \sum_{k=1}^{n} \tau_k \phi(z_k), \qquad \forall n \le n_*,$$

with $\sum_{k=1}^{n_*} \tau_k \le T_*$. Then there exists a constant $C_* > 0$, which is independent of $\tau_j > 0$ but dependent of T_*, satisfies

$$z_{n+1} + \sum_{k=1}^{n} \tau_k w_k \le C_*, \qquad 1 \le n \le n_*.$$

3. Main results of variable time-step IMEX-BDF2 SAV scheme

In this section, we state the main results on the global optimal error estimate for the 2D variable time-step IMEX-BDF2 SAV scheme in Theorem 3.1, and the local optimal error estimate for the 3D case in Theorem 3.2. The proofs of these results are based on the stability properties given in Theorem 2.2 and will be deferred to Section 5.

Theorem 3.1. *Let $d = 2$, $T > 0$, $u_0 \in V \cap H_p^m$ with $m \geq 3$, and u be the solution of (1.1). Assume $u \in C(0,T;H_p^m)$ with $m \geq 3$, $\frac{\partial^j u}{\partial t^j} \in L^2(0,T;H_p^2)$ with $1 \leq j \leq 2$, $\frac{\partial^3 u}{\partial t^3} \in L^1(0,T;H_p^1)$. Denote by $\tau_k := t^k - t^{k-1}$ the kth time-step size, by $\tau := \max_{1\leq k\leq M} \tau_k$ the maximum time-step size. Then for $\tau \leq \frac{1}{2+4C_0^2}$ and $N \geq 4C_\Pi^2 + 2$, we have*

$$\|\bar{U}_N^n - u(t^n)\|_1^2, \quad \|U_N^n - u(t^n)\|_1^2 \ \leq \ C\tau^4 + CN^{2(1-m)},$$

and

$$\sum_{q=1}^{n} \tau_q \|\bar{U}_N^q - u(t^q)\|_2^2, \quad \sum_{q=1}^{n} \tau_q \|U_N^q - u(t^q)\|_2^2 \ \leq \ C\tau^4 + CN^{2(2-m)},$$

where the constants C_0, C_Π and C are independent of τ, N.

As pointed out in [11], it is no longer possible to obtain a global error estimate in the 3D case. But, the related local error estimate of 3D case in [11] can still be established for variable time-step scheme. Here we again use the DOC kernels to overcome the analysis difficulties raised by variable time-step.

Theorem 3.2. *Let $d = 3$, $T > 0$, $u_0 \in V \cap H_p^m$ with $m \geq 3$, u be the solution of (1.1). Assume $u \in C(0,T;H_p^m)$ with $m \geq 3$, $\frac{\partial^j u}{\partial t^j} \in L^2(0,T;H_p^2)$ with $1 \leq j \leq 2$, and $\frac{\partial^3 u}{\partial t^3} \in L^1(0,T;H_p^1)$. Given T_* such that $0 < T_* < \int_{\mathcal{F}}^{\infty} \mathrm{d}z/\phi(z)$ (ϕ will be given later in the proof), denote by $\tau_k := t^k - t^{k-1}$ the kth time-step size, by $\tau := \max_{1\leq k\leq M} \tau_k$ the maximum time-step size, and $\mathcal{F}$ is a positive constant that only depends on real solution u. If $\sum_{k=1}^{n_*} \tau_k \leq T_*$, $\tau \leq \frac{1}{2+4C_0^2}$ and $N \geq 4C_\Pi^2 + 2$, it holds that*

$$\begin{aligned} \left\|\bar{U}_N^n - u(t^n)\right\|_1^2, \ \|U_N^n - u(t^n)\|_1^2 &\leq C\tau^4 + CN^{2(1-m)}, \qquad \forall n \leq n_*, \\ \sum_{q=1}^{n} \tau_q \|\bar{U}_N^q - u(t^q)\|_2^2, \quad \sum_{q=1}^{n} \tau_q \|U_N^q - u(t^q)\|_2^2 &\leq C\tau^4 + CN^{2(2-m)}, \qquad \forall n \leq n_*, \end{aligned}$$

where the constants C_0, C_Π and C are independent of τ, N.

The error analysis for the pressure p is exactly the same as the one in [11]. We present the theorem and omit its proof here.

Theorem 3.3. *Under the same assumptions of Theorem 3.1 in 2D and Theorem 3.2 in 3D, there hold*

$$\|p_N^n - p(\cdot,t^n)\|^2 \leq \begin{cases} C\tau^{2k} + CN^{2(1-m)}, & \forall n \geq 0, \quad d = 2, \\ C\tau^{2k} + CN^{2(1-m)}, & \forall n \leq n_*, \quad d = 3, \end{cases} \tag{3.1}$$

and

$$\sum_{q=1}^{n} \tau_q \|\nabla(p_N^q - p(\cdot,t^q))\|^2 \leq \begin{cases} C\tau^{2k} + CN^{2(2-m)}, & \forall n \geq 0, \quad d = 2, \\ C\tau^{2k} + CN^{2(2-m)}, & \forall n \leq n_*, \quad d = 3, \end{cases} \tag{3.2}$$

where p_N^{n+1} is computed from (2.12), C is a constant independent of τ and N, n_* is the biggest integer satisfies $\sum_{k=1}^{n_*} \tau_k \leq T_*$, and T_* is defined in Theorem 3.2.

TABLE 1. Errors and temporal convergence orders with $M = 2^{(4:8)}$.

M	H^1-error	Order
16	1.6129e−04	–
32	4.0055e−05	2.01
64	9.9799e−06	2.00
128	2.4907e−06	2.00
256	6.2222e−07	2.00

4. Numerical results

We now present three examples to verify the effectiveness and convergence of variable time-step IMEX-BDF2 SAV scheme (2.10). In Example 1, we use the benchmark problems given in [5,11] to investigate the convergence order of our proposed scheme. In Example 2, we design an adaptive time-stepping strategy, and then construct a benchmark problem with different time scales to investigate the effectiveness of our variable time-step scheme, and also present several simulations to investigate the influence of the parameters in the adaptive time-stepping strategy on the CPU-time, H^1-error and the minimum time step-size. In Example 3, we use our adaptive time-stepping strategy to deal with the double shear layer problem in [5,11]. The experimental results show that, in the same total CPU time, we can successfully solve some problems by our variable time-step IMEX-BDF2 SAV scheme, but, which *fails* to be solved by the constant time-step IMEX-BDF2 SAV scheme.

Example 1. We first consider the computation of the N-S equation (1.1) in $\Omega = (-\pi, \pi)^2$ with periodic boundary condition, a benchmark problem in [5], by constructing an exact solution satisfying

$$\begin{aligned} u_1(x, y) &= -\cos(x)\sin(y)\exp(-2\nu t);\\ u_2(x, y) &= \sin(x)\cos(y)\exp(-2\nu t);\\ p(x, y) &= 0. \end{aligned}$$

In the simulations, we take $\nu = 1$, $N = 16$, and divide the time interval $[0, 1]$ by a variable time mesh with $\tau = \mathcal{O}(1/M)$. The maximum H^1 errors of u_h^n are listed in Table 1, which shows the second-order convergence of the numerical simulations. According to Theorem 3.1, the optimal error estimate in H^1-norm is at least $\|u_h^n - u(t^n)\|_1 = \mathcal{O}(1/N^2 + \tau^2)$, which is consistent with the results of our experiments.

To make our results more convincing, we also consider the convergence order of a benchmark problem with external forces provided in [11]. Consider (1.1) in $\Omega = (-1, 1)^2$ such that the exact solution satisfies

$$\begin{aligned} u_1(x, y) &= \pi\exp(\sin(\pi(x+1)))\exp(\sin(\pi(y+1)))\cos(\pi(y+1))\sin^2(t);\\ u_2(x, y) &= -\pi\exp(\sin(\pi(x+1)))\exp(\sin(\pi(y+1)))\cos(\pi(x+1))\sin^2(t);\\ p(x, y) &= \exp(\cos(\pi(x+1))\sin(\pi(y+1)))\sin^2(t). \end{aligned}$$

In the simulations, we take $\nu = 1$, $N = 40$, and partition the time interval $[0, 1]$ by a variable-time-mesh with $\tau = \mathcal{O}(1/M)$. The maximum H^1 errors of u_h^n are listed in Table 2, which again shows the second-order convergence in time.

Overall, both the results in Tables 1 and 2 confirm the analysis results in Theorem 3.1.

Example 2. The main advantage of the variable time-step scheme is its ability to efficiently capture the dynamics of numerical solutions in different time scales. This enables us to obtain smaller errors using less CPU time than the constant time-step scheme. Therefore, how to design an appropriate adaptive time-stepping strategy plays an important role in for the variable time-step scheme.

TABLE 2. Errors and temporal convergence orders with $M = 10 \cdot 2^{(4:8)}$.

M	H^1-error	Order
160	1.7524e−00	–
320	3.7830e−01	2.15
640	8.8126e−02	2.07
1280	2.1281e−02	2.03
2560	5.2295e−03	2.02

Several common strategies have been used to construct adaptive time stepping schemes. For example, adaptive time-stepping strategies using energy have been well studied in literatures [22,23]. However some mathematical indicators are not suitable for designing adaptive time-stepping strategies for N-S equations. In fact, when spatial scalar is fixed and $\nu \to 0$, it will reduces to energy conservation for inviscid flows, and the energy dissipation is almost negligible. For the 2D fluid problem we consider here (*i.e.*, $d = 2$ and $\nu = 0$), it will satisfy the Euler equation, the vorticity of the fluid system is also conserved. Therefore, it is not ideal to apply the changes of energy and vorticity directly to the adaptive time-stepping strategy for N-S equations. The adaptive time-stepping strategy in [1] uses a posteriori time error indicator that related to the changes in the adjacent time-step H^1-norm of velocity, and has successfully simulated the unsteady N-S problems.

In this paper, we will design an adaptive time-stepping strategy based on the changes of velocity. A similar strategy is discussed in [1]. Our strategy is motivated by an idea that the errors at two adjacent steps should be roughly equal to each other, and an observation that the changes of velocity can effectively reflect the choice of the adaptive time steps. Specifically, our adaptive time strategy is given by

$$\tau_n = \begin{cases} (1-\alpha)\tau_{n-1}, & \dfrac{\mathcal{E}\big(U_N^{n-1} - U_N^{n-2}\big)}{\mathcal{E}\big(U_N^{n-1}\big)} > \epsilon, \\ \min\{(1+\alpha)\tau_{n-1}, \tau_{\max}\}, & \dfrac{\mathcal{E}\big(U_N^{n-1} - U_N^{n-2}\big)}{\mathcal{E}\big(U_N^{n-1}\big)} \le \epsilon, \end{cases} \tag{4.1}$$

where the parameters ϵ and α are the given positive constant to adjust the change of time steps.

Noting $\frac{\mathcal{E}\big(U_N^{n-1} - U_N^{n-2}\big)}{\tau_{n-1}^2 \mathcal{E}\big(U_N^{n-1}\big)} \approx \frac{\mathcal{E}\big(u_t^{n-1}\big)}{\mathcal{E}(u^{n-1})}$, the above strategy insures $\tau_{n-1} \approx \sqrt{\epsilon \frac{\mathcal{E}(u^{n-1})}{\mathcal{E}(u_t^{n-1})}}$. For the practical simulations, the first step-size τ_1 is set to be small, but our strategy can guarantee that τ_n will evolve to a proper size in a few steps based on the dynamics of the solution itself.

As a benchmark example, we construct an exact solution having different time scales to verify the superiority of the variable time-step IMEX-BDF2 SAV scheme. In fact, by considering the relative speed of change $\|u_t\|/\|u\| = f(t)$, we can choose $f(t)$ to be a function that is extremely large at some time, such as $\frac{G}{1+G^2(t-t^0)^2}$, where G is a large number. We now design the benchmark problem over $\Omega = (-1,1)^2$ with the following exact solution

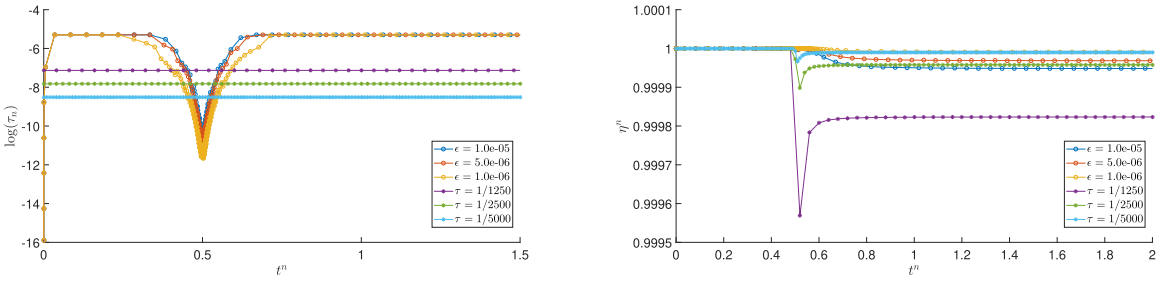
$$\begin{aligned} u_1(x,y) &= \frac{\pi}{100} \exp(\sin(\pi x)) \exp(\sin(\pi y)) \cos(\pi y) \cdot \exp(\arctan(100(t-0.5))); \\ u_2(x,y) &= -\frac{\pi}{100} \exp(\sin(\pi x)) \exp(\sin(\pi y)) \cos(\pi x) \cdot \exp(\arctan(100(t-0.5))); \\ p(x,y) &= 0. \end{aligned} \tag{4.2}$$

In the simulations, we take $\nu = 1$ and $N = 16$. The problem will be solved by using the variable and constant time-step strategies, respectively. In the adaptive time-stepping strategy (4.1), we choose $\tau_1 = 1.25\text{e}{-7}$, $\alpha = 0.2$ and $\tau_{\max} = 5.0\text{e}{-03}$. For variable time-step scheme, we choose $\epsilon = 1.0\text{e}{-05}$, $5.0\text{e}{-06}$ and $1.0\text{e}{-06}$ respectively to solve the problem (4.2) until $t = 4$. The step sizes at different time levels and the evolution of η^n in the experiments above are plotted in Figures 1a and 1b respectively.

TABLE 3. Errors and parameters of experiments for solving (4.2).

ϵ	M	τ	H^1-error	CPU-time (s)
1.0e−05	1758	5.0000e−03	1.3250e−04	10.01
5.0e−06	2153	5.0000e−03	8.0934e−05	11.96
1.0e−06	3859	5.0000e−03	2.4751e−05	22.95
–	5000	8.0000e−04	4.5017e−04	27.37
–	10 000	4.0000e−04	1.0693e−04	51.07
–	20 000	2.0000e−04	2.6069e−05	104.13

Notes. In Table 3, the first 3 lines are experiment results using variable time-step and the last 3 lines are results using constant time-step. We re-emphasize that $\tau := \max_{1\le k\le M} \tau_k$ here.



(a) (b)

FIGURE 1. (Example 2) Experimental results by variable time-step IMEX-BDF2 SAV scheme (ϵ = 1.0e−05, 5.0e−06, 1.0e−06) and constant time-step IMEX-BDF2 SAV scheme ($\tau = 1/1250, 1/2500, 1/5000$). (a) Time step-size at different time level. (b) Evolution of η^n.

As a contrast, we solve the same problem using constant-step sizes $\tau = 1/1250$, $1/2500$ and $1/5000$, respectively. The numerical results are listed in Table 3. Table 3 shows that the variable time-step scheme gives a lower H^1-error with lower computational costs, while the constant time-step scheme gives a large H^1-error at the almost same computational cost. For example, the H^1-error of variable time-step scheme with $\epsilon = 1.0\text{e}{-06}$ and the constant time-step scheme with $\tau = 2.0\text{e}{-04}$ is similar, but the CPU-time of constant time-step scheme is almost five times than the variable time-step one. This can also be noticed by observing the evolution of η^n over time in Figure 1b, and it is noticed that η^n is much closer to 1 for variable time-step scheme, while the η^n is much farther than 1 for constant time-step scheme. Figure 1a shows that the variable time-step IMEX-BDF2 SAV scheme captures changes over different time scales very well, and by setting a lower ϵ, one can get a more accurate numerical result.

We further investigate how the parameters in the adaptive time-stepping strategy influence the CPU-time, H^1-error and the minimum time step-size. Table 4 shows the CPU-time, H^1-error and the minimum time step-size (since the initial step-size is very small, we only consider steps after the 100th step here) for different ϵ. In this experiment, all numerical solutions are obtained by fixing $N = 16$, $\alpha = 0.2$ and $\tau_{\max} = 5.0\text{e}{-03}$. One can observe in Table 4 that as ϵ becomes smaller, the numerical error and the minimum time-step size also become smaller, which implies one can get more accurate numerical results by lowering ϵ. However, smaller ϵ may result in longer CPU-time, and how to select appropriate ϵ in actual computation is worthy of further study.

Table 5 shows the CPU-time and H^1-error for different parameter $\tau_{\max}$ in adaptive time-stepping strategy. Numerical solutions were obtained by setting $N = 16, \alpha = 0.2$ and $\epsilon = 4.0\text{e}{-05}$.

TABLE 4. Numerical results by choosing different ϵ for solving (4.2).

ϵ	M	$\min\{\tau_{n>100}\}$	H^1-error	CPU-time (s)
1.024e−04	2184	8.2739e−05	1.1592e−05	13.44
2.560e−05	2422	4.0770e−05	1.8281e−05	13.09
6.400e−06	2944	2.0414e−05	1.3505e−05	15.93
1.600e−06	4055	1.0216e−05	7.0986e−06	22.09
4.000e−07	6364	5.0671e−06	3.1276e−06	34.36
1.000e−07	11 095	2.5346e−06	1.2053e−06	63.99

Notes. In Table 4, numerical solutions obtained by setting $N = 16$, $\alpha = 0.2$ and $\tau_{\max} = 5.0\text{e}{-03}$.

TABLE 5. Numerical results by choosing different $\tau_{\max}$ for solving (4.2).

$\tau_{\max}$	M	$\min\{\tau_{n>100}\}$	H^1-error	CPU-time (s)
$+\infty$	566	5.1389e−05	7.1947e−04	3.96
3.2e−02	564	5.1143e−05	8.2145e−04	3.67
8.0e−03	628	5.1783e−05	4.4551e−04	4.15
2.0e−03	937	5.1442e−05	1.4252e−04	5.46
5.0e−04	2321	5.1244e−05	1.7853e−05	12.62
1.25e−04	8125	5.1470e−05	8.7122e−07	57.32

Notes. In Table 5, numerical solutions obtained by setting $N = 16, \alpha = 0.2$ and $\epsilon = 4.0\text{e}{-05}$.

One can clearly observe that when the $\tau_{\max}$ gets smaller, one can have a more accurate result. By setting an appropriate $\tau_{\max}$, one can get a much smaller error at a negligible cost in CPU-time than with no $\tau_{\max}$. But, as the $\tau_{\max}$ becomes smaller, the variable time-step scheme under this strategy will gradually degenerate into a constant time-step scheme, and we will spend much more CPU-time to obtain a smaller H^1-error.

From the above results, one can see that our adaptive time-stepping strategy (4.1) is suitable for the variable time-step IMEX-BDF2 SAV scheme, which significantly improves the computational efficiency for this benchmark problem.

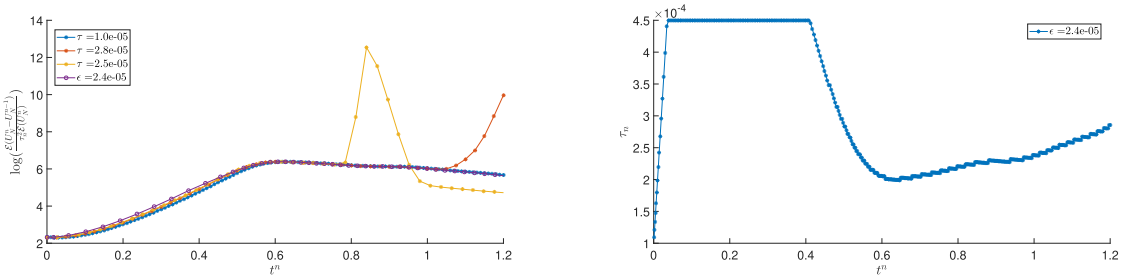
Example 3. We also use our adaptive time-stepping strategy to carry out experiments on the well known e double shear layer problem [5,11]. Consider the N-S equation (1.1) in $\Omega = (-0.5, 0.5)^2$ with periodic boundary conditions and the initial conditions given by

$$u_1(x, y, 0) = \begin{cases} \tanh(\rho(y + 0.25)), & y \le 0, \\ \tanh(\rho(0.25 - y)), & y > 0, \end{cases} \qquad u_2(x, y, 0) = -\delta \sin(2\pi x),$$

where ρ determines the slope of the shear layer and δ represents the size of the perturbation. As the same setting as [11], we fix $\delta = 0.05$, $\rho = 100$ and $\nu = 0.00005$.

In [11] it has been shown that the constant time-step IMEX-BDF2 SAV scheme can compute a reasonably correct solution at $T = 1.2$ by taking $\tau = 2.5\text{e}{-04}$ and $N = 128$, and when the step-size is $\tau = 3.0\text{e}{-04}$, the solution at $T = 1.2$ will blow up. However, through a large number of experiments and observations, we find that it may not be sufficient to use only the vorticity contours as in [11], or to use the magnitude of η deviation from 1 to determine whether the numerical results are correctly computed. In fact, if we examine $\frac{\mathcal{E}(u_t^n)}{\mathcal{E}(u^n)} \approx \frac{\mathcal{E}(U_N^n - U_N^{n-1})}{\tau_n^2 \mathcal{E}(U_N^n)}$ in the computation of the solution, when using step-size $\tau = 2.5\text{e}{-04}$, we can see that the velocity solved by the constant time-step scheme is not accurate enough at $T = 1.2$. However, by using the variable time-step IMEX-BDF2 SAV scheme, we may compute correct and accurate results with less computational cost.

In the simulations, we take $N = 128$. The problem is solved by using the variable and constant time-step schemes, respectively. With the adaptive time-stepping strategy (4.1), we choose $\tau_1 = 1.0\text{e}{-04}$ and $\tau_{\max} =$



(a) (b)

FIGURE 2. (Example 3) Experimental results by variable time-step IMEX-BDF2 SAV scheme ($\epsilon = 2.4\text{e}{-05}$) and constant time-step IMEX-BDF2 SAV scheme ($\tau = 1.0\text{e}{-04}, 2.8\text{e}{-04}, 2.5\text{e}{-04}$). (a) Relative change of velocity at different time level. (b) Time step sizes at different time level used in variable stepsize IMEX-BDF2 SAV scheme.

$4.5\text{e}{-04}$, $\epsilon = 2.4\text{e}{-05}$ and $\alpha = 0.01$ to solve the double shear layer problem until $T = 1.2$. Figure 2a plots the numerical results $\frac{\mathcal{E}(u_t^n)}{\mathcal{E}(u^n)}$. Figure 2a shows that the variable time-step scheme can compute the correct relative change of velocity on $[0, 1.2]$ with less computational cost, while the relative change of velocity computed by the constant time-step scheme with the same or even higher computational cost will blow up. For example, the numerical result of constant time-step scheme using $\tau = 2.8\text{e}{-04}$ blows up at around 0.8. Meanwhile, for the variable time-step scheme, the average step-size used in the simulation from 0 to 1.2 is $2.803\text{e}{-04}$, but it can compute the correct relative change on $[0, 1.2]$. We further consider the experiment in [11], *i.e.*, using step-size $\tau = 2.5\text{e}{-04}$ to solve this problem. Although its η^n does not deviate much from 1 [11], and the vorticity contour is correctly simulated, we can still see the relative change of velocity blows up after 1.05. However, the variable time-step scheme still works. The result of variable time-step scheme with less computational cost behaves better, and the relative change of velocity maintains correct over the time interval $[0, 1.2]$. As shown in Figure 2b, it can be seen that the variable time-step IMEX-BDF2 SAV scheme well captures the changes of different time scales in the double shear layer problem. This is also the reason why it can obtain correct numerical results at a lower computational cost than using the constant time-step scheme.

As a contrast, we solve the same problem using constant-step sizes $\tau_n = 1.0\text{e}{-04}$, $2.5\text{e}{-04}$ and $2.8\text{e}{-04}$, respectively, and the relative changes of velocity at different time level, *i.e.*, $\frac{\mathcal{E}(U_N^n - U_N^{n-1})}{\tau_n^2 \mathcal{E}(U_N^n)} \approx \frac{\mathcal{E}(u_t^n)}{\mathcal{E}(u^n)}$, are also plotted in Figure 2a. Table 6 shows the results of constant time-step IMEX-BDF2 SAV scheme by using $\tau = 1.0\text{e}{-04}$ as the reference solution, and H^1-error of velocity. It can be seen that the error of variable step-size scheme is smaller.

In order to make it easier for the readers to realize the difference in results, we compute the vorticity contours of the difference between numerical results according to the following equation

$$\nabla \times \left(U_N^M - U_{\text{ref}}^M\right).$$

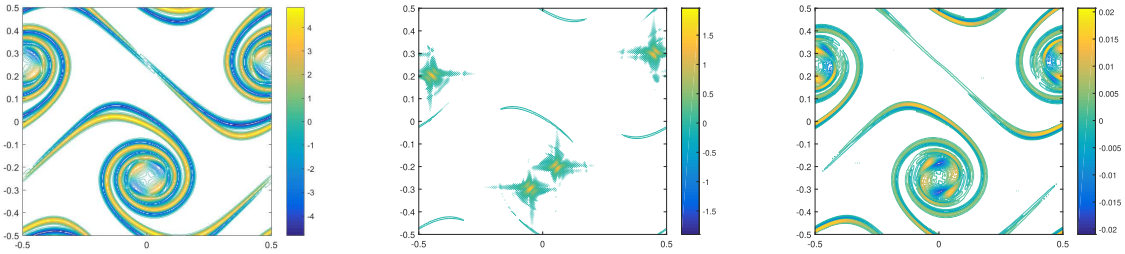
Here U_N^M are results using constant step-size $2.5\text{e}{-04}$ and $2.8\text{e}{-04}$ and variable step-size when setting $\epsilon = 2.4\text{e}{-05}$, and U_{ref}^M is the reference solution, *i.e.*, result using constant step-size $1.0\text{e}{-04}$. The Figure 3 shows the vorticity contours of the difference between numerical solution U_N^M and reference solution U_{ref}^M at $T = 1.2$.

5. Proofs of Theorems 3.1 and 3.2

We now present the proofs of Theorems 3.1 and 3.2. We carry out a complete rigorous analysis of the error estimate in the 2D case, and only present the part of the proof for the 3D case which is different from the 2D case.

TABLE 6. Errors and parameters of experiments for solving double shear layer problem.

ϵ	M	τ_{average}	H^1-error	CPU-time (s)
2.4e−05	4281	2.803e−04	9.5663e−03	284.64
–	4287	2.8e−04	7.5719e−00	275.42
–	4801	2.5e−04	5.4291e−01	312.23
–	12001	1.0e−04	–	790.29



(a) (b) (c)

FIGURE 3. (Example 3) Vorticity contours of different numerical solutions and reference solution at $T = 1.2$. (a) Constant time-step scheme using $\tau = 2.8\text{e}{-}04$. (b) Constant time-step scheme using $\tau = 2.5\text{e}{-}04$. (c) variable time-step scheme using $\epsilon = 2.4\text{e}{-}05$.

We first introduce the following denotations for errors

$$E^l := \bar{U}_N^l - u(t^l), \quad e^l := U_N^l - u(t^l), \quad e_\Pi^l := \Pi_N u(t^l) - u(t^l), \quad E_N^l := \Pi_N E^l, \quad e_N^l := \Pi_N e^l.$$

For brevity, the proof for the following error estimate is divided into three steps, and the technique of DOC kernels is mainly used for the theoretical analysis in the second step.

5.1. Proof of Theorem 3.1

Proof. Similarly as in [11], the main task is to establish by induction

$$|1 - \xi^q| \le C_0\tau + C_\Pi N^{2-m} \quad \forall q \le T/\tau. \tag{5.1}$$

Clearly $|1-\xi^0| \le C_0\tau + C_\Pi N^{2-m}$ is satisfied. Now we assume $|1-\xi^q| \le C_0\tau + C_\Pi N^{2-m}$ $(\forall q \le n)$, and we want to show $|1-\xi^{n+1}| \le C_0\tau + C_\Pi N^{2-m}$ where C_0, C_Π will be determined later.

Step 1: prove the bounds of $\|\nabla\bar{U}_N^q\|$, $\sum_{l=1}^q \tau_l\|\Delta\bar{U}_N^l\|^2$ and $\sum_{l=1}^q \tau_l\|\Delta U_N^l\|^2$, $\forall q \le n$.

When $\tau \le \frac{1}{4C_0^2} < \frac{1}{2}$, $N \ge 4C_\Pi^2 > 2$, the direct calculation from the above assumption shows that

$$\begin{aligned} 1 - \frac{1}{4C_0} - \frac{N^{3-m}}{4C_\Pi} \le |\xi^q| &\le 1 + \frac{1}{4C_0} + \frac{N^{3-m}}{4C_\Pi}, && \forall q \le n, \\ (1-\xi^q)^2 &\le \frac{\tau}{2} + \frac{N^{5-2m}}{2}, && \forall q \le n, \\ \frac{1}{2} &\le |\eta^q| < 2, && \forall q \le n. \end{aligned} \tag{5.2}$$

Noticing we have proved in Theorem 2.2 that there exists a constant $F > 0$ that is decided by γ^0 only and satisfies $\|U_N^q\|_1 \le F$, so it is easy to get the bound for $\|\bar{U}_N^q\|_1$:

$$\|\bar{U}_N^q\|_1 = \left|\frac{1}{\eta^q}\right| \|U_N^q\|_1 \le 2F, \quad \forall q \le n.$$

Besides, according to theorem 2.2, we have $\gamma^0 \ge \nu \sum_{q=1}^n \tau_q \xi^q \|\Delta \bar{U}_N^q\|^2$, which derives the bounds for $\sum_{l=1}^q \tau_l \|\Delta \bar{U}_N^l\|^2$ and $\sum_{l=1}^q \tau_l \|\Delta U_N^l\|^2$:

$$\nu \sum_{q=1}^n \tau_q \|\Delta \bar{U}_N^q\|^2 \le \frac{\gamma^0}{\min_{1\le q\le n} |\xi^q|} \le 4\gamma^0, \qquad C_0 \ge 1, \tag{5.3}$$

$$\nu \sum_{q=1}^n \tau_q \|\Delta U_N^q\|^2 \le 16\gamma^0, \qquad C_0 \ge 1. \tag{5.4}$$

Step 2: the estimates of $\|\nabla E_N^{n+1}\|$ and $\sum_{l=1}^{n+1} \tau_l \|\Delta E_N^l\|^2$.

We can get the following equality by subtracting the SAV scheme (2.10) from N-S equation (2.5):

$$(\mathcal{D}_2 E^{q+1}, v) + \nu(\nabla E^{q+1}, \nabla v) = (R_2^q, v) + (Q_2^q, v), \quad \forall v \in S_N, \quad 0 \le q \le n, \tag{5.5}$$

where

$$R_2^q = -\mathcal{D}_2 u(\cdot, t^{q+1}) + u_t(t^{q+1}) \quad \text{and} \quad Q_2^q = A(B_2 U_N^q \cdot \nabla B_2 U_N^q) - A\big(u(t^{q+1}) \cdot \nabla u(t^{q+1})\big).$$

By multiplying (5.5) by DOC kernels, and summing them from $q = 0$ to $q = l - 1$, we have

$$(\nabla_\tau E^l, v) - \nu\left(\sum_{j=1}^l \theta_{l-j}^{(l)} \Delta E^j, v\right) = \sum_{j=1}^l \left(\theta_{l-j}^{(l)}\left(R_2^{j-1} + Q_2^{j-1}\right), v\right), \quad \forall v \in S_N, \quad 1 \le l \le n+1. \tag{5.6}$$

By setting $v = -\Delta E_N^l (E_N^l := \Pi_N E^l)$, one has

$$\frac{\|\nabla E_N^l\|^2 - \|\nabla E_N^{l-1}\|^2}{2} + \nu\left(\sum_{j=1}^l \theta_{l-j}^{(l)} \Delta E_N^j, \Delta E_N^l\right) \le -\sum_{j=1}^l \left(\theta_{l-j}^{(l)}\left(R_2^{j-1} + Q_2^{j-1}\right), \Delta E_N^l\right), \quad 1 \le l \le n+1. \tag{5.7}$$

Taking the sum of (5.7) form $l = 1$ to $n + 1$, and using lemma 2.5 and corollary 2.4, we have

$$\begin{aligned} &\frac{\|\nabla E_N^{n+1}\|^2 - \|\nabla E_N^0\|^2}{2} + \frac{\nu C_r}{2} \sum_{l=1}^{n+1} \tau_l \|\Delta E_N^l\|^2 + \frac{\nu}{2} \sum_{j=1}^{n+1} \frac{\left\|\sum_{l=j}^{n+1} \theta_{l-j}^{(l)} \Delta E_N^l\right\|^2}{\tau_j} \\ &\qquad \le -\sum_{l=1}^{n+1} \sum_{j=1}^l \left(\theta_{l-j}^{(l)}\left(R_2^{j-1} + Q_2^{j-1}\right), \Delta E_N^l\right). \end{aligned} \tag{5.8}$$

For the first term of the right hand side, the truncation error R_2^{j-1} satisfies the following estimate:

$$-\sum_{l=1}^{n+1} \sum_{j=1}^l \left(\theta_{l-j}^{(l)} R_2^{j-1}, \Delta E_N^l\right) = \sum_{l=1}^{n+1} \sum_{j=1}^l \left(\theta_{l-j}^{(l)} \nabla R_2^{j-1}, \nabla E_N^l\right) \le \sum_{l=1}^{n+1} \|\nabla E_N^l\| \sum_{j=1}^l \theta_{l-j}^{(l)} \left\|\nabla R_2^{j-1}\right\|. \tag{5.9}$$

For the second term, we have the following decomposition:

$$\begin{aligned}
Q_2^q &= A(B_2U_N^q\cdot\nabla B_2U_N^q) - A\big(u\big(t^{q+1}\big)\cdot\nabla u\big(t^{q+1}\big)\big)\\
&= A(B_2U_N^q\cdot\nabla B_2U_N^q) - A(B_2u(t^q)\cdot\nabla B_2u(t^q)) + A(B_2u(t^q)\cdot\nabla B_2u(t^q)) - A\big(u\big(t^{q+1}\big)\cdot\nabla u\big(t^{q+1}\big)\big)\\
&:= Q_{21}^q + Q_{22}^q,\\
Q_{21}^q &= A(B_2U_N^q\cdot\nabla B_2(U_N^q - u(t^q))) + A(B_2(U_N^q - u(t^q))\cdot\nabla B_2u(t^q)),\\
Q_{22}^q &= A\big(\big(B_2u(t^q) - u\big(t^{q+1}\big)\big)\cdot\nabla B_2u(t^q)\big) + A\big(u(t^{q+1})\cdot\nabla\big(B_2u(t^q) - u\big(t^{q+1}\big)\big)\big).
\end{aligned} \tag{5.10}$$

Recalling the definition $e^q := U_N^q - u(t^q)$, and applying the (2.3), we have:

$$\begin{aligned}
-\sum_{l=1}^{n+1}\sum_{j=1}^{l}\Big(\theta_{l-j}^{(l)}Q_{21}^{j-1},\Delta E_N^l\Big) &= -\sum_{j=1}^{n+1}\left(Q_{21}^{j-1},\sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\Delta E_N^l\right)\\
&\le C_1\sum_{j=1}^{n+1}\Big(\Big\|B_2U_N^{j-1}\Big\|_2 + \big\|B_2u\big(t^{j-1}\big)\big\|_2\Big)\big\|B_2e^{j-1}\big\|_1\left\|\sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\Delta E_N^l\right\|\\
&\le \sum_{j=1}^{n+1}\left(\nu\frac{\big\|\sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\Delta E_N^l\big\|^2}{2\tau_j} + \frac{\tau_jC_1^2}{2\nu}\Big(\Big\|B_2U_N^{j-1}\Big\|_2 + \big\|B_2u\big(t^{j-1}\big)\big\|_2\Big)^2\big\|B_2e^{j-1}\big\|_1^2\right)\\
&\le \nu\sum_{j=1}^{n+1}\frac{\big\|\sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\Delta E_N^l\big\|^2}{2\tau_j} + \sum_{j=1}^{n}\tau_{j+1}C_2\Big(\Big\|B_2U_N^j\Big\|_2^2 + \big\|B_2u\big(t^j\big)\big\|_2^2\Big)\big\|e^{j-1}\big\|_1^2\\
&\quad + \sum_{j=1}^{n+1}\tau_jC_2\Big(\Big\|B_2U_N^{j-1}\Big\|_2^2 + \big\|B_2u\big(t^{j-1}\big)\big\|_2^2\Big)\|e^{j-1}\|_1^2,
\end{aligned} \tag{5.11}$$

where $C_2 := C_1^2(1+r_{\max})^2/\nu$.

In (5.11), the estimate of $\|e^{j-1}\|_1$ satisfies:

$$\begin{aligned}
\big\|e^{j-1}\big\|_1 = \Big\|U_N^{j-1} - u\big(t^{j-1}\big)\Big\|_1 &\le \Big\|E_N^{j-1}\Big\|_1 + \Big\|e_\Pi^{j-1}\Big\|_1 + \Big\|\bar U_N^{j-1}\big(1-\xi^{j-1}\big)^2\Big\|_1\\
&\le \Big\|E_N^{j-1}\Big\|_1 + C\big\|u\big(t^{j-1}\big)\big\|_m N^{1-m} + 4F\big(C_0^2\tau^2 + C_\Pi^2N^{4-2m}\big).
\end{aligned} \tag{5.12}$$

For the Q_{22}^q part, we have the following estimate:

$$-\sum_{l=1}^{n+1}\sum_{j=1}^{l}\theta_{l-j}^{(l)}\Big(Q_{22}^{j-1},\Delta E_N^l\Big) \le C_3\sum_{l=1}^{n+1}\big\|E_N^l\big\|_1\sum_{j=1}^{l}\theta_{l-j}^{(l)}\big\|B_2u\big(t^{j-1}\big) - u\big(t^j\big)\big\|_2\big(\big\|B_2u\big(t^{j-1}\big)\big\|_2 + \big\|u\big(t^j\big)\big\|_2\big), \tag{5.13}$$

where $C_3 = C$ is the constant in inequalities (2.3).

Substituting (5.9), (5.11)–(5.13) into (5.8), we have

$$\begin{aligned}
&\frac{\big\|\nabla E_N^{n+1}\big\|^2 - \big\|\nabla E_N^0\big\|^2}{2} + \frac{\nu\lambda}{2}\sum_{l=1}^{n+1}\tau_l\big\|\Delta E_N^l\big\|^2\\
&\quad\le \sum_{l=1}^{n+1}\big\|E_N^l\big\|_1\sum_{j=1}^{l}\theta_{l-j}^{(l)}\Big(\Big\|\nabla R_2^{j-1}\Big\| + C_3\big\|B_2u\big(t^{j-1}\big) - u\big(t^j\big)\big\|_2\big(\big\|B_2u\big(t^{j-1}\big)\big\|_2 + \big\|u\big(t^j\big)\big\|_2\big)\Big)
\end{aligned}$$

$$+2C_2\sum_{j=1}^{n}\left(\tau_{j+1}\left(\left\|B_2U_N^j\right\|_2^2+\left\|B_2u(t^j)\right\|_2^2\right)+\tau_j\left(\left\|B_2U_N^{j-1}\right\|_2^2+\left\|B_2u(t^{j-1})\right\|_2^2\right)\right)\|E_N^{j-1}\|_1^2$$
$$+C_4\left(C_0^2\tau^2+C_\Pi^2N^{4-2m}+N^{1-m}\right)^2+2C_2\tau_{n+1}\left(\|B_2U_N^n\|_2^2+\|B_2u(t^n)\|_2^2\right)\|E_N^n\|_1^2, \tag{5.14}$$

where

$$C_4:=\max\left\{64F^2C_2,4C^2\sup_{t\in[0,T]}\|u(t)\|_m^2\right\}\cdot\sum_{j=1}^{n+1}\tau_j\left(\left\|B_2U_N^{j-1}\right\|_2^2+\left\|B_2u(t^{j-1})\right\|_2^2\right).$$

We point out that C_4 is bounded thanks to (5.4).

By applying the modified Grönwall's inequality in Lemma 2.7 to (5.14), we can prove:

$$\begin{aligned}&\|E_N^{n+1}\|_1^2+\nu\lambda\sum_{l=1}^{n+1}\tau_l\|\Delta E_N^l\|^2\\&\le\exp\left(8C_2\sum_{j=1}^{n+1}\tau_j(\left\|B_2U_N^{j-1}\right\|_2^2+\left\|B_2u(t^{j-1})\right\|_2^2)\right)\left(\sqrt{C_4}(C_0^2\tau^2+C_\Pi^2N^{4-2m}+N^{1-m})\right.\\&\left.+2\sum_{l=1}^{n+1}\sum_{j=1}^{l}\theta_{l-j}^{(l)}(\|\nabla R_2^{j-1}\|+C_3(2+2r_{\max})\sup_{t\in[0,T]}\|u(t)\|_2\|B_2u(t^{j-1})-u(t^j)\|_2)\right)^2.\end{aligned}\tag{5.15}$$

The summation terms on the righthand side can be bounded as follows.

$$\begin{aligned}\sum_{j=1}^{n+1}\tau_j\left(\left\|B_2U_N^{j-1}\right\|_2^2+\left\|B_2u(t^{j-1})\right\|_2^2\right)&\le C\sum_{j=1}^{n+1}\tau_j\left(\left\|U_N^{j-1}\right\|_2^2+\left\|u(t^{j-1})\right\|_2^2\right)\\&\le C\left(r_{\max}\sum_{j=1}^{n}\tau_j\left\|U_N^j\right\|_2^2+\tau\|U_N^0\|_2^2+\sum_{j=1}^{n+1}\tau_j\|u(t^{j-1})\|_2^2\right)\\&\le C\left(r_{\max}\frac{16\gamma^0}{\nu}+r_{\max}TM^2+\tau\|u_0\|_2^2\right)+C\sum_{j=1}^{n+1}\tau_j\|u(t^{j-1})\|_2^2,\end{aligned}\tag{5.16}$$

where we used (5.4) in the last step. The last term in (5.16) can be bounded by $\int_0^T\|u_t(\xi)\|_2^2\,\mathrm{d}\xi$ and $\int_0^T\|u(\xi)\|_2^2\,\mathrm{d}\xi$ in the following way:

$$\begin{aligned}\sum_{j=1}^{n+1}\tau_j\|u(t^{j-1})\|_2^2-\int_0^{t^{n+1}}\|u(s)\|_2^2\,\mathrm{d}s&\le\sum_{j=1}^{n+1}\int_{t^{j-1}}^{t^j}\left|\|u(t^{j-1})\|_2^2-\|u(s)\|_2^2\right|\mathrm{d}s\\&\le2\sum_{j=1}^{n+1}\int_{t^{j-1}}^{t^j}\sqrt{\int_{t^{j-1}}^{t^j}\|u(\xi)\|_2^2\,\mathrm{d}\xi\int_{t^{j-1}}^{t^j}\|u_t(\xi)\|_2^2\,\mathrm{d}\xi}\,\mathrm{d}s\\&\le\tau\left[\int_0^T\|u(\xi)\|_2^2\,\mathrm{d}\xi+\int_0^T\|u_t(\xi)\|_2^2\,\mathrm{d}\xi\right].\end{aligned}$$

Hence, there exists a constan $C_5>0$ such that

$$\sum_{j=1}^{n+1}\tau_j\left(\left\|B_2U_N^{j-1}\right\|_2^2+\left\|B_2u(t^{j-1})\right\|_2^2\right)\le C_5. \tag{5.17}$$

On the other hand, thanks to Lemma 2.8, we have

$$\sum_{l=1}^{n+1}\sum_{j=1}^{l}\theta_{l-j}^{(l)}\left\|\nabla R_2^{j-1}\right\| \le \frac{\tau^2}{2}\int_0^T \|\partial_{ttt}\nabla u\|\,\mathrm{d}t + 3\tau^2\int_0^T \|\partial_{ttt}\nabla u\|\,\mathrm{d}t + (\tau_1+\tau)\int_0^{t^1}\|\partial_{tt}\nabla u\|\,\mathrm{d}t \le C_6\tau^2. \tag{5.18}$$

Similarly,

$$\begin{aligned}
\sum_{l=1}^{n+1}\sum_{j=1}^{l}\theta_{l-j}^{(l)}\big\|B_2u(t^{j-1})-u(t^j)\big\|_2 &= \sum_{j=1}^{n+1}\big\|B_2u(t^{j-1})-u(t^j)\big\|_2\sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\\
&= \big\|B_1u(t^0)-u(t^1)\big\|_2\sum_{l=1}^{n+1}\theta_{l-1}^{(l)} + \sum_{j=2}^{n+1}\big\|B_2u(t^{j-1})-u(t^j)\big\|_2\sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\\
&\le 2\tau\int_0^{t^1}\|\partial_t u\|_2\,\mathrm{d}t + 2\tau\sum_{j=2}^{n+1}(1+r_{\max})\left\|\int_{t^{j-1}}^{t^j}(s-t^{j-1})\partial_{tt}u(s)\,\mathrm{d}s\right\|_2\\
&\quad + 2\tau\sum_{j=2}^{n+1}r_{\max}\left\|\int_{t^{j-2}}^{t^j}(s-t^{j-2})\partial_{tt}u(s)\,\mathrm{d}s\right\|_2\\
&\le 2\tau\int_0^{t^1}\|\partial_t u\|_2\,\mathrm{d}t + 10(1+r_{\max})\tau^2\int_0^T\|\partial_{tt}u\|_2\,\mathrm{d}s \le C_7\tau^2.
\end{aligned} \tag{5.19}$$

Using (5.17)–(5.19), we derive from (5.15) that

$$\begin{aligned}
\big\|E_N^{n+1}\big\|_1 &\le C_8\big[(C_0^2+1)\tau^2 + C_\Pi^2N^{4-2m} + N^{1-m}\big],\\
\sum_{l=1}^{n+1}\tau_l\big\|\Delta E_N^l\big\|^2 &\le C_9\big[(C_0^2+1)\tau^2 + C_\Pi^2N^{4-2m} + N^{1-m}\big]^2,
\end{aligned}$$

where C_8, C_9 are constants independent of C_0, C_Π. By simple computation, one also finds

$$\begin{aligned}
\big\|E^{n+1}\big\|_1 &\le C_8\big[(C_0^2+1)\tau^2 + C_\Pi^2N^{4-2m} + N^{1-m}\big] + CN^{1-m},\\
\sum_{l=1}^{n+1}\tau_l\|\Delta E^l\|^2 &\le C_9\big[(C_0^2+1)\tau^2 + C_\Pi^2N^{4-2m} + N^{1-m}\big]^2 + CN^{2(2-m)},\\
\big\|\bar U_N^{n+1}\big\|_1^2, \sum_{l=1}^{n+1}\tau_l\big\|\Delta\bar U_N^l\big\|^2 &\le 2C_{10} + 2C_9\big[(C_0^2+1)\tau^2 + C_\Pi^2N^{4-2m} + N^{1-m}\big]^2 + 2CN^{2(2-m)}\\
&\le 2C_{10} + 2C_9\left(\frac{1}{8}+\frac{1}{4}+2^{3-2m}+2^{1-m}\right)^2 + 2C2^{2(2-m)} := C_{11},
\end{aligned} \tag{5.20}$$

where C_{10} is a constant depending on the exact solution $u(t)$ only. The proof here can be easily generalized to prove the boundedness of $\sum_{l=1}^{n+1}\tau_l\|\bar U_N^l\|_2^2$. The bound here is independent of C_0, C_Π and only decided by the parameters $T, \nu,\ r_{\max},\ \Omega$ and the exact solution u. For simplicity, we denote all the up bound here as C_{11}.

Step 3: Estimate of $|1-\xi^{n+1}|$.

If we take $s^n := \gamma^n - \gamma(t^n)$, then the equation for $\{s^q\}$ can be written as

$$s^{q+1}-s^q = \tau_{q+1}\nu\left(\big\|\Delta u(t^{q+1})\big\|^2 - \frac{\gamma^{q+1}}{\mathcal{E}\big(\bar U_N^{q+1}\big)+1}\big\|\Delta\bar U_N^{q+1}\big\|^2\right) + T^q, \quad \forall q \le n, \tag{5.21}$$

where T^q represents the truncation error as

$$T^q = \gamma(t^q) - \gamma(t^{q+1}) + \tau_{q+1}\gamma_t(t^{q+1}) = \int_{t^q}^{t^{q+1}} (s - t^q)\gamma_{tt}(s)\,\mathrm{d}s. \tag{5.22}$$

Taking the sum of (5.21) for q from 0 to n, and noting that $s^0 = 0$, we have

$$s^{n+1} = \nu \sum_{q=0}^{n} \tau_{q+1} \left(\left\|\Delta u\left(t^{q+1}\right)\right\|^2 - \frac{\gamma^{q+1}}{\mathcal{E}\left(\bar{U}_N^{q+1}\right) + 1} \left\|\Delta \bar{U}_N^{q+1}\right\|^2 \right) + \sum_{q=0}^{n} T^q. \tag{5.23}$$

We bound the right-hand side of (5.23) as follows. On the one hand, by direct calculation, we have

$$\gamma_{tt} = \int_\Omega \left((\nabla u)_t^2 + \nabla u (\nabla u)_{tt} \right) \mathrm{d}x, \tag{5.24}$$

then from (5.22), we have

$$|T^q| \le \tau_{q+1} \int_{t^q}^{t^{q+1}} |\gamma_{tt}|\,\mathrm{d}s \le \tau_{q+1} \int_{t^q}^{t^{q+1}} \left(\|u_t\|_1^2 + \|u_{tt}\|_1 \|u\|_1 \right) \mathrm{d}s, \quad \forall q \le n.$$

On the other hand, by triangular inequality, we have

$$\begin{aligned} &\left| \left\|\Delta u\left(t^{q+1}\right)\right\|^2 - \frac{\gamma^{q+1}}{\mathcal{E}\left(\bar{U}_N^{q+1}\right) + 1} \left\|\Delta \bar{U}_N^{q+1}\right\|^2 \right| \\ &\quad \le \left\|\Delta u\left(t^{q+1}\right)\right\|^2 \left| 1 - \frac{\gamma^{q+1}}{\mathcal{E}\left(\bar{U}_N^{q+1}\right) + 1} \right| + \frac{\gamma^{q+1}}{\mathcal{E}\left(\bar{U}_N^{q+1}\right) + 1} \left| \left\|\Delta u\left(t^{q+1}\right)\right\|^2 - \left\|\Delta \bar{U}_N^{q+1}\right\|^2 \right| \\ &\quad := K_1^q + K_2^q. \end{aligned} \tag{5.25}$$

It follows from Theorem 2.2 that

$$\begin{aligned} K_1^q &\le C_u \left| 1 - \frac{\gamma^{q+1}}{\mathcal{E}\left(\bar{U}_N^{q+1}\right) + 1} \right| \\ &= C_u \left| \frac{\gamma\left(t^{q+1}\right)}{\mathcal{E}[u(t^{q+1})] + 1} - \frac{\gamma^{q+1}}{\mathcal{E}[u(t^{q+1})] + 1} \right| + C_u \left| \frac{\gamma^{q+1}}{\mathcal{E}[u(t^{q+1})] + 1} - \frac{\gamma^{q+1}}{\mathcal{E}\left(\bar{U}_N^{q+1}\right) + 1} \right| \\ &\le C_u \left(\left| s^{q+1} \right| + \gamma^0 \left| \mathcal{E}\left[u\left(t^{q+1}\right)\right] - \mathcal{E}\left(\bar{U}_N^{q+1}\right) \right| \right), \quad \forall q \le n \end{aligned} \tag{5.26}$$

with $C_u = \sup_{t\in[0,T]} \|\Delta u(t)\|^2$, and

$$\begin{aligned} K_2^q &\le \gamma^0 \left| \left\|\Delta \bar{U}_N^{q+1}\right\|^2 - \left\|\Delta u(t^{q+1})\right\|^2 \right| \\ &\le \gamma^0 \left\|\Delta \bar{U}_N^{q+1} - \Delta u\left(t^{q+1}\right)\right\| \left(\left\|\Delta \bar{U}_N^{q+1}\right\| + \left\|\Delta u\left(t^{q+1}\right)\right\| \right) \\ &\le \gamma^0 \left\|\Delta \bar{U}_N^{q+1}\right\| \left\|\Delta E^{q+1}\right\| + \gamma^0 C_u \left\|\Delta E^{q+1}\right\|, \qquad \forall q \le n. \end{aligned} \tag{5.27}$$

We derive from the definition of $\mathcal{E}(u)$ that

$$\left|\mathcal{E}\big(u(t^{q+1})\big)-\mathcal{E}\Big(\bar{U}_N^{q+1}\Big)\right|\le\frac{1}{2}\Big(\big\|\nabla u(t^{q+1})\big\|+\Big\|\nabla\bar{U}_N^{q+1}\Big\|\Big)\Big\|\nabla u(t^{q+1})-\nabla\bar{U}_N^{q+1}\Big\|\le\sqrt{C_{11}}\big\|\nabla E^{q+1}\big\|. \tag{5.28}$$

It follows from (5.20) and the Cauchy–Schwarz inequality that

$$\begin{aligned}\sum_{q=0}^{n}\tau_{q+1}\Big\|\Delta\bar{U}_N^{q+1}\Big\|\big\|\Delta E^{q+1}\big\|&\le\left(\sum_{q=0}^{n}\tau_{q+1}\Big\|\Delta\bar{U}_N^{q+1}\Big\|^2\sum_{q=0}^{n}\tau_{q+1}\big\|\Delta E^{q+1}\big\|^2\right)^{1/2}\\&\le\sqrt{C_{11}}\sqrt{C_9[(C_0^2+1)\tau^2+C_\Pi^2N^{4-2m}+N^{1-m}]^2+CN^{2(2-m)}}.\end{aligned} \tag{5.29}$$

Now we are ready to estimate s^{n+1}. Combining (5.25)–(5.29) with (5.23), we have

$$\begin{aligned}\big|s^{n+1}\big|&\le\nu\sum_{q=0}^{n}\tau_{q+1}\left|\big\|\nabla u(t^{q+1})\big\|^2-\frac{\gamma^{q+1}}{E(\bar{U}^{q+1})+1}\Big\|\nabla\bar{U}_N^{q+1}\Big\|^2\right|+\sum_{q=0}^{n}|T^q|\\&\le\nu C_u\sum_{q=0}^{n}\tau_{q+1}\big|s^{q+1}\big|+\nu C_u\Big(\sqrt{C_{11}}+\gamma^0\Big)\sum_{q=0}^{n}\tau_{q+1}\big\|E^{q+1}\big\|_2+\nu\gamma^0\sum_{q=0}^{n}\tau_{q+1}\Big\|\Delta\bar{U}_N^{q+1}\Big\|\big\|\Delta E^{q+1}\big\|\\&\quad+\tau\int_0^{t^{n+1}}\Big(\|u_t\|_1^2+\|u_{tt}\|_1\|u\|_1\Big)\,\mathrm{d}s\\&\le\nu\Big(\gamma^0\sqrt{C_{11}}+C_u\gamma^0\sqrt{T}+C_u\sqrt{TC_{11}}\Big)\sqrt{C_9[(C_0^2+1)\tau^2+C_\Pi^2N^{4-2m}+N^{1-m}]^2+CN^{2(2-m)}}\\&\quad+\nu C_u\sum_{q=0}^{n}\tau_{q+1}\big|s^{q+1}\big|+C_u\tau.\end{aligned} \tag{5.30}$$

Finally, applying the discrete Grönwall's inequality in Lemma 2.6 to (5.30) and taking $\tau<\frac{1}{2\nu C_u}$, we obtain

$$\begin{aligned}\big|s^{n+1}\big|&\le C_{12}\exp(2\nu C_ut^n)\left(\sqrt{C_9[(C_0^2+1)\tau^2+C_\Pi^2N^{4-2m}+N^{1-m}]^2+CN^{2(2-m)}}+\tau\right)\\&\le C_{13}\sqrt{C_9}(1+C_0^2)\tau^2+C_{13}\Big(\sqrt{C_9}(C_\Pi^2N^{4-2m}+N^{1-m})+CN^{2-m}\Big)+C_{13}\tau,\end{aligned} \tag{5.31}$$

where $C_{12}:=2\max\{\nu(\gamma^0\sqrt{C_{11}}+C_u\gamma^0\sqrt{T}+C_u\sqrt{TC_{11}}),C_u\}$, and the definition of C_{13} is similar, both of the two constants are independent of τ,N,C_Π and C_0.

According to the (5.20), (5.26), (5.28), (5.31) and $m\ge3$, we have

$$\begin{aligned}\big|1-\xi^{n+1}\big|&\le C_{14}\big(\big|\mathcal{E}\big[u(t^{n+1})\big]-\mathcal{E}\big(\bar{U}_N^{n+1}\big)\big|+\big|s^{n+1}\big|\big)\\&\le C_{14}\Big(\sqrt{C_{11}}\big\|\nabla E^{n+1}\big\|+\big|s^{n+1}\big|\Big)\\&\le C_{14}\sqrt{C_{11}}\big(C_8\big[\big(C_0^2+1\big)\tau^2+C_\Pi^2N^{4-2m}+N^{1-m}\big]+CN^{1-m}\big)\\&\quad+C_{14}\Big(C_{13}\sqrt{C_9}(1+C_0^2)\tau^2+C_{13}\Big(\sqrt{C_9}(C_\Pi^2N^{4-2m}+N^{1-m})+CN^{2-m}\Big)+C_{13}\tau\Big)\\&\le C_{15}\big(1+(1+C_0^2)\tau\big)\tau+C_{15}\big(C_\Pi^2N^{2-m}+1\big)N^{2-m},\end{aligned} \tag{5.32}$$

where the constants $C_{14}:=C_u$ and C_{15} are independent of C_0, C_Π , τ and N.

We now define C_0,C_Π. We can choose $C_0=2C_{15}$ and $\tau\le\frac{1}{1+C_0^2}$ to obtain

$$C_{15}\big(1+(1+C_0^2)\tau\big)\le2C_{15}=C_0, \tag{5.33}$$

and since $m \geq 3$, we can choose $C_\Pi = 2C_{15}$ and $N \geq C_\Pi^2$ to obtain

$$C_{15}\big(C_\Pi^2 N^{2-m} + 1\big) \leq 2C_{15} = C_\Pi. \tag{5.34}$$

By mathematical induction, we proved the following inequalities

$$|1 - \xi^n| \leq C_0\tau + C_\Pi N^{2-m}, \quad \forall n \geq 1,$$

under $\tau \leq \frac{1}{2+4C_0^2}$ and $N \geq 4C_\Pi^2 + 2$. The induction process is completed.

We point out that the constraint before (5.31) needs $\tau \leq \frac{1}{2\nu C_u}$, which naturally satisfies $\tau \leq \frac{1}{2+4C_0^2}$.

By the induction process, we obtain from (5.20) following the global error estimate in 2D case:

$$\begin{aligned} \|\bar{U}_N^n - u(\cdot, t^n)\|_1 &\leq C\tau^2 + CN^{1-m}, \\ \sum_{l=1}^{n} \tau_l \|\bar{U}_N^l - u(\cdot, t^l)\|_2^2 &\leq C\tau^4 + CN^{2(2-m)}, \quad \forall n \geq 1. \end{aligned} \tag{5.35}$$

It remains to estimate e^n. We derive from (2.10) and (5.20) that

$$\big\|U_N^n - \bar{U}_N^n\big\|_1^2 \leq |\eta^n - 1|^2 \big\|\bar{U}_N^n\big\|_1^2 \leq |\eta^n - 1|^2 C_{11}, \tag{5.36}$$

and

$$\begin{aligned} \sum_{q=1}^{n} \tau_q \big\|U_N^q - \bar{U}_N^q\big\|_2^2 &\leq \sum_{q=1}^{n} \tau_q |\eta^q - 1|^2 \big\|\bar{U}_N^q\big\|_2^2 \\ &\leq \max_q |\eta^q - 1|^2 \sum_{q=1}^{n} \tau_q \Big\|\bar{U}_N^{q+1}\Big\|_2^2 \\ &\leq \max_q |\eta^q - 1|^2 C_{11}. \end{aligned} \tag{5.37}$$

The η^n in (5.36) and (5.37) holds

$$|\eta^n - 1| \leq 2C_0^2\tau^2 + 2C_\Pi^2 N^{2(2-m)}, \qquad \forall 1 \leq n. \tag{5.38}$$

Therefore, we derive from (2.10), (5.36), (5.37), (5.38) and the triangle inequality that

$$\|e^n\|_1^2 \leq 2\|E^n\|_1^2 + 2\big\|U_N^n - \bar{U}_N^n\big\|_1^2 \leq C\tau^4 + CN^{2(1-m)}, \quad \forall 1 \leq n,$$

and

$$\sum_{q=1}^{n} \tau_q \|e^q\|_2^2 \leq 2\sum_{q=1}^{n} \tau_q \|E^q\|_2^2 + 2\sum_{q=1}^{n} \tau_q \big\|U_N^q - \bar{U}_N^q\big\|_2^2 \leq C\tau^4 + CN^{2(2-m)}, \quad \forall 1 \leq n$$

under the condition $\tau \leq \frac{1}{2+4C_0^2}$ and $N \geq 4C_\Pi^2 + 2$. The proof is completed since we have proved (5.20). □

5.2. Proof of Theorem 3.2

Proof. A main difficulty in the three dimensional case is that Theorem 2.2 only provides L^2-stability for U_N^n, while in the two dimensional case, H^1-stability for U_N^n is provided in Theorem 2.2. Thus, our main task here is to prove the stability of U_N^n in H^1-norm. The rest of the proof is similar to the one for Theorem 3.1.

We proceed by induction. Assuming $|1 - \xi^q| \leq C_0\tau + C_\Pi N^{2-m}$, $\forall q \leq n$, we shall prove $|1 - \xi^{n+1}| \leq C_0\tau + C_\Pi N^{2-m}$ where C_0, C_Π will be determined below, but we may need them greater than 1 without loss of generality.

Clearly, it is satisfied for $n = 0$.

Step 1. When $\tau \le \min\{\frac{1}{4C_0^2}, \frac{1}{2}\}$, $N \ge \max\{4C_\Pi^2, 2\}$, we have (5.2), namely.

$$1 - \frac{1}{4C_0} - \frac{N^{3-m}}{4C_\Pi} \le |\xi^q| \le 1 + \frac{1}{4C_0} + \frac{N^{3-m}}{4C_\Pi}, \qquad \forall q \le n,$$
$$(1-\xi^q)^2 \le \frac{\tau}{2} + \frac{N^{5-2m}}{2}, \qquad \forall q \le n,$$
$$\frac{1}{2} \le |\eta^q| < 2, \qquad \forall q \le n.$$

Multiplying the first equation in (2.10) by DOC kernels, and summing up from $q = 0$ to $l-1$, we get

$$(\nabla_\tau \bar{U}_N^l, v) - \nu\left(\sum_{j=1}^{l} \theta_{l-j}^{(l)} \Delta \bar{U}_N^j, v\right) = \sum_{j=1}^{l}\Big(\theta_{l-j}^{(l)} A\Big(B_2\Big(U_N^{j-1}\Big)\cdot\nabla B_2\Big(U_N^{j-1}\Big)\Big), v\Big), \quad \forall v \in S_N,\ 1 \le l \le n+1. \tag{5.39}$$

Setting $v = -\Delta \bar{U}_N^l$, we get

$$\frac{\|\nabla \bar{U}_N^l\|^2 - \|\nabla \bar{U}_N^{l-1}\|^2}{2} + \nu\left(\sum_{j=1}^{l} \theta_{l-j}^{(l)} \Delta \bar{U}_N^j, \Delta \bar{U}_N^l\right) \le -\sum_{j=1}^{l}\Big(\theta_{l-j}^{(l)} A\Big(B_2\Big(U_N^{j-1}\Big)\cdot\nabla B_2\Big(U_N^{j-1}\Big)\Big), \Delta \bar{U}_N^l\Big). \tag{5.40}$$

Taking the sum of (5.40) form $l = 1$ to $n+1$, and using Lemma 2.5 and Corollary 2.4, we have

$$\begin{aligned}\frac{\|\nabla \bar{U}_N^{n+1}\|^2 - \|\nabla \bar{U}_N^0\|^2}{2} &+ \frac{\nu C_r}{2}\sum_{l=1}^{n+1}\tau_l\|\Delta \bar{U}_N^l\|^2 + \frac{\nu}{2}\sum_{j=1}^{n+1}\frac{\big\|\sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\Delta \bar{U}_N^l\big\|^2}{\tau_j} \\ &\le -\sum_{l=1}^{n+1}\sum_{j=1}^{l}\Big(\theta_{l-j}^{(l)} A\Big(B_2\Big(U_N^{j-1}\Big)\cdot\nabla B_2\Big(U_N^{j-1}\Big)\Big), \Delta \bar{U}_N^l\Big).\end{aligned} \tag{5.41}$$

Applying the inequality in (2.3), the right term of (5.8) can be estimated by

$$\begin{aligned}&-\sum_{l=1}^{n+1}\sum_{j=1}^{l}\Big(\theta_{l-j}^{(l)} A\Big(B_2\Big(U_N^{j-1}\Big)\cdot\nabla B_2\Big(U_N^{j-1}\Big)\Big), \Delta \bar{U}_N^l\Big) \\ &\quad= -\sum_{j=1}^{n+1}\left(A\Big(B_2\Big(U_N^{j-1}\Big)\cdot\nabla B_2\Big(U_N^{j-1}\Big)\Big), \sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\Delta \bar{U}_N^l\right) \\ &\quad\le C\sum_{j=1}^{n+1}\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1\Big\|\nabla B_2\Big(U_N^{j-1}\Big)\Big\|_{1/2}\Big\|\sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\Delta \bar{U}_N^l\Big\| \\ &\quad\le C\sum_{j=1}^{n+1}\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1^{3/2}\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_2^{1/2}\Big\|\sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\Delta \bar{U}_N^l\Big\| \\ &\quad\le C\sum_{j=1}^{n+1}\tau_j\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1^3\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_2 + \frac{\nu}{2}\sum_{j=1}^{n+1}\frac{\big\|\sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\Delta \bar{U}_N^l\big\|^2}{\tau_j} \\ &\quad\le C\sum_{j=1}^{n+1}\tau_j\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1^4 + C\sum_{j=1}^{n+1}\tau_j\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1^3\Big\|\Delta B_2\Big(U_N^{j-1}\Big)\Big\| + \frac{\nu}{2}\sum_{j=1}^{n+1}\frac{\big\|\sum_{l=j}^{n+1}\theta_{l-j}^{(l)}\Delta \bar{U}_N^l\big\|^2}{\tau_j}.\end{aligned} \tag{5.42}$$

Similarly,

$$\begin{aligned}\sum_{j=1}^{n+1}\tau_j\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1^3\Big\|\Delta B_2\Big(U_N^{j-1}\Big)\Big\| &\le \frac{1}{2\epsilon}\sum_{j=1}^{n+1}\tau_j\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1^6+\frac{\epsilon}{2}\sum_{j=1}^{n+1}\tau_j\Big\|\Delta B_2\Big(U_N^{j-1}\Big)\Big\|^2\\ &\le \frac{1}{2\epsilon}\sum_{j=1}^{n+1}\tau_j\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1^6+C_{r_{\max}}\frac{\epsilon}{2}\sum_{j=1}^{n+1}\tau_j\Big\|\Delta U_N^{j-1}\Big\|^2\\ &\le \frac{1}{2\epsilon}\sum_{j=1}^{n+1}\tau_j\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1^6+2C_{r_{\max}}\epsilon\sum_{j=1}^{n+1}\tau_j\Big\|\Delta \bar U_N^{j-1}\Big\|^2\\ &\le \frac{1}{2\epsilon}\sum_{j=1}^{n+1}\tau_j\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1^6+2C_{r_{\max}}\epsilon\sum_{j=1}^{n}\tau_j\Big\|\Delta \bar U_N^{j}\Big\|^2+2C_{r_{\max}}\epsilon\tau\big\|\Delta \bar U_N^0\big\|^2.\end{aligned}\tag{5.43}$$

Choosing $\epsilon=\frac{\nu C_r}{8C_{r_{\max}}}$ and combining (5.41)–(5.43) together, we have

$$\begin{aligned}&\frac{\big\|\nabla \bar U_N^{n+1}\big\|^2-\big\|\nabla \bar U_N^0\big\|^2}{2}+\frac{\nu C_r}{4}\sum_{l=1}^{n+1}\tau_l\big\|\Delta \bar U_N^l\big\|^2\\ &\quad\le C\sum_{j=1}^{n+1}\tau_j\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1^4+C\sum_{j=1}^{n+1}\tau_j\Big\|B_2\Big(U_N^{j-1}\Big)\Big\|_1^6+C\tau\big\|\Delta \bar U_N^0\big\|^2\\ &\quad\le C\sum_{j=1}^{n}\tau_j\Big\|\bar U_N^{j}\Big\|_1^4+C\sum_{j=1}^{n}\tau_j\Big\|\bar U_N^{j}\Big\|_1^6+C\tau\Big(\big\|\Delta \bar U_N^0\big\|^2+\big\|\bar U_N^0\big\|_1^4+\big\|\bar U_N^0\big\|_1^6\Big).\end{aligned}$$

Thanks to the Poincaré's inequality, there holds the following inequality:

$$\big\|\bar U_N^{n+1}\big\|_1^2+\frac{\nu C_r}{2}\sum_{l=1}^{n+1}\tau_l\big\|\Delta \bar U_N^l\big\|^2\le C\sum_{j=1}^{n}\tau_j\left(\Big\|\bar U_N^{j}\Big\|_1^4+\Big\|\bar U_N^{j}\Big\|_1^6\right)+C_u.$$

Here C_u is a constant decided by u_0. By applying lemma 2.9, and choosing $\phi(z)=z^2+z^3$, there exist $0<T_*<\int_{\mathcal{F}}^{\infty}\frac{\mathrm{d}z}{\phi(z)}$ and $C_*>0$ such that

$$\big\|\bar U_N^{q}\big\|_1^2+\frac{\nu C_r}{2}\sum_{l=1}^{q}\tau_l\big\|\Delta \bar U_N^l\big\|^2\le C_*,\qquad q\le n.$$

With the above bound, we can then prove the desired result by following similar procedures in Steps 2 and 3 in the proof of Theorem 3.1. □

6. Concluding remarks

We considered in this paper an energy stable variable time-step IMEX-BDF2 SAV scheme for the Navier–Stokes equations. Based on the energy stability, we proved the sharp global optimal error estimates by using DOC kernels through rigorous mathematical induction process. We point out that our results are obtained without any restriction on the ratio of time step and spatial grid sizes.

We introduced a suitable adaptive time-stepping strategy, and provided numerical examples to verify that the variable time-step size IMEX-BDF2 SAV scheme can effectively obtain correct numerical solutions with lower computational cost compared with a corresponding constant time-step scheme.

We only considered the analysis for the Navier–Stokes equations with periodic boundary conditions, although a similar variable time-step IMEX-BDF2 SAV scheme can be developed for the Navier–Stokes equation with non-periodic boundary conditions based on the constant step size scheme in [29]. However, the analysis for the non-periodic case is much more difficult due to the lack of *a priori* bound on the pressure approximation. But the many techniques that we developed in this paper will be useful in the future study of the variable time-step IMEX-BDF2 SAV scheme for the Navier–Stokes equation with non-periodic boundary conditions.

Acknowledgements. J. Shen was partially supported by NSFC grant 11971407. Jiwei Zhang was partially supported by in NSFC under grants No. 12171376, 2020-JCJQ-ZD-029, and Natural Science Foundation of Hubei Province No. 2019CFA007. Y. Di was partially supported by NSFC grants 12271048, 12171042, and the Guangdong Key Laboratory of Interdisciplinary Research and Application for Data Science No. 2022B1212010006.

CRediT authorship contribution statement. All authors contributed equally to this paper.

References

[1] S. Berrone and M. Marro, Space-time adaptive simulations for unsteady Navier–Stokes problems. *Comput. Fluids* **38** (2009) 1132–1144.
[2] C. Canuto and A. Quarteroni, Approximation results for orthogonal polynomials in sobolev spaces. *Math. Comput.* **38** (1982) 67–86.
[3] W. Chen, X. Wang, Y. Yan and Z. Zhang, A second order BDF numerical scheme with variable steps for the Cahn–Hilliard equation. *SIAM J. Numer. Anal.* **57** (2019) 495–525.
[4] M.O. Deville, P.F. Fischer and E.H. Mund, High-Order Methods for Incompressible Fluid Flow. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press (2002).
[5] Y. Di, R. Li, T. Tang and P. Zhang, Moving mesh finite element methods for the incompressible Navier–Stokes equations. *SIAM J. Sci. Comput.* **26** (2005) 1036–1056.
[6] L. Failer and T. Wick, Adaptive time-step control for nonlinear fluid-structure interaction. *J. Comput. Phys.* **366** (2018) 448–477.
[7] V. Girault and P.A. Raviart, Finite Element Approximation of the Navier–Stokes Equations. Springer, Berlin, Heidelberg (1979).
[8] R. Glowinski, Finite element methods for incompressible viscous flow. *Handb. Numer. Anal.* **9** (2003) 3–1176.
[9] M.D. Gunzburger, Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice, and Algorithms. *Computer Science and Scientific Computing*. Elsevier, San Diego (1989).
[10] A. Hay, S. Etienne, A. Garon and D. Pelletier, Time-integration for ALEsimulations of fluid-structure interaction problems: stepsize and order selection based on the BDF. *Comput. Methods Appl. Mech. Eng.* **295** (2015) 172–195.
[11] F. Huang and J. Shen, Stability and error analysis of a class of high-order IMEX schemes for Navier–Stokes equations with periodic boundary conditions. *SIAM J. Numer. Anal.* **59** (2021) 2926–2954.
[12] G. Jannoun, E. Hachem, J. Veysset and T. Coupez, Anisotropic meshing with time-stepping control for unsteady convection-dominated problems. *Appl. Math. Modell.* **39** (2015) 1899–1916.
[13] D.A. Kay, P.M. Gresho, D. Griffiths and D.J. Silvester, Adaptive time-stepping for incompressible flow. Part II: Navier–Stokes equations. *SIAM J. Sci. Comput.* **32** (2010) 111–128.
[14] H. Liao, T. Tang and T. Zhou, On energy stable, maximum-principle preserving,second-order BDF scheme with variable steps for the Allen-Cahn equation. *SIAM J. Numer. Anal.* **58** (2020) 2294–2314.
[15] H. Liao, X. Song, T. Tang and T. Zhou, Analysis of the second-order BDF scheme with variable steps for the molecular beam epitaxial model without slope selection. *Sci. China Math.* **64** (2021) 887–902.
[16] H. Liao, B. Ji and L. Zhang, An adaptive BDF2 implicit time-stepping method for the phase field crystal model. *IMA J. Numer. Anal.* **42** (2022) 649–679.
[17] J. Liu and R.L. Pego, Stable discretization of magnetohydrodynamics in bounded domains. *Commun. Math. Sci.* **8** (2010) 235–251.
[18] Y. Ma, J. Zhang and C. Zhao, The unconditionally optimal H1-norm error estimate of a semi-implicit galerkin FEMs VSBDF2 scheme for solving semilinear parabolic equations. Preprint (2022).
[19] P. Moin and K. Mahesh, Direct numerical simulation: a tool in turbulence research. *Ann. Rev. Fluid Mech.* **30** (1998) 539–578.
[20] S.A. Orszag and G.S. Patterson, Numerical simulation of three-dimensional homogeneous isotropic turbulence. *Phys. Rev. Lett.* **28** (1972) 76–79.
[21] R. Peyret, Spectral Methods for Incompressible Viscous Flow. Springer (2002).
[22] Z. Qiao, Z. Zhang and T. Tang, An adaptive time-stepping strategy for the molecular beam epitaxy models. *SIAM J. Sci. Comput.* **33** (2011) 1395–1414.
[23] Z. Qiao, Z. Sun and Z. Zhang, The stability and convergence of two linearized finite difference schemes for the nonlinear epitaxial growth model. *Numer. Methods Part. Differ. Equ.* **28** (2012) 1893–1915.

[24] Z. She, E. Jackson and S.A. Orszag, Structure and dynamics of homogeneous turbulence: models and simulations. *Proc. R. Soc. London. Ser. A: Math. Phys. Sci.* **434** (1991) 101–124.
[25] R. Temam, Navier–Stokes Equations and Nonlinear Functional Analysis. SIAM, Philadelphia (1982).
[26] W. Wang, Y. Chen and H. Fang, On the variable two-step imex BDF method for parabolic integro-differential equations with nonsmooth initial data arising in finance. *SIAM J. Numer. Anal.* **57** (2019) 1289–1217.
[27] W. Wang, M. Mao and Z. Wang, Stability and error estimates for the variable step-size BDF2 method for linear and semilinear parabolic equations. *Adv. Comput. Math.* **47** (2021) 1–28.
[28] W. Wang, Z. Wang, and M. Mao, Linearly implicit variable step-size BDF schemes with fourier pseudospectral approximation for incompressible Navier–Stokes equations. *Appl. Numer. Math.* **172** (2022) 393–412.
[29] K. Wu, F. Huang and J. Shen, A new class of higher-order decoupled schemes for the incompressible Navier–Stokes equations and applications to rotating dynamics. *J. Comput. Phys.* **458** (2022) 16.
[30] J. Zhang and C. Zhao, Sharp error estimate of BDF2 scheme with variable time steps for linear reaction-diffusion equations. *J. Math.* **41** (2021) 471–488.
[31] C. Zhao, L. Liu, Y. Ma and J. Zhang, Unconditionally optimal error estimate of a linearized variable-time-step BDF2 scheme for nonlinear parabolic equations. Preprint: arxiv:2201.06008 (2022).