Vol. 42 (2022)
No. 5

数 学 杂 志
J. of Math. (PRC)

SHARP ERROR ESTIMATE OF BDF2 SCHEME WITH VARIABLE TIME STEPS FOR MOLECULAR BEAM EPITAXIAL MODELS WITHOUT SLOP SELECTION

ZHANG Ji-wei[1], ZHAO Cheng-chao[2]

(*1. School of Mathematics and Statistics, and Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan 430072, China*)

(*2. Beijing Computational Science Research Center, Beijing 100193, China*)

Abstract: The stability and convergence of two-step backward differentiation formula (BDF2) with variable time steps still remain incomplete for solving the molecular beam epitaxial model without slope selection. In this paper, we first prove the proposed BDF2 scheme to preserve a modified energy dissipation law under a new adjacent time-step ratio condition: $r_k := \tau_k/\tau_{k-1} \leq 4.8645 - \delta$, where $\delta > 0$ is a given arbitrarily small constant. After that, we introduce the recently developed techniques of the discrete orthogonal convolution (DOC) and discrete complementary convolution (DCC) kernels, and present the robust and sharp second-order convergence of the BDF2 scheme with the new ratio condition: $r_k \leq 4.8645 - \delta$. The robustness means the convergence does not need other constrained condition on the time steps except for $r_k \leq 4.8645 - \delta$. In addition, our analysis shows that the first-order BDF1 scheme for the start step is enough to ensure the globally optimal convergence order. This is, the choice of BDF1 scheme for the start step does not bring the loss of global second-order convergence. Numerical examples are provided to demonstrate the theoretical analysis.

Keywords: BDF2 with variable time steps; the discrete orthogonal convolution (DOC) kernels; the discrete complementary convolution (DCC) kernels; the error convolution structure (ECS); sharp error estimate; MBE models

2010 MR Subject Classification: 65M06; 65M12.

Document code: A **Article ID:** 0255-7797(2022)05-0377-25

1 Introduction

In this paper, we revisit the two-step backward differentiation formula (BDF2) with variable time-steps for solving the molecular beam epitaxial (MBE) model [6, 9] without slope selection

$$\begin{aligned} &u_t + \varepsilon\Delta^2 u + \nabla\cdot\boldsymbol{f}(\nabla u) = 0, \quad \mathbf{x}\in\Omega, t\in(0,T], \\ &u(\boldsymbol{x},0) = u_0(\boldsymbol{x}), \qquad\qquad\quad \boldsymbol{x}\in\bar{\Omega}, \end{aligned} \tag{1.1}$$

* **Received date:** 2022-08-20 **Accepted date:** 2022-09-10

Foundation item: Supported by NSFC (12171376, 2020-JCJQ-ZD-029); Natural Science Foundation of Hubei Province (2019CFA007); the Fundamental Research Funds for the Central Universities (2042021kf0050).

Biography: Zhang Jiwei (1979 –), male, professor, major in the computational and applied mathematics.

with the periodic boundary conditions. Here the periodic solution $u = u(\boldsymbol{x}, t)$ represents the scaled height function of a thin film in a co-moving frame, the fourth-order term models surface diffusion with a surface diffusion constant $\varepsilon > 0$ and the nonlinear force vector $\boldsymbol{f}(\mathbf{v}) := \mathbf{v}/(1 + |\mathbf{v}|^2)$ models the well-known Ehrlich-Schwoebel effect.

The MBE model (1.1) has been widely applied in various fields such as physics, biology, ecology and chemistry [7, 22], and can be derived from the gradient flow with the following energy functional in the $L^2(\Omega)$ inner product

$$E(u) = \int_\Omega \left(\frac{\varepsilon}{2} |\Delta u|^2 - \frac{1}{2} \ln(1 + |\nabla u|^2) \right) \mathrm{d}\boldsymbol{x}. \tag{1.2}$$

The logarithmic term $-\frac{1}{2}\ln(1 + |\nabla u|^2)$ in the energy functional (1.2) can be bounded by zero but unbounded blew, which means the logarithmic term has no relative minima. The well-posedness of problem (1.1) is studied by Li and Liu [16] using the perturbation analysis and Galerkin spectral approximations.

Recently, to investigate the evolution process of thin-film epitaxial growth, various numerical schemes for MBE model (1.1) have been developed including the first and second order convex splitting schemes [2, 24], the nonlinear Crank-Nicolson type scheme [21], the stabilized semi-implicit scheme [28] and so on. However, the analysis in those mentioned literatures was based on uniform time steps.

A feature of phase field models is that the solutions admit multiple time scales, namely, the dynamics evolves on a fast time scale at the beginning and coarsening evolves slowly on a time later. In this situation, the coarse-grained and refined time steps are useful to capture the multi-scale dynamics according to the slow and fast change of the solution itself. Thus, the BDF2 scheme with variable time steps is a good choice due to its strong stability for solving stiff or differential-algebraic problems [5, 10, 11, 23, 25, 26].

The BDF2 scheme with variable time steps has been widely developed [1, 3, 5, 13, 20, 29] for the stability and convergence analysis, including linear diffusion problems [20, 29], semilinear parabolic problems [5, 13] and the Cahn-Hilliard (CH) equation [3]. More specifically for the stability analysis of linear diffusion equations, twenty years ago Becker [1] presented the bound under the adjacent time-step ratio condition $0 < r_k := \tau_k/\tau_{k-1} \leq (2 + \sqrt{13})/3 \approx 1.868$ that

$$\|u^n\| \leq C \exp(C\Gamma_n) \big(\|u_0\| + \sum_{j=1}^{n} \tau_j \left\| f^j \right\| \big) \quad \text{for} \quad n \geq 1,$$

where $\Gamma_n := \sum_{k=2}^{n-2} \max\{0, r_k - r_{k+2}\}$. The result is also given in Thomée's classical book [25, Lemma 10.6]. As shown in [25] and [3], the magnitudes of Γ_n can be bounded [25, pp. 175] and unbounded [3, Remark 4.1] by choosing certain step-ratio sequence and vanishing step sizes. Emmrich [5] extends the Becker's condition to $0 < r_k \leq 1.91$, but still keeps the undesirable factor $\exp(C\Gamma_n)$. Recently, Liao and Zhang [20] introduce the technique of the discrete orthogonal convolution (DOC) kernels, and improve Grigorieff's stability condition

[8] nearly forty years ago (one also refers to [4] and [11, Section III.5] a classical book by Hairer *et al.*) from $0 \le r_k \le 1+\sqrt{2}$ to $0 \le r_k \le (3+\sqrt{17})/2 \approx 3.561$. However, the second-order convergence in [20] suffers from an extra restriction condition $|\mathfrak{R}_p| \le N_0 \ll N$ with the index set

$$\mathfrak{R}_p = \left\{ k \;\middle|\; 1+\sqrt{2} \le r_k \le (3+\sqrt{17})/2 \right\}. \tag{1.3}$$

While the stability and convergence analysis of BDF2 with variable time steps has brought the great challenge for linear problems, the analysis for nonlinear problems is even hard and still has a great progress. For instance, Chen *et al.* [3] replace $\exp(C\mathrm{T}_n)$ in Becker's estimate with a bounded factor $\exp(Ct_n)$ with $0 < r_k \le 1.53$ for CH equations. Liao *et al.* [19] consider MBE model (1.1) with variable-time-steps BDF2 scheme, and obtain the second-order convergence under the ratio condition $0 < r_k < 3.561$, but they still require an additional condition $|\mathfrak{R}_p| \le N_0 \ll N$.

The aim of this paper is to achieve the robust and sharp second-order error estimate for the variable time-steps BDF2 scheme under a new ratio condition $0 < r_k < r_{\max} \approx 4.8645$. Under this new ratio condition, we first prove the BDF2 scheme with BDF1 as starting step to preserve a modified energy dissipation law. After that, we carefully analyze the positive definiteness of discrete convolution kernels [20], and then introduce the discrete complementary convolution (DCC) kernels (defined in (4.21)) and the error convolution structure (**ECS**) with the BDF2 kernels (see Lemma 5.3), and finally obtain the sharp second-order convergence given as

$$\begin{aligned}\|e_h^n\| \le & 2\exp(16\mathcal{Q}_\delta^2 t_{n-1}/\varepsilon)(\|e_h^0\| + 2C_u t_n h^2 + \sum_{k=1}^{n} \tau_k^2 \int_{t_{k-1}}^{t_k} \|u_{ttt}\|\,\mathrm{d}t \\ & + 2t_n \max_{1\le k\le n} \tau_k \int_{t_{k-1}}^{t_k} \|u_{ttt}\|\,\mathrm{d}t + 4\tau \int_0^{t_1} \|u_{tt}\|\,\mathrm{d}t).\end{aligned} \tag{1.4}$$

For brevity, we list the adjacent time step ratio condition as

$$\textbf{A1}: \quad 0 < r_k \le r_{\max} - \delta \quad \text{for any small constant } 0 < \delta < r_{\max} \text{ and } 2 \le k \le N,$$

where the maximum ratio $r_{\max} = \frac{1}{6}\left(\sqrt[3]{1196-12\sqrt{177}} + \sqrt[3]{1196+12\sqrt{177}}\right) + \frac{4}{3} \approx 4.8645$ is the root of the cubic equation

$$r_{\max}^3 = (2r_{\max}+1)^2. \tag{1.5}$$

The $\mathcal{Q}_\delta$ in (1.4) is a constant depending on the choice of adjacent time steps, and has a upper bound of the form $\mathcal{O}(1/\delta)$, where the parameter δ is given in **A1** and generally taken as a given small constant, for example $\delta = 0.1$ or other any small constant.

Comparing with recent results in [19], our second-order convergence is sharp and robust with the new ratio condition **A1**. The robustness means the convergence does not suffer from other extra conditions on the time step sizes, like the constrained condition $|\mathfrak{R}_p| \le N_0 \ll N$ in [19], expect for **A1**. In addition, our analysis shows that the sharp second-order convergence

is consistent to the first-order BDF1 scheme for the first step solution u^1. It is the first time to make clear that the BDF1 scheme as start step to compute u^1 is enough to guarantee the second-order convergence of BDF2 schemes with variable time steps for MBE models. Numerical examples are provided to demonstrate our theoretical analysis.

The remainder is organized as follows. In section 2, we present the fully discrete scheme with variable time steps by using the finite difference method in space and BDF2 scheme in time. The solvability of the BDF2 scheme and the energy stability are presented in section 3. In section 4, we introduce the concepts of DOC and DCC kernels, and also present the properties of DOC kernels and DCC kernels. In section 5, we give the stability and second-order convergence analysis. Numerical simulations are carried out in section 6. We end the paper with a conclusion.

2 Setting

2.1 Numerical scheme

We take the generally variable time grids $0 = t_0 < t_1 < t_2 < \cdots < t_N = T$ and denote the kth time-step size by $\tau_k := t_k - t_{k-1}$ and the maximum time step size by $\tau := \max_{1\le k\le N} \tau_k$. The adjacent time-step ratio is defined by

$$r_k = \frac{\tau_k}{\tau_{k-1}}, \qquad 2 \le k \le N.$$

Set $u^k = u(\cdot, t_k)$ and the difference operator $\nabla_\tau u^k = u^k - u^{k-1}$ for $1 \le k \le N$. The BDF1 and BDF2 formulas with variable time steps are defined respectively by

$$\mathcal{D}_1 u^n = \frac{1}{\tau_n}\nabla_\tau u^n, \quad \mathcal{D}_2 u^n = \frac{1+2r_n}{\tau_n(1+r_n)}\nabla_\tau u^n - \frac{r_n^2}{\tau_n(1+r_n)}\nabla_\tau u^{n-1}.$$

Set the discrete convolution kernels $b_{n-k}^{(n)}$ as $b_0^{(1)} := 1/\tau_1$ and

$$b_0^{(n)} = \frac{1+2r_n}{\tau_n(1+r_n)}, \quad b_1^{(n)} = -\frac{r_n^2}{\tau_n(1+r_n)} \quad \text{and} \quad b_j^{(n)} = 0 \text{ for } n, j \ge 2. \tag{2.6}$$

Thus, we may reformulate the BDF1 and BDF2 into a unified discrete convolution form

$$\mathcal{D}_2 u^n := \sum_{k=1}^{n} b_{n-k}^{(n)} \nabla_\tau u^k, \qquad n \ge 1. \tag{2.7}$$

The spacial domain $\Omega = (0, L)^2$ considered here is approximated by a uniform grid $h = L/M$ for a positive integer M, and the discrete domains are denoted by

$$\Omega_h := \{\boldsymbol{x}_h : (ih, jh), 1 \le i, j \le M-1\}, \quad \bar{\Omega}_h := \{\boldsymbol{x}_h : (ih, jh), 0 \le i, j \le M\}.$$

The partial derivatives $\partial_x w$ and $\partial_{xx} w$ are respectively approximated by the following operators

$$\Delta_x w_{i,j} := (w_{i+1,j} - w_{i-1,j})/(2h), \quad \delta_x^2 := (w_{i+1,j} - 2w_{i,j} + w_{i-1,j})/h^2.$$

The operators $\Delta_y w_{i,j}$ and $\delta_y^2 w_{i,j}$ can be defined similarly. Moreover, the discrete gradient operator and the discrete Laplacian operator are accordingly defined by

$$\nabla_h w_{i,j} := (\Delta_x w_{i,j}, \Delta_y w_{i,j})^T, \quad \Delta_h w_{i,j} := (\delta_x^2 + \delta_y^2) w_{i,j}.$$

For the vector $\boldsymbol{u}_{i,j} = (u_{i,j}^1, u_{i,j}^2)^T$, the discrete divergence is defined by

$$\nabla_h \cdot \boldsymbol{u}_{i,j} := \Delta_x u_{i,j}^1 + \Delta_y u_{i,j}^2.$$

By using the finite difference method in space and BDF2 scheme in time, we have the fully discrete scheme with variable time steps as

$$\mathcal{D}_2 u_h^n + \varepsilon \Delta_h^2 u_h^n + \nabla_h \cdot \boldsymbol{f}(\nabla_h u_h^n) = 0, \quad \text{for} \quad 1 \le n \le N. \tag{2.8}$$

3 Solvability and Energy Stability

Define the space of L-periodic grid functions as

$$\mathcal{V}_h = \{v_h | v_h \text{ is L-periodic for } \mathbf{x}_h \in \bar{\Omega}_h\}.$$

For any $v, w \in \mathcal{V}_h$, the discrete L^2 inner product and norm are defined by

$$\langle v, w \rangle := h^2 \sum_{\mathbf{x}_h \in \Omega_h} v_h w_h, \quad \left\| v \right\|^2 := \langle v, v \rangle .$$

The discrete norms $\left\| \nabla_h v \right\|$ and $\left\| \Delta_h v \right\|$ are defined by

$$\|\nabla_h v\| := \sqrt{h^2 \sum_{\boldsymbol{x}_h \in \Omega_h} |\nabla_h v_h|^2}, \quad \|\Delta_h v\| := \sqrt{h^2 \sum_{\boldsymbol{x}_h \in \Omega_h} |\Delta_h v_h|^2}.$$

For any $v, w \in \mathcal{V}_h$, one has the discrete Green's formula with periodic boundary conditions

$$\langle -\nabla_h \cdot \nabla_h v, w \rangle = \langle \nabla_h v, \nabla_h w \rangle. \tag{3.9}$$

Lemma 3.1 ([21]) For any grid function $v \in \mathcal{V}_h$ and $\epsilon > 0$, we have

$$\|\nabla_h v\|^2 \le \langle -\Delta_h v, v \rangle \le \|\Delta_h v\| \|v\| \le \epsilon \|\Delta_h v\|^2 + \frac{1}{4\epsilon} \|v\|^2. \tag{3.10}$$

3.1 Unique Solvability

Theorem 3.2 If the time-step sizes $\tau_n \le 4\varepsilon$, the BDF2 time-stepping scheme (2.8) is convex and uniquely solvable.

The solvability of BDF2 scheme can be established by introducing a discrete energy functional G on the space $\mathcal{V}_h$:

$$G[z] = \frac{1}{2} b_0^{(n)} \left\| z - u_h^{n-1} \right\|^2 + b_1^{(n)} \left\langle \nabla_\tau u^{n-1}, z \right\rangle \frac{\varepsilon}{2} \left\| \Delta_h z \right\|^2 - \frac{1}{2} \langle \ln(1 + |\nabla_h z|^2), 1 \rangle.$$

The detailed proof for the solvability of BDF2 scheme is given in [19] and the key technique is referred to [27] .

3.2 Discrete Energy Dissipation Law

We now consider the energy stability of BDF2 scheme (2.8). To do so, we first present the positive definiteness of discrete convolution kernels $b_{n-k}^{(n)}$ in the following lemma.

Lemma 3.2 Assume the time-step ratio r_k satisfies **A1** . For any real sequence $\{w_k\}_{k=1}^n$, it holds for $\epsilon_* = \sqrt[3]{12}\big(\sqrt[3]{\sqrt{177}+9} - \sqrt[3]{\sqrt{177}-9}\,\big)/6 \approx 0.4534$ and for any small constant $0 < \delta < r_{\max}$ (see **A1**) that

$$2w_k \sum_{j=1}^{k} b_{k-j}^{(k)} w_j \geq \frac{r_{k+1}}{(1+r_{k+1})} \frac{w_k^2}{\epsilon_* \tau_k} - \frac{r_k}{(1+r_k)} \frac{w_{k-1}^2}{\epsilon_* \tau_{k-1}} + \frac{\delta w_k^2}{(1+r_{\max})^2 \epsilon_* \tau_k}, \quad k \geq 2, \tag{3.11}$$

$$2\sum_{k=1}^{n} w_k \sum_{j=1}^{k} b_{k-j}^{(k)} w_j \geq \sum_{k=1}^{n} \frac{\delta w_k^2}{(1+r_{\max})^2 \epsilon_* \tau_k} \geq 0, \quad \text{for } n \geq 1. \tag{3.12}$$

Proof Denote the multi-variable functions

$$\mathfrak{F}(x, y, \epsilon) = \frac{2\epsilon + 4\epsilon x - \epsilon^2 x^2}{(1+x)} - \frac{y}{(1+y)}, \quad \text{for } x, y, \epsilon \geq 0.$$

It follows from [29, Lemma 2.1] and the proof of [29, Lemma 2.2] that

$$2w_k \sum_{j=1}^{k} b_{k-j}^{(k)} w_j \geq \frac{r_{k+1}}{(1+r_{k+1})} \frac{w_k^2}{\epsilon_* \tau_k} - \frac{r_k}{(1+r_k)} \frac{w_{k-1}^2}{\epsilon_* \tau_{k-1}} + \mathfrak{F}(r_k, r_{k+1}, \epsilon_*) \frac{w_k^2}{\epsilon_* \tau_k}, \quad k \geq 2, \tag{3.13}$$

and

$$\frac{2\epsilon_* + 4\epsilon_* r_k - \epsilon_*^2 r_k^2}{1+r_k} \geq \frac{r_{\max}}{1+r_{\max}}, \quad \forall\, 0 < r_k \leq r_{\max}.$$

Hence, for any $0 < r_k \leq r_{\max} - \delta$, one has

$$\mathfrak{F}(r_k, r_{k+1}, \epsilon_*) \geq \frac{r_{\max}}{1+r_{\max}} - \frac{r_{\max} - \delta}{1 + r_{\max} - \delta} = \frac{\delta}{(1+r_{\max}-\delta)(1+r_{\max})} \geq \frac{\delta}{(1+r_{\max})^2}, \quad k \geq 2$$

where the monotony of the function $h(x) = x/(1+x)$ is used. Inserting above inequality to (3.13), one immediately has the inequality (3.11). Summing the inequality (3.11) from 1 to n, one has

$$\begin{aligned}
2\sum_{k=1}^{n} w_k \sum_{j=1}^{k} b_{k-j}^{(k)} w_j &\geq \frac{2}{\tau_1} w_1^2 + \frac{r_{n+1}}{(1+r_{n+1})} \frac{w_n^2}{\epsilon_* \tau_n} - \frac{r_2}{(1+r_2)} \frac{w_1^2}{\epsilon_* \tau_1} + \sum_{k=2}^{n} \frac{\delta w_k^2}{(1+r_{\max})^2 \epsilon_* \tau_k} \\
&\geq \left(2\epsilon_* - \frac{r_{\max} - \delta}{1 + r_{\max} - \delta}\right) \frac{w_1^2}{\epsilon_* \tau_1} + \sum_{k=2}^{n} \frac{\delta w_k^2}{(1+r_{\max})^2 \epsilon_* \tau_k} \\
&\geq \left(2 - \frac{r_{\max}}{(1+r_{\max})\epsilon_*}\right) \frac{w_1^2}{\tau_1} + \sum_{k=1}^{n} \frac{\delta w_k^2}{(1+r_{\max})^2 \epsilon_* \tau_k} \\
&\geq \sum_{k=1}^{n} \frac{\delta w_k^2}{(1+r_{\max})^2 \epsilon_* \tau_k}, \quad n \geq 1.
\end{aligned}$$

The proof is complete.

We now consider the energy stability of BDF2 scheme (2.8) by defining the modified discrete energy

$$E^n := \frac{r_{n+1}}{2(1+r_{n+1})\epsilon_*\tau_n}\|\nabla_\tau u^n\|^2 + \frac{\varepsilon}{2}\|\Delta_h u^n\|^2 - \frac{1}{2}\langle \ln(1+|\nabla_h u^n|^2), 1\rangle, \quad n \geq 1, \tag{3.14}$$

and the initial energy $E^0 := \frac{\varepsilon}{2}\|\Delta_h u^0\|^2 - \frac{1}{2}\langle \ln(1+|\nabla_h u^0|^2), 1\rangle$.

To establish the energy dissipation law, we need the time-step ratio r_k to hold **A1** and the time step size τ_n to satisfy

$$\tau_n \leq 4\varepsilon \min\{2 - \frac{r_{\max}}{(1+r_{\max})\epsilon_*}, \frac{\delta}{(1+r_{\max})^2\epsilon_*}\}. \tag{3.15}$$

Theorem 3.2 Assume the time-step ratio condition **A1** holds with the time-step condition (3.15), then the discrete energy E_n defined in (3.14) satisfies

$$E^n \leq E^{n-1} \leq E^0, \quad n \geq 1.$$

Proof Taking the inner product on both sides of (2.8) with $\nabla_\tau u^n$, one has

$$\langle \mathcal{D}_2 u^n, \nabla_\tau u^n\rangle + \varepsilon\langle \Delta_h^2 u^n, \nabla_\tau u^n\rangle + \langle \nabla_h \cdot \boldsymbol{f}(\nabla_h u^n), \nabla_\tau u^n\rangle = 0, \quad \text{for} \quad 1 \leq n \leq N. \tag{3.16}$$

Due to the periodic boundary conditions, the summation by parts argument holds, which implies

$$\varepsilon\langle \Delta_h^2 u^n, \nabla_\tau u^n\rangle = \varepsilon\langle \Delta_h u^n, \Delta_h \nabla_\tau u^n\rangle = \frac{\varepsilon}{2}(\|\Delta_h u^n\|^2 - \|\Delta_h u^{n-1}\|^2 + \|\Delta_h \nabla_\tau u^n\|^2). \tag{3.17}$$

By using the inequality $\frac{x}{1+x} \leq \ln(1+x)$ with $x = (\nabla_\tau|\nabla_h u^n|^2)/(1+|\nabla_h u^{n-1}|^2)$, one has

$$\frac{\nabla_\tau|\nabla_h u^n|^2}{1+|\nabla_h u^n|^2} \leq \ln(\frac{1+|\nabla_h u^n|^2}{1+|\nabla_h u^{n-1}|^2}),$$

which together with the discrete Green's formula (3.9) and the inequality (3.10) with $\epsilon = \varepsilon$ imply that

$$\begin{aligned}
&\langle \nabla_h \cdot \boldsymbol{f}(\nabla_h u^n), \nabla_\tau u^n\rangle = -\langle \boldsymbol{f}(\nabla_h u^n), \nabla_h \nabla_\tau u^n\rangle \\
=&-\langle \frac{\nabla_\tau|\nabla_h u^n|^2}{2(1+|\nabla_h u^n|^2)}, 1\rangle - \langle \frac{|\nabla_h \nabla_\tau u^n|^2}{2(1+|\nabla_h u^n|^2)}, 1\rangle \\
\geq& -\frac{1}{2}\langle \ln(1+|\nabla_h u^n|^2), 1\rangle + \frac{1}{2}\langle \ln(1+|\nabla_h u^{n-1}|^2), 1\rangle - \frac{1}{2}\|\nabla_h \nabla_\tau u^n\|^2 \\
\geq& -\frac{1}{2}\langle \ln(1+|\nabla_h u^n|^2), 1\rangle + \frac{1}{2}\langle \ln(1+|\nabla_h u^{n-1}|^2), 1\rangle - \frac{\varepsilon}{2}\|\Delta_h \nabla_\tau u^n\|^2 - \frac{1}{8\varepsilon}\|\nabla_\tau u^n\|^2.
\end{aligned} \tag{3.18}$$

For $n \geq 2$, it follows from Lemma 3.2 and the time-step condition (3.15) that

$$\begin{aligned}
\langle \mathcal{D}_2 u^n, \nabla_\tau u^n\rangle &\geq \frac{r_{n+1}}{2(1+r_{n+1})}\frac{\|\nabla_\tau u^n\|^2}{\epsilon_*\tau_n} - \frac{r_n}{2(1+r_n)}\frac{\|\nabla_\tau u^{n-1}\|^2}{\epsilon_*\tau_{n-1}} + \frac{\delta\|\nabla_\tau u^n\|^2}{2(1+r_{\max})^2\epsilon_*\tau_n} \\
&\geq \frac{r_{n+1}}{2(1+r_{n+1})}\frac{\|\nabla_\tau u^n\|^2}{\epsilon_*\tau_n} - \frac{r_n}{2(1+r_n)}\frac{\|\nabla_\tau u^{n-1}\|^2}{\epsilon_*\tau_{n-1}} + \frac{1}{8\varepsilon}\|\nabla_\tau u^n\|^2.
\end{aligned} \tag{3.19}$$

Hence, by inserting (3.17)-(3.19) into (3.16), we have

$$E^n \leq E^{n-1}, \quad n \geq 2.$$

For $n = 1$, it follows from the condition **A1** that

$$\frac{r_2}{(1+r_2)\epsilon_*} \leq \frac{r_{\max}}{(1+r_{\max})\epsilon_*} \leq 2,$$

which together with the time step condition (3.15) imply that

$$\langle \mathcal{D}_2 u^1, \nabla_\tau u^1 \rangle = \frac{1}{\tau_1}\|\nabla_\tau u^1\|^2 \geq \frac{r_2}{2(1+r_2)\epsilon_*\tau_1}\|\nabla_\tau u^1\|^2 + \frac{1}{8\varepsilon}\|\nabla_\tau u^1\|^2.$$

Thus, by inserting the above inequality and (3.17)-(3.18) into (3.16), we have

$$E^1 \leq E^0.$$

The proof is complete.

4 The DOC and DCC Kernels and Their Properties

4.1 The DCC Kernels and DOC Kernels

To obtain stability analysis of BDF2 scheme (2.8), we introduce the discrete complementary convolution (DCC) kernels $p_{n-j}^{(n)}$ such that

$$\sum_{j=1}^{n} p_{n-j}^{(n)} \mathcal{D}_2 u^j = \sum_{j=1}^{n} p_{n-j}^{(n)} \sum_{l=1}^{j} b_{j-l}^{(j)} \nabla_\tau u^l = \sum_{l=1}^{n} \nabla_\tau u^l \sum_{j=l}^{n} p_{n-j}^{(n)} b_{j-l}^{(j)} = u^n - u^0, \quad \forall n \geq 1. \tag{4.20}$$

As the identity (4.20) holds for all $n \geq 1$, it only requires

$$\sum_{j=k}^{n} p_{n-j}^{(n)} b_{j-k}^{(j)} \equiv 1, \quad \forall 1 \leq k \leq n,\ 1 \leq n \leq N. \tag{4.21}$$

The idea of DCC kernels has been successfully applied to the stability analysis for subdiffusion problems [14, 15, 17] and reaction-diffusion problem [29].

We now introduce the discrete orthogonal convolution (DOC) kernels by

$$\sum_{j=k}^{n} \theta_{n-j}^{(n)} b_{j-k}^{(j)} = \delta_{nk}, \quad \text{for all } 1 \leq k \leq n, \tag{4.22}$$

where δ_{nk} is the Kronecker delta symbol with $\delta_{nk} = 1$ if $n = k$ and $\delta_{nk} = 0$ if $n \neq k$. From (4.22), it holds

$$\sum_{j=1}^{k} \theta_{k-j}^{(k)} \mathcal{D}_2 u^j = \sum_{l=1}^{k} \nabla_\tau u^l \sum_{j=l}^{k} \theta_{k-j}^{(k)} b_{j-l}^{(j)} = u^k - u^{k-1}, \quad 1 \leq k \leq N. \tag{4.23}$$

The DCC and DOC kernels have a close connection. In fact, summing from 1 to n with k on both sides of (4.23), and then exchanging the order of summation, we have

$$\sum_{j=1}^{n} \mathcal{D}_2 u^j \sum_{k=j}^{n} \theta_{k-j}^{(k)} = u^n - u^0. \tag{4.24}$$

One can compare the identity (4.24) with (4.20) to find that (also see [29])

$$p_{n-j}^{(n)} = \sum_{k=j}^{n} \theta_{k-j}^{(k)} \quad (1 \le j \le n). \tag{4.25}$$

From (4.25), the direct calculation leads to another relation between DOC and DCC kernels

$$\theta_0^{(n)} = p_0^{(n)}, \quad \theta_{n-k}^{(n)} = p_{n-k}^{(n)} - p_{n-k-1}^{(n)} \quad (1 \le k \le n-1). \tag{4.26}$$

4.2 Some Previous Properties

To establish the stability and error estimate, here we streamline the useful results in [20, 29].

Lemma 4.1 [20, Lemma 2.2] Assume the BDF2 kernels $b_{n-k}^{(n)}$ defined in (2.6) are positive definite. Then the DOC kernels $\theta_{n-k}^{(n)}$ defined in (4.22) are also positive definite. This is, for any real sequence $\{\omega_j\}_{j=1}^{n}$, it holds that

$$\sum_{k=1}^{n} \omega_k \sum_{j=1}^{k} \theta_{k-j}^{(k)} \omega_j \ge 0, \quad \forall n \ge 1.$$

Lemma 4.2 [20, Corollary 2.1] The DOC kernels $\theta_{n-j}^{(n)}$ have the following properties:

$$\theta_{n-j}^{(n)} > 0, \quad \text{for any} \quad 1 \le j \le n, 1 \le n \le N, \tag{4.27}$$

$$\sum_{j=1}^{n} \theta_{n-j}^{(n)} = \tau_n, \qquad \text{for } n \ge 1. \tag{4.28}$$

Proposition 4.1 [29, Proposition 2.2] Let τ be the maximum time step size and r_* be any given positive constant. If the time step ratio satisfies $0 < r_k \le r_*$, then the DCC kernels $p_{n-k}^{(n)}$ defined in (4.21) satisfy

$$p_{n-j}^{(n)} = \sum_{k=j}^{n} \frac{\tau_k(1+r_j)}{1+2r_j} \prod_{i=j+1}^{k} \frac{r_i}{1+2r_i}, \quad 2 \le j \le n, \tag{4.29}$$

$$p_{n-1}^{(n)} = \sum_{k=1}^{n} \tau_k \prod_{i=2}^{k} \frac{r_i}{1+2r_i}, \tag{4.30}$$

$$\sum_{j=1}^{n} p_{n-j}^{(n)} = t_n, \tag{4.31}$$

$$p_{n-j}^{(n)} \le \sum_{k=j}^{n} \tau_k \left(\frac{r_*}{1+2r_*} \right)^{k-j} \le \sum_{k=j}^{n} \frac{\tau_k}{2^{k-j}} \le 2\tau, \tag{4.32}$$

where $\prod_{i=j+1}^{k} = 1$ for $j \geq k$ is defined.

4.3 Some New Properties of DOC Kernels

The BDF2 kernels $b_{n-k}^{(n)}$ and DOC kernels $\theta_{n-k}^{(n)}$ defined in (2.6) and (4.22) respectively can be represented as the following matrix forms [18]

$$B_2 := \begin{pmatrix} b_0^{(1)} & & & \\ b_1^{(2)} & b_0^{(2)} & & \\ & \ddots & \ddots & \\ & & b_1^{(n)} & b_0^{(n)} \end{pmatrix}_{n\times n} \quad \text{and} \quad \Theta_2 := \begin{pmatrix} \theta_0^{(1)} & & & \\ \theta_1^{(2)} & \theta_0^{(2)} & & \\ \vdots & \ddots & \ddots & \\ \theta_{n-1}^{(n)} & \cdots & \theta_1^{(n)} & \theta_0^{(n)} \end{pmatrix}_{n\times n}.$$

It follows from the definition of DOC kernels $\theta_{n-k}^{(n)}$ in (4.22) that

$$\Theta_2 = B_2^{-1}. \tag{4.33}$$

Assume **A1** holds, then Lemmas 3.2 and 4.1 imply the real symmetric matrices

$$B := B_2 + B_2^T \quad \text{and} \quad \Theta := \Theta_2 + \Theta_2^T$$

are positive definite.

Define the diagonal matrix $\Lambda_\tau := \mathrm{diag}(\sqrt{\tau_1}, \cdots, \sqrt{\tau_n})$ and

$$\tilde{B}_2 := \Lambda_\tau B_2 \Lambda_\tau = \begin{pmatrix} \tilde{b}_0^{(1)} & & & \\ \tilde{b}_1^{(2)} & \tilde{b}_0^{(2)} & & \\ & \ddots & \ddots & \\ & & \tilde{b}_1^{(n)} & \tilde{b}_0^{(n)} \end{pmatrix}_{n\times n,} \tag{4.34}$$

with

$$\tilde{b}_0^{(1)} = 1, \quad \tilde{b}_0^{(k)} = \frac{1+2r_k}{1+r_k}, \quad \tilde{b}_1^{(k)} = -\frac{r_k^{3/2}}{1+r_k}, \quad 2 \leq k \leq n.$$

Moreover, we define the real symmetric matrix $\tilde{B} := \tilde{B}_2 + \tilde{B}_2^T$, which has the following properties.

Lemma 4.3 Assume **A1** holds, the minimum eigenvalue of $\tilde{B}$ can be bounded by

$$\lambda_{\min}(\tilde{B}) \geq \min_{1\leq k\leq n} \tilde{R}(r_k, r_{k+1}) \geq C_\delta,$$

where

$$\tilde{R}(x,y) = \frac{2+4x-x^{3/2}}{1+x} - \frac{y^{3/2}}{1+y}, \quad 0 < x, y \leq r_{\max}, \tag{4.35}$$

$$C_\delta = \min\{\tilde{R}(0, r_{\max}-\delta), \tilde{R}(r_{\max}-\delta, r_{\max}-\delta)\}. \tag{4.36}$$

Thus $\tilde{B}$ is positive definite and there exists a non-singular upper triangular matrix L such that

$$\tilde{B} = \Lambda_\tau B \Lambda_\tau = L^T L \qquad \text{or} \qquad B = (L\Lambda_\tau^{-1})^T L \Lambda_\tau^{-1}.$$

For brevity, we leave the detailed proof in the Appendix.

Remark 1 It follows from the **A1** and the monotony of $\tilde{R}(x,y)$ for $x>1, y>0$ that

$$\begin{aligned}
&\tilde{R}(0, r_{\max}-\delta) \geq \tilde{R}(0, r_{\max}) = 2 - \frac{r_{\max}^{3/2}}{1+r_{\max}} > 0\\
&\tilde{R}(r_{\max}-\delta, r_{\max}-\delta) \geq \tilde{R}(r_{\max}, r_{\max}-\delta) = \frac{r_{\max}^{3/2}}{1+r_{\max}} - \frac{(r_{\max}-\delta)^{3/2}}{1+r_{\max}-\delta}\\
&\qquad = \frac{\sqrt{r_{\max}}(r_{\max}+r_{\max}^2-\delta r_{\max}) - \sqrt{r_{\max}-\delta}(r_{\max}+r_{\max}^2-\delta-\delta r_{\max})}{(1+r_{\max})(1+r_{\max}-\delta)}\\
&\qquad \geq \frac{\sqrt{r_{\max}}}{(1+r_{\max})^2}\delta, \quad \text{for} \quad 0<\delta<r_{\max}-1,
\end{aligned}$$

which implies the positive constant C_δ depends on δ. Moreover, if $0<\delta\leq r_{\max}-4$, one has

$$\tilde{R}(0, r_{\max}-\delta) \geq \tilde{R}(r_{\max}-\delta, r_{\max}-\delta).$$

Then, for any $0<\delta\leq r_{\max}-4$, the constant C_δ can be estimated by

$$C_\delta = \tilde{R}(r_{\max}-\delta, r_{\max}-\delta) \geq \frac{\sqrt{r_{\max}}}{(1+r_{\max})^2}\delta.$$

Thus, the lower bound of C_δ is $\mathcal{O}(\delta)$ for small δ.

The next Lemma gives an upper bound of the maximum singular value of $\tilde{B}_2$.

Lemma 4.4 If **A1** holds, then the maximum eigenvalue of the real symmetric matrix $\tilde{B}_2^T\tilde{B}_2$ can be bounded by

$$\lambda_{\max}(\tilde{B}_2^T\tilde{B}_2) \leq \hat{R}(r_{\max}-\delta, r_{\max}-\delta) \leq \frac{4r_{\max}^3}{(1+r_{\max})^2} < 14,$$

where $\tilde{B}_2$ is defined by (4.34) and $\hat{R}(x,y)$ is defined by

$$\hat{R}(x,y) := \frac{(1+2x)(1+2x+x^{3/2})}{(1+x)^2} + \frac{y^{3/2}(1+2y+y^{3/2})}{(1+y)^2}.$$

Again for brevity, we leave the detailed proof in the Appendix.

To deal with the nonlinear term $\nabla_h\cdot\boldsymbol{f}(\nabla_h u^n)$, we now introduce the following matrices

$$\widehat{\boldsymbol{B}} := B\otimes I_2, \quad \widehat{\boldsymbol{\Theta}} := \Theta\otimes I_2, \quad \widetilde{\boldsymbol{B_2}} := \tilde{B}_2\otimes I_2, \quad \widetilde{\boldsymbol{B}} := \tilde{B}\otimes I_2, \quad \widehat{\boldsymbol{\Lambda_\tau}} = \Lambda_\tau\otimes I_2, \quad \widehat{\boldsymbol{L}} := L\otimes I_2.$$

One can use Lemma 4.3 to derive that

$$\|\widehat{\boldsymbol{L}}^{-1}\|_2^2 = \lambda_{\max}((\widehat{\boldsymbol{L}}^T\widehat{\boldsymbol{L}})^{-1}) = \lambda_{\min}^{-1}((L^TL)\otimes I_2) = \lambda_{\min}^{-1}(\widetilde{\boldsymbol{B}}).$$

For convenience, we define the vector norm $\|\cdot\|_2$ by $\|\boldsymbol{u}\|_2 = \sqrt{\boldsymbol{u}^T\boldsymbol{u}}$ and the associated matrix norm by $\|A\|_2 := \sqrt{\lambda_{\max}(A^TA)}$. We also define

$$\mathcal{Q}_\delta := \max_{n\geq 1}\|\widetilde{\boldsymbol{B_2}}\|_2^2\|\widehat{\boldsymbol{L}}^{-1}\|_2^4 = \max_{n\geq 1}\frac{\lambda_{\max}(\widetilde{\boldsymbol{B_2}}^T\widetilde{\boldsymbol{B_2}})}{\lambda_{\min}^2(\widetilde{\boldsymbol{B}})}. \tag{4.37}$$

Note that under the codition **A1**, Lemmas 4.3 and 4.4 imply the positive constant $\mathcal{Q}_\delta < \frac{14}{C_\delta}$, where C_δ is defined by (4.36). By taking $\delta = 1.303$, one can obtain that $\mathcal{Q}_\delta < 39$, which is consistent with the one in [19].

We now present several important Lemmas, which plays a key role in dealing with the nonlinear term $\nabla_h \cdot \boldsymbol{f}(\nabla_h u^n)$, and leave the proofs in the Appendix for brevity due to their similarity with [19].

Lemma 4.5 Assume **A1** holds. For the positive definite matrix $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{B}}_2^{-1}\widehat{\boldsymbol{B}}(\widehat{\boldsymbol{B}}_2^{-1})^T$, and any vector sequences $\{\boldsymbol{v}^k\}_{k=1}^n, \{\boldsymbol{w}^k\}_{k=1}^n, \boldsymbol{v}^k, \boldsymbol{w}^k \in \mathbb{R}^2$, it holds

$$\sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}(\boldsymbol{v}^k)^T\boldsymbol{w}^j \le \frac{\epsilon}{2}\boldsymbol{v}^T\widehat{\boldsymbol{\Theta}}\boldsymbol{v} + \frac{1}{2\epsilon}\boldsymbol{w}^T\widehat{\boldsymbol{B}}^{-1}\boldsymbol{w}, \quad \forall \epsilon > 0, \tag{4.38}$$

where $\boldsymbol{v} = ((\boldsymbol{v}^1)^T, \cdots, (\boldsymbol{v}^n)^T)^T$ and $\boldsymbol{w} = ((\boldsymbol{w}^1)^T, \cdots, (\boldsymbol{w}^n)^T)^T$.

Lemma 4.6 *(12, Lemma 3.5])* For any $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^2$, there exists a symmetric matrix $Q_{\boldsymbol{f}} \in \mathbb{R}^{2\times 2}$ such that

$$\boldsymbol{f}(\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{w}) = Q_{\boldsymbol{f}}(\boldsymbol{v} - \boldsymbol{w}).$$

The eigenvalues λ_1, λ_2 of $Q_{\boldsymbol{f}}$ satisfy $-\frac{1}{8} \le \lambda_1, \lambda_2 \le 1$. Consequently, it holds that

$$|\boldsymbol{f}(\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{w})| \le |\boldsymbol{v} - \boldsymbol{w}|. \tag{4.39}$$

We now give another important lemma as follows.

Lemma 4.7 Assume **A1** holds, then for any vector sequence $\boldsymbol{v}^k = (v_1^k, v_2^k)^T, \boldsymbol{w}^k = (w_1^k, w_2^k)^T, \boldsymbol{z}^k = (z_1^k, z_2^k)^T$ with $1 \le k \le n$ and any $\epsilon > 0$, it holds that

$$\sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}(\boldsymbol{z}^k)^T[\boldsymbol{f}(\boldsymbol{v}^j + \boldsymbol{w}^j) - \boldsymbol{f}(\boldsymbol{v}^j)] \le \sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}\left[\epsilon(\boldsymbol{z}^k)^T\boldsymbol{z}^j + \frac{\mathcal{Q}_\delta}{\epsilon}(\boldsymbol{w}^k)^T\boldsymbol{w}^j\right],$$

where the positive constant $\mathcal{Q}_\delta$ is defined in (4.37). Moreover,

$$\sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}(\boldsymbol{z}^k)^T[\boldsymbol{f}(\boldsymbol{v}^j + \boldsymbol{z}^j) - \boldsymbol{f}(\boldsymbol{v}^j)] \le 2\sqrt{\mathcal{Q}_\delta}\sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}(\boldsymbol{z}^j)^T\boldsymbol{z}^k.$$

5 The Stability and Convergence Analysis for BDF2 Scheme

In this section, we consider the stability and convergence analysis for the BDF2 scheme (2.8), which requires a discrete Grönwall inequality given as follows.

Lemma 5.1 Assume that $\lambda > 0$ 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 that

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

5.1 L^2-norm Stability

We now consider the L^2-norm stability analysis of BDF2 scheme (2.8) with variable time steps. It is noted that the inequality (3.10) plays an important role in [21] to obtain the L^2-norm error estimate on uniform time steps. But the method developed in [21] fails to prove the L^2-norm estimate if the DOC technique is used in this paper. In other words, the inequality (3.10) can not be used to have L^2-norm estimate since the DOC technique will lead to the cross inner product $\langle \phi^k, \phi^j\rangle$ for different time levels. Alternatively, we here develop a new inequality to fill this gap as follows.

Lemma 5.2 Let the matrix $\boldsymbol{\beta} := \boldsymbol{\beta}_2 + \boldsymbol{\beta}_2^T$ be positive definite with $\boldsymbol{\beta}_2 := (\beta_{k,j}), 1 \le j, k \le n, \beta_{k,j} = 0$ if $j > k$. Assume $\beta_{k,j} \ge 0$ for any $1 \le k, j \le n$, then it holds

$$\sum_{k=1}^{n}\sum_{j=1}^{k}\beta_{k,j}\langle \nabla_h\phi_h^j, \nabla_h\phi_h^k\rangle \le \sum_{k=1}^{n}\sum_{j=1}^{k}\beta_{k,j}\langle -\Delta_h\phi_h^j, \phi_h^k\rangle. \tag{5.40}$$

Proof For notation convenience, we denote the operator $\mathcal{L}_h := \nabla_h\cdot\nabla_h - \Delta_h$. It follows from the positive definiteness of the matrix $\boldsymbol{\beta}$ and the standard Cholesky decomposition that there exists a non-singular upper triangular matrix A such that $\boldsymbol{\beta} = A^TA$. Denote the matrix $A = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n)$, where $\boldsymbol{a}_l = (a_{1,l}, \cdots, a_{n,l})^T, 1 \le l \le n$ and $a_{l,k} = 0$ if $l > k$. It is easy to verify that $\beta_{k,j} = \boldsymbol{a}_k^T\boldsymbol{a}_j = \sum_{l=1}^{n} a_{l,k}a_{l,j}, k \ne j$, $\beta_{k,k} = \boldsymbol{a}_k^T\boldsymbol{a}_k/2 = \sum_{l=1}^{n} a_{l,k}^2/2, 1 \le k \le n$. The discrete Green's formula and the identity $\langle \mathcal{L}_hv, w\rangle = \langle v, \mathcal{L}_hw\rangle$ yield

$$\begin{aligned}
\frac{1}{2}\sum_{l=1}^{n}\langle \mathcal{L}_h(\sum_{k=1}^{n}a_{l,k}\phi_h^k), \sum_{k=1}^{n}a_{l,k}\phi_h^k\rangle &= \frac{1}{2}\sum_{k=1}^{n}(\sum_{l=1}^{n}a_{l,k}^2)\langle \mathcal{L}_h\phi_h^k, \phi_h^k\rangle + \sum_{k=2}^{n}\sum_{j=1}^{k-1}(\sum_{l=1}^{n}a_{l,k}a_{l,j})\langle \mathcal{L}_h\phi_h^k, \phi_h^j\rangle \\
&= \sum_{k=1}^{n}\beta_{k,k}\langle \mathcal{L}_h\phi_h^k, \phi_h^k\rangle + \sum_{k=2}^{n}\sum_{j=1}^{k-1}\beta_{k,j}\langle \mathcal{L}_h\phi_h^k, \phi_h^j\rangle \\
&= \sum_{k=1}^{n}\sum_{j=1}^{k}\beta_{k,j}\langle \mathcal{L}_h\phi_h^k, \phi_h^j\rangle.
\end{aligned}$$

It follows from the discrete Green's formula and the inequality (3.10) that $\langle \mathcal{L}_hv, v\rangle \ge 0$ for any $v \in \mathbb{V}_h$, which implies

$$\sum_{k=1}^{n}\sum_{j=1}^{k}\beta_{k,j}\langle \mathcal{L}_h\phi_h^k, \phi_h^j\rangle \ge 0.$$

The proof is complete.

We now consider the L^2-norm stability of BDF2 with variable time steps.

Theorem 5.1 Assume **A1** holds and the maximum time step $\tau \leq 16\mathcal{Q}_\delta^2/\varepsilon$. Then the BDF2 scheme (2.8) is stable in L^2-norm, and has the estimate

$$\|\hat{u}_h^n - u_h^n\| \leq 2\exp(16\mathcal{Q}_\delta t_{n-1}/\varepsilon)\|\hat{u}_h^0 - u_h^0\|, \tag{5.41}$$

where u_h^n and $\hat{u}_h^n$ are the solutions to (2.8) with the initial values u_h^0 and $\hat{u}_h^0$, respectively.

Proof Let $\phi_h^n := \hat{u}_h^n - u_h^n$ $(0 \leq n \leq N)$ be the solution perturbation for $\boldsymbol{x}_h \in \bar{\Omega}_h$. The BDF2 scheme (2.8) implies that ϕ_h^n solves

$$\mathcal{D}_2\phi_h^j + \varepsilon\Delta_h^2\phi_h^j + \nabla_h \cdot (\boldsymbol{f}(\nabla_h\hat{u}_h^j) - \boldsymbol{f}(\nabla_h u_h^j)) = 0.$$

Multiplying both sides of the above identity by $\theta_{k-j}^{(k)}$ and summing j from 1 to k, one can use the identity (4.23) to obtain

$$\nabla_\tau\phi_h^k + \varepsilon\sum_{j=1}^k \theta_{k-j}^{(k)}\Delta_h^2\phi_h^j + \sum_{j=1}^k \theta_{k-j}^{(k)}\nabla_h \cdot (\boldsymbol{f}(\nabla_h\hat{u}_h^j) - \boldsymbol{f}(\nabla_h u_h^j)) = 0.$$

Taking the inner product of the above identity with $2\phi^k$ and summing k from 1 to n, one has

$$\begin{aligned}\|\phi_h^n\|^2 - \|\phi_h^0\|^2 + 2\varepsilon\sum_{k=1}^n\sum_{j=1}^k \theta_{k-j}^{(k)}\langle\Delta_h\phi_h^j, \Delta_h\phi_h^k\rangle \\ \leq \quad 2\sum_{k=1}^n\sum_{j=1}^k \theta_{k-j}^{(k)}\langle\boldsymbol{f}(\nabla_h\hat{u}_h^j) - \boldsymbol{f}(\nabla_h u_h^j), \nabla_h\phi_h^k\rangle,\end{aligned} \tag{5.42}$$

where the discrete Green's formula (3.9) and $2a^2 - 2ab \geq a^2 - b^2$ are used. We now deal with the rightmost term of (5.42). In view of $\nabla_h\hat{u}_h^j = \nabla_h u_h^j + \nabla_h\phi_h^j$, one can apply Lemmas 5.2 and 4.7 to obtain

$$\begin{aligned}&2\sum_{k=1}^n\sum_{j=1}^k\theta_{k-j}^{(k)}\langle(\boldsymbol{f}(\nabla_h\hat{u}_h^j) - \boldsymbol{f}(\nabla_h u_h^j)), \nabla_h\phi_h^k\rangle \\ &\qquad \leq 4\sqrt{\mathcal{Q}_\delta}\sum_{k=1}^n\sum_{j=1}^k \theta_{k-j}^{(k)}\langle\nabla_h\phi_h^j, \nabla_h\phi_h^k\rangle \\ &\qquad \leq 4\sqrt{\mathcal{Q}_\delta}\sum_{k=1}^n\sum_{j=1}^k \theta_{k-j}^{(k)}\langle-\Delta_h\phi_h^j, \phi_h^k\rangle \qquad (5.43) \\ &\qquad \leq 2\varepsilon\sum_{k=1}^n\sum_{j=1}^k \theta_{k-j}^{(k)}\langle\Delta_h\phi_h^j, \Delta_h\phi_h^k\rangle + \frac{8\mathcal{Q}_\delta^2}{\varepsilon}\sum_{k=1}^n\sum_{j=1}^k \theta_{k-j}^{(k)}\langle\phi_h^j, \phi_h^k\rangle. \qquad (5.44)\end{aligned}$$

The last inequality holds by taking $\epsilon = \varepsilon/(2\sqrt{\mathcal{Q}_\delta})$ in Lemma 4.7 since Lemma 4.7 still holds for the linear function $\boldsymbol{f}(\boldsymbol{v}) := \boldsymbol{v}$. Inserting above inequality to (5.42), one can use the discrete Cauchy-Schwartz inequality to obtain

$$\|\phi_h^n\|^2 - \|\phi_h^0\|^2 \leq \frac{8\mathcal{Q}_\delta^2}{\varepsilon}\sum_{k=1}^n\sum_{j=1}^k \theta_{k-j}^{(k)}\langle\phi_h^j, \phi_h^k\rangle \leq \frac{8\mathcal{Q}_\delta^2}{\varepsilon}\sum_{k=1}^n\sum_{j=1}^k \theta_{k-j}^{(k)}\|\phi_h^j\|\phi_h^k\|.$$

Select a n_0 such that $\|\phi_h^{n_0}\| = \max_{0\le k\le n}\|\phi_h^k\|$. Then for the time level n_0, we have

$$\|\phi_h^{n_0}\|^2 \le \|\phi_h^0\|\|\phi_h^{n_0}\| + \frac{8\mathcal{Q}_\delta^2}{\varepsilon}\|\phi_h^{n_0}\|\sum_{k=1}^{n_0}\|\phi_h^k\|\sum_{j=1}^{k}\theta_{k-j}^{(k)},$$

which together with Lemma 4.2 imply that

$$\|\phi_h^n\| \le \|\phi_h^{n_0}\| \le \|\phi_h^0\| + \frac{8\mathcal{Q}_\delta^2}{\varepsilon}\|\sum_{k=1}^{n_0}\tau_k\|\phi_h^k\| \le \|\phi_h^0\| + \frac{8\mathcal{Q}_\delta^2}{\varepsilon}\|\sum_{k=1}^{n}\tau_k\|\phi_h^k\|.$$

With the help of the maximum time step $\tau \le 16\mathcal{Q}_\delta^2/\varepsilon$, we arrive at

$$\|\phi_h^n\|\| \le 2\|\phi_h^0\| + \frac{16\mathcal{Q}_\delta^2}{\varepsilon}\sum_{k=1}^{n}\tau_k\|\phi_h^k\|.$$

Thus, the proof is completed by using the Grönwall inequality in Lemma 5.1.

5.2 Consistency and Convergence

Let $e_h^n := u(t_n, \boldsymbol{x}_h) - u_h^n$ $(n \ge 0)$ be the error between the numerical solution u_h^n and exact solution $u(\boldsymbol{x}_h, t_n)$ for $\boldsymbol{x}_h \in \bar{\Omega}_h$. Then e_h^n solves the error function

$$\mathcal{D}_2 e_h^j + \varepsilon\Delta_h^2 e_h^j + \nabla_h\cdot(\boldsymbol{f}(\nabla_h(e_h^j + u_h^j)) - \boldsymbol{f}(\nabla_h u_h^j)) = \xi_h^j + \eta_h^j, \quad \text{for} \quad 1 \le j \le N, \tag{5.45}$$

where $\eta_h^n := \mathcal{D}_2 u(\boldsymbol{x}_h, t_j) - u_t(\boldsymbol{x}_h, t_j)$ $(1 \le j \le N)$ denotes the temporal truncation error and $\xi_h^j := C_u h^2$ denotes the spatial truncation error. Here C_u is a constant depending on the exact solution u, but independent of the mesh sizes h and τ_j.

The following Lemma implies that the temporal truncation error η_h^j can be divided into a convolution part and a rest part, which plays a key role in simplifying the complicate calculation for the estimate of convergence order when using the techniques of DOC and DCC kernels.

Lemma 5.3 *[29, Lemma 3.2])* Denote the temporal truncation error by $\eta_h^j := \mathcal{D}_2 u(t_j) - u_t(t_j)$ $(1 \le j \le N)$ and set

$$\begin{aligned} G^l &:= -\frac{1}{2}\int_{t_{l-1}}^{t_l}(t - t_{l-1})^2 u_{ttt}\,\mathrm{d}t, \quad 1 \le l \le N,\\ R^j &:= -\frac{1}{2}b_1^{(j)}\tau_{j-1}\int_{t_{j-1}}^{t_j}(2(t - t_{j-1}) + \tau_{j-1})u_{ttt}\,\mathrm{d}t, \quad 2 \le j \le n,\\ R^1 &:= \frac{1}{2\tau_1}\int_0^{t_1} t^2 u_{ttt}\,\mathrm{d}t - \frac{1}{\tau_1}\int_0^{t_1} t u_{tt}\,\mathrm{d}t. \end{aligned} \tag{5.46}$$

Then it holds that

$$\eta_h^j = \sum_{l=1}^{j} b_{j-l}^{(j)} G^l + R^j, \quad 1 \le j \le n. \tag{5.47}$$

Theorem 5.2 Let $u(t,\boldsymbol{x})$ and u_h^n be the solutions to problem (1.1) and discrete problem (2.8), respectively. Assume **A1** holds and the maximum time step $\tau \le \varepsilon/16\mathcal{Q}_\delta$, then the discrete solution u_h^n of BDF2 scheme (2.8) is convergent in L^2-norm in the sense that

$$\begin{aligned}\|e_h^n\| \le & 2\exp(16\mathcal{Q}_\delta^2 t_{n-1}/\varepsilon)(\|e_h^0\| + 2C_u t_n h^2 + \sum_{k=1}^{n}\tau_k^2\int_{t_{k-1}}^{t_k}\|u_{ttt}\|\,\mathrm{d}t \\ & + 2t_n \max_{1\le k\le n}\tau_k\int_{t_{k-1}}^{t_k}\|u_{ttt}\|\,\mathrm{d}t + 4\tau\int_0^{t_1}\|u_{tt}\|\,\mathrm{d}t).\end{aligned}$$

Proof Multiplying both sides of (5.45) by $\theta_{k-j}^{(k)}$, and summing j from 1 to k, then taking the inner product with e_h^k on both sides and summing k from 1 to n, one has

$$\begin{aligned}&\|e_h^n\|^2 - \|e_h^0\|^2 + 2\varepsilon\sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}\langle\Delta_h e_h^j, \Delta_h e_h^k\rangle \\ &\quad\le 2\sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}\langle \boldsymbol{f}(\nabla_h(e_h^j+u_h^j)) - \boldsymbol{f}(\nabla_h u_h^j), \nabla_h e_h^k\rangle + 2\sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}\langle\xi_h^j+\eta_h^j, e_h^k\rangle \\ &\quad := I_1 + I_2 + I_3,\end{aligned} \tag{5.48}$$

where the discrete Green's formula (3.9) and $2a^2 - 2ab \ge a^2 - b^2$ are used. The first term I_1 has been estimated by (5.44). One can use Lemma 4.2 and the discrete Cauchy–Schwartz inequality to obtain

$$I_2 \le 2t_n \max_{1\le k\le n}\|\xi_h^k\|\|e_h^k\| \le 2C_u t_n h^2 \max_{1\le k\le n}\|e_h^k\|. \tag{5.49}$$

We now consider I_3. By using Lemma 5.3 and exchanging the order of summation, we have

$$\begin{aligned}I_3 &= 2\sum_{k=1}^{n}\langle\sum_{j=1}^{k}\theta_{k-j}^{(k)}\sum_{l=1}^{j}b_{j-l}^{(j)}G^l, e_h^k\rangle + 2\sum_{k=1}^{n}\langle\sum_{j=1}^{k}\theta_{k-j}^{(k)}R^j, e_h^k\rangle \\ &= 2\sum_{k=1}^{n}\langle\sum_{l=1}^{k}G^l\sum_{j=l}^{k}\theta_{k-j}^{(k)}b_{j-l}^{(j)}, e_h^k\rangle + 2\sum_{k=1}^{n}\langle\sum_{j=1}^{k}\theta_{k-j}^{(k)}R^j, e_h^k\rangle \\ &= 2\sum_{k=1}^{n}\langle G^k, e_h^k\rangle + 2\sum_{k=1}^{n}\langle\sum_{j=1}^{k}\theta_{k-j}^{(k)}R^j, e_h^k\rangle, \\ &\le 2\sum_{k=1}^{n}\|G^k\|\|e_h^k\| + 2\sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}\|R^j\|\|e_h^k\|.\end{aligned}$$

Inserting (5.44), (5.49) and above inequality into (5.48), one has

$$\begin{aligned}\|e_h^n\|^2 \le & \|e_h^0\|^2 + \frac{8\mathcal{Q}_\delta^2}{\varepsilon}\sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}\|e_h^j\|\|e_h^k\| \\ & + 2C_u t_n h^2\max_{1\le k\le n}\|e_h^k\| + 2\sum_{k=1}^{n}\|G^k\|\|e_h^k\| + 2\sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}\|R^j\|\|e_h^k\|.\end{aligned}$$

Choose n_0 such that $\|e_h^{n_0}\| = \max_{1\le k\le n}\|e_h^n\|$. Then, we have

$$\begin{aligned}\|e_h^{n_0}\|^2 \le&\|e_h^0\|\|e_h^{n_0}\| + \frac{8\mathcal{Q}_\delta^2}{\varepsilon}\|e_h^{n_0}\|\sum_{k=1}^{n_0}\sum_{j=1}^{k}\theta_{k-j}^{(k)}\|e_h^k\|\\ &+ 2C_u t_n h^2\|e_h^{n_0}\| + 2\|e_h^{n_0}\|\sum_{k=1}^{n_0}\|G^k\| + 2\|e_h^{n_0}\|\sum_{k=1}^{n_0}\sum_{j=1}^{k}\theta_{k-j}^{(k)}\|R^j\|.\end{aligned}$$

Thus, by exchanging the order of summation, we arrive at

$$\begin{aligned}\|e_h^n\| \le \|e_h^{n_0}\| \le \|e_h^0\| + \frac{8\mathcal{Q}_\delta^2}{\varepsilon}\sum_{k=1}^{n}\tau_k\|e_h^k\|\\ + 2C_u t_n h^2 + 2\sum_{k=1}^{n}\|G^k\| + 2\sum_{j=1}^{n}p_{n-j}^{(k)}\|R^j\|.\end{aligned}$$

One may use Lemma 4.2 and the assumption of the maximum time step $\tau \le \varepsilon/16\mathcal{Q}_\delta$ to obtain

$$\begin{aligned}\|e_h^n\| \le 2\|e_h^0\| + \frac{16\mathcal{Q}_\delta^2}{\varepsilon}\sum_{k=1}^{n-1}\tau_k\|e_h^k\|\\ + 4C_u t_n h^2 + 4\sum_{k=1}^{n}\|G^k\| + 4\sum_{j=1}^{n}p_{n-j}^{(k)}\|R^j\|,\end{aligned}$$

which together with Lemma 4.2 and the discrete Grönwall inequality in Lemma 5.1 derives

$$\|e_h^n\| \le 2\exp(16\mathcal{Q}_\delta^2 t_{n-1}/\varepsilon)(\|e_h^0\| + 2C_u t_n h^2 + 2\sum_{k=1}^{n}\|G^k\| + 2\sum_{k=1}^{n}p_{n-k}^{(n)}\|R^k\|). \tag{5.50}$$

Note that G^l, R^j defined in Lemma 5.3 satisfy

$$\begin{aligned}\|G^l\| &\le \frac{\tau_l^2}{2}\int_{t_{l-1}}^{t_l}\|u_{ttt}\|\,\mathrm{d}t,\\ \|R^j\| &\le \frac{r_j}{2+2r_j}(2\tau_j+\tau_{j-1})\int_{t_{j-1}}^{t_j}\|u_{ttt}\|\,\mathrm{d}t,\\ &\le \tau_j\int_{t_{j-1}}^{t_j}\|u_{ttt}\|\,\mathrm{d}t, \quad 2\le j\le n.\end{aligned}$$

Then it follows from Proposition 4.1 that

$$\begin{aligned}\sum_{k=1}^{n}p_{n-k}^{(n)}\|R^k\| &= \sum_{k=2}^{n}p_{n-k}^{(n)}\|R^k\| + p_{n-1}^{(n)}\|\xi_1\|\\ &\le \sum_{k=1}^{n}p_{n-k}^{(n)}\tau_j\int_{t_{j-1}}^{t_j}\|u_{ttt}\|\,\mathrm{d}t + p_{n-1}^{(n)}\int_0^{t_1}\|u_{tt}\|\,\mathrm{d}t\\ &\le t_n\max_{1\le k\le n}\tau_k\int_{t_{k-1}}^{t_k}\|u_{ttt}\|\,\mathrm{d}t + 2\tau\int_0^{t_1}\|u_{tt}\|\,\mathrm{d}t.\end{aligned}$$

Thus, we derive that

$$\begin{aligned}\|e_h^n\| \leq & 2\exp(16\mathcal{Q}_\delta^2 t_{n-1}/\varepsilon)(\|e_h^0\| + 2C_u t_n h^2 + \sum_{k=1}^{n} \tau_k^2 \int_{t_{k-1}}^{t_k} \|u_{ttt}\| \,\mathrm{d}t \\ & + 2t_n \max_{1\leq k\leq n} \tau_k \int_{t_{k-1}}^{t_k} \|u_{ttt}\| \,\mathrm{d}t + 4\tau \int_0^{t_1} \|u_{tt}\| \,\mathrm{d}t). \end{aligned} \tag{5.51}$$

The proof is complete.

Remark 2 Recently, [19] gives a result for the error estimate of the molecular beam epitaxial model (1.1) in form of

$$\|e_h^n\| \leq C_u \exp(16\mathcal{M}_\tau^2 t_{n-1}/\varepsilon)(\|e_h^0\| + t_n\tau_1 + t_n(\tau^2 + h^2)). \tag{5.52}$$

with $0 \leq r_k \leq 3.561$. One can see that the right-hand-side second term is the first-order convergence for large t_n. To obtain the second-order convergence in [19], it requires another restriction condition $|\mathfrak{R}_p| \leq N_0 \ll N$, where the index set $|\mathfrak{R}_p|$ defined by (1.3). Our result in Theorem 5.2 shows the robust second-order (optimal) convergence remains valid to a new ratio condition $0 < r_k \leq r_{\max} \approx 4.8645 - \delta$, where $\delta > 0$ is a arbitrarily small constant. The robustness means that the convergence does not need other conditions on the time step like the constrained condition $|\mathfrak{R}_p| \leq N_0 \ll N$.

6 Numerical Experiments

We now present numerical example to investigate the convergence order of BDF2 with variable time steps, which is tested on random time meshes. To do so, we set the computational domain $\Omega = (0, S)^2$ with S a positive constant, the final time $T = 1$, and consider the following exterior-forced MBE model

$$u_t = -\varepsilon\Delta^2 u - \nabla \cdot \boldsymbol{f}(\nabla u) + g(\boldsymbol{x}, t). \tag{6.53}$$

By choosing a suitable function g, the exact solution to (6.53) is constructed as follows

$$u = \cos(t)\sin(\frac{2\pi x}{S})\sin(\frac{2\pi y}{S}).$$

The random time meshes are given by $\tau_k = T\chi_k/C$, where $C = \sum_{k=1}^{N} \chi_k$ with χ_k randomly drawn from the uniform distribution on $(0, 1)$ and the number of spatial meshes is chosen by $M = N$. In each run, the error $e(N) = \|u(T) - u^N\|$ and the numerical rate of convergence at the final time $T = 1$ are recorded in Tables 1 and 2, in which the maximum time step τ and maximum adjacent time step ratio are also listed, where the convergence rate is calculated by

$$\text{Order} = \log_2(e(N)/e(2N)).$$

Table 1 Numerical accuracy on random time mesh for $S = 2\pi, \varepsilon = 1$

N	$e(N)$	Order	τ	max r_k
64	1.9293e-03	–	0.0280807	48.0321
128	4.8909e-04	1.980	0.0152735	98.0471
256	1.2007e-04	2.026	0.00745376	1584.01
512	3.0633e-05	1.971	0.00383547	430.559

Table 2 Numerical accuracy on random time mesh for $S = 2, \varepsilon = 4$

N	$e(N)$	Order	τ	max r_k
64	8.6542e-04	–	0.0300289	191.022
128	2.1622e-04	2.001	0.0147087	418.406
256	5.4044e-05	2.000	0.0078714	395.573
512	1.3506e-05	2.001	0.00374001	604.021

As shown in Tables 1 and 2, even though the time step is randomly chosen beyond the constrained condition **A1**, the BDF2 scheme with variable time steps is robustly stable and convergent in the second order. In addition for the simulations, the first-step BDF1 does not bring the loss of accuracy and is consistent with our theoretical analysis.

7 Conclusion

The BDF2 scheme with variable time steps is considered to solve the MBE model without slope selection. Our proposed BDF2 scheme with BDF1 for the first step is proved to preserve the discrete energy dissipation law under a new adjacent time-step ratio condition: $r_k := \tau_k/\tau_{k-1} \leq 4.8645 - \delta$, where $\delta > 0$ is a arbitrarily small constant. By using the techniques of DOC and DCC kernels, we achieve the robust and sharp second-order error estimate under the condition **A1**.

Our second-order convergence analysis shows two folds: (i) the variable time steps under condition **A1** will not influence on the convergence order; (ii) the BDF1 scheme for the first step is consistent to the globally optimal convergence order. This conclusion removes the doubt of the classical choice of the first level solution with first-order consistent BDF1 scheme for the sharp second-order convergence. Numerical results demonstrate the theoretical analysis.

Acknowledgements

The numerical simulations in this work have been done on the supercomputing system in the Supercomputing Center of Wuhan University. The authors would like to thank Professor Tao Tang for his valuable suggestions.

Appendix

The Proof of Lemma 4.3:

By using the inequality $2ab \le a^2 + b^2$, one has

$$\begin{aligned} 2w_k \sum_{j=1}^{k} \tilde{b}_{k-j}^{(k)} w_j &= 2\tilde{b}_0^{(k)} w_k^2 + 2\tilde{b}_1^{(k)} w_k w_{k-1} \ge (2\tilde{b}_0^{(k)} + \tilde{b}_1^{(k)}) w_k^2 + \tilde{b}_1^{(k)} w_{k-1}^2 \\ &= \frac{2 + 4r_k - r_k^{3/2}}{(1+r_k)} w_k^2 - \frac{r_k^{3/2}}{(1+r_k)} w_{k-1}^2 \\ &= \frac{2 + 4r_k - r_k^{3/2}}{(1+r_k)} w_k^2 - \frac{r_k^{3/2}}{(1+r_k)} w_{k-1}^2 \\ &= \frac{r_{k+1}^{3/2}}{(1+r_{k+1})} w_k^2 - \frac{r_k^{3/2}}{(1+r_k)} w_{k-1}^2 + \tilde{R}(r_k, r_{k+1}) w_k^2, \quad k \ge 2. \end{aligned}$$

Note that $\partial_x \tilde{R}(x,y) = \frac{1}{2}(1+x)^{-2}(1-\sqrt{x})(x+\sqrt{x}+4)$. Hence, $\tilde{R}(x,y)$ is increasing in $(0,1)$ and decreasing in $(1, r_{\max})$ with respect to x. And it is easy to verify that $\tilde{R}(x,y)$ is decreasing in $(0, r_{\max})$ with respect to y. Thus, $\tilde{R}(x,y)$ attains its minimum at $(x,y) = (0, r_{\max} - \delta)$ or $(x,y) = (r_{\max}, r_{\max} - \delta)$, namely,

$$\min_{0<x,y\le r_{\max}-\delta} \tilde{R}(x,y) = \min\{\tilde{R}(0, r_{\max}-\delta), \tilde{R}(r_{\max}-\delta, r_{\max}-\delta)\} = C_\delta > 0.$$

Due to the symmetry of matrix $\tilde{B}$, for any $\boldsymbol{v} \in \mathbb{R}^n$, we have

$$\lambda_{\min}(\tilde{B}) = \min_{\boldsymbol{v}\in\mathbb{R}^n} \frac{\boldsymbol{v}^T \tilde{B} \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}} = \min_{\boldsymbol{v}\in\mathbb{R}^n} \frac{2}{\boldsymbol{v}^T \boldsymbol{v}} \sum_{k=1}^{n} w_k \sum_{j=1}^{k} \tilde{b}_{k-j}^{(k)} w_j \ge \min_{1\le k\le n} \tilde{R}(r_k, r_{k+1}) \ge C_\delta,$$

which implies that the real symmetric matrix $\tilde{B}$ is positive definite. The last claim holds by applying the standard Cholesky decomposition to $\tilde{B}$. The proof is complete.

The proof of Lemma 4.4:

The direct calculation from the definition of $\tilde{B}_2$ (4.34) produces that

$$\tilde{B}_2^T \tilde{B}_2 = \begin{pmatrix} d_0^{(1)} & d_1^{(2)} & & & \\ d_1^{(2)} & d_0^{(2)} & d_1^{(3)} & & \\ & \ddots & \ddots & \ddots & \\ & & d_1^{(n-1)} & d_0^{(n-1)} & d_1^{(n)} \\ & & & d_1^{(n)} & d_0^{(n)} \end{pmatrix}_{n\times n},$$

where the elements $d_0^{(k)}$ and $d_1^{(k)}$ are given by (set $r_1 \equiv 0$)

$$\begin{aligned} d_0^{(n)} &= (\frac{1+2r_n}{1+r_n})^2, \quad d_0^{(k)} = (\frac{1+2r_k}{1+r_k})^2 + \frac{r_{k+1}^3}{(1+r_{k+1})^2}, \quad 1 \le k \le n-1 \\ d_1^{(k)} &= -\frac{r_k^{3/2}(1+2r_k)}{(1+r_k)^2}, \quad 1 \le k \le n. \end{aligned}$$

Using the Gerschgorin circle theorem, one can obtain the upper bound of the maximum eigenvalue of matrix $\tilde{B}_2^T\tilde{B}_2$ as

$$\begin{aligned}\lambda_{\max}(\tilde{B}_2^T\tilde{B}_2) &\le \max_{1\le k\le n-1}\{d_0^{(k)}-d_1^{(k)}-d_1^{(k+1)},\ d_0^{(n)}-d_1^{(n)}\}\\ &= \max_{1\le k\le n-1}\{\hat{R}(r_k,r_{k+1}),\hat{R}(r_n,0)\}\le \max_{1\le k\le n}\hat{R}(r_k,r_{k+1}).\end{aligned}\tag{7.54}$$

It is easy to verify that $\hat{R}(x,y)$ is increasing with respect to x and y. Hence, by the time-step ratio condition **A1** , we have

$$\begin{aligned}\lambda_{\max}(\tilde{B}_2^T\tilde{B}_2) &\le \max_{1\le k\le n}\hat{R}(r_k,r_{k+1})\le \hat{R}(r_{\max}-\delta,r_{\max}-\delta)\\ &\le \hat{R}(r_{\max},r_{\max}) = \frac{4r_{\max}^3}{(1+r_{\max})^2}\approx 13.3880<14.\end{aligned}$$

The proof is complete

The Proof of Lemma 4.5:

The first claim holds by a simple calculation that

$$\widehat{\boldsymbol{B}}_2^{-1}\widehat{\boldsymbol{B}}(\widehat{\boldsymbol{B}}_2^{-1})^T=\widehat{\boldsymbol{B}}_2^{-1}(\widehat{\boldsymbol{B}}_2^T+\widehat{\boldsymbol{B}}_2)(\widehat{\boldsymbol{B}}_2^{-1})^T=(\widehat{\boldsymbol{B}}_2^{-1})^T+(\widehat{\boldsymbol{B}}_2^{-1})=\widehat{\boldsymbol{\Theta}}.$$

We now prove the second claim. It follows from Lemma 4.3 that

$$\widehat{\boldsymbol{B}}=(\widehat{\boldsymbol{L}}\widehat{\boldsymbol{\Lambda}}_\tau^{-1})^T\widehat{\boldsymbol{L}}\widehat{\boldsymbol{\Lambda}}_\tau^{-1},$$

together with the Young's inequality, one has

$$\begin{aligned}\sum_{k=1}^{n}\sum_{j=1}^{k}\theta_{k-j}^{(k)}(\boldsymbol{v}^k)^T\boldsymbol{w}^j &= \boldsymbol{w}^T\widehat{\boldsymbol{\Theta}}_2^T\boldsymbol{v}=\boldsymbol{w}^T(\widehat{\boldsymbol{L}}\widehat{\boldsymbol{\Lambda}}_\tau^{-1})^{-1}\widehat{\boldsymbol{L}}\widehat{\boldsymbol{\Lambda}}_\tau^{-1}\widehat{\boldsymbol{\Theta}}_2^T\boldsymbol{v}\\ &\le \frac{\epsilon}{2}\boldsymbol{v}^T(\widehat{\boldsymbol{L}}\widehat{\boldsymbol{\Lambda}}_\tau^{-1}\widehat{\boldsymbol{\Theta}}_2^T)^T\widehat{\boldsymbol{L}}\widehat{\boldsymbol{\Lambda}}_\tau^{-1}\widehat{\boldsymbol{\Theta}}_2^T\boldsymbol{v}+\frac{1}{2\epsilon}\boldsymbol{w}^T(\widehat{\boldsymbol{L}}\widehat{\boldsymbol{\Lambda}}_\tau^{-1})^{-1}(\widehat{\boldsymbol{L}}\widehat{\boldsymbol{\Lambda}}_\tau^{-1})^{-T}\boldsymbol{w}\\ &\le \frac{\epsilon}{2}\boldsymbol{v}^T\widehat{\boldsymbol{\Theta}}\boldsymbol{v}+\frac{1}{2\epsilon}\boldsymbol{w}^T\widehat{\boldsymbol{B}}^{-1}\boldsymbol{w}.\end{aligned}$$

The proof is complete.

The Proof of Lemma 4.6:

Consider the gradient matrix of $\boldsymbol{f}$ with respect to $\boldsymbol{v}=(v_1,v_2)$, namely

$$\nabla\boldsymbol{f}(\boldsymbol{v})=\frac{1}{1+|\boldsymbol{v}|^2}\begin{pmatrix}1-v_1^2+v_2^2 & -2v_1v_2\\ -2v_1v_2 & 1-v_1^2+v_2^2\end{pmatrix}.$$

By calculating the eigenvalues of $\nabla\boldsymbol{f}$, one has

$$\mu_1(\boldsymbol{v})=\frac{1-|\boldsymbol{v}|^2}{(1+|\boldsymbol{v}|^2)^2},\quad \mu_2(\boldsymbol{v})=\frac{1}{1+|\boldsymbol{v}|^2},$$

with $-\frac{1}{8} \leq \mu_1(\boldsymbol{v}) \leq 1$ and $0 < \mu_2(\boldsymbol{v}) \leq 1$. The application of Taylor expansion yields

$$\boldsymbol{f}(\boldsymbol{v}) = \boldsymbol{f}(\boldsymbol{w}) + \int_0^1 \nabla \boldsymbol{f}(\theta \boldsymbol{v} + (1-\theta)\boldsymbol{w})\, \mathrm{d}\theta (\boldsymbol{v} - \boldsymbol{w}).$$

From the symmetry of $\nabla \boldsymbol{f}(\boldsymbol{v})$, there exists an orthogonal matrix $\begin{pmatrix} a_\theta & b_\theta \\ c_\theta & d_\theta \end{pmatrix}$ such that

$$\begin{pmatrix} \lambda_1 & \\ & \lambda_2 \end{pmatrix} = \int_0^1 \begin{pmatrix} a_\theta & b_\theta \\ c_\theta & d_\theta \end{pmatrix} \begin{pmatrix} \mu_1(\boldsymbol{\xi}_\theta) & \\ & \mu_2(\boldsymbol{\xi}_\theta) \end{pmatrix} \begin{pmatrix} a_\theta & c_\theta \\ b_\theta & d_\theta \end{pmatrix} \mathrm{d}\theta,$$

where $\boldsymbol{\xi}_\theta = \theta \boldsymbol{v} + (1-\theta)\boldsymbol{w}$, and λ_1, λ_2 are the eigenvalues of matrix $\int_0^1 \nabla \boldsymbol{f}(\boldsymbol{\xi}_\theta)\, \mathrm{d}\theta$, and $\mu_1(\boldsymbol{\xi}_\theta), \mu_2(\boldsymbol{\xi}_\theta)$ are the eigenvalues of matrix $\nabla \boldsymbol{f}(\boldsymbol{\xi}_\theta)$. Without loss of generality, here we assume $\mu_1(\boldsymbol{\xi}_\theta) \leq \mu_2(\boldsymbol{\xi}_\theta)$. It is easy to verify

$$\lambda_1 = \int_0^1 (a_\theta^2 \mu_1(\xi_\theta) + b_\theta^2 \mu_2(\xi_\theta))\, \mathrm{d}\theta, \quad \lambda_2 = \int_0^1 (c_\theta^2 \mu_1(\xi_\theta) + d_\theta^2 \mu_2(\xi_\theta))\, \mathrm{d}\theta.$$

The orthogonality of matrix $\begin{pmatrix} a_\theta & b_\theta \\ c_\theta & d_\theta \end{pmatrix}$ yields $a_\theta^2 + b_\theta^2 = 1, c_\theta^2 + d_\theta^2 = 1$, which implies $\mu_1(\xi_\theta) \leq \lambda_1, \lambda_2 \leq \mu_2(\xi_\theta)$. Hence, we have $Q_{\boldsymbol{f}} = \int_0^1 \nabla \boldsymbol{f}(\theta \boldsymbol{v} + (1-\theta)\boldsymbol{w})\, \mathrm{d}\theta$ and its eigenvalues λ_1, λ_2 satisfy $-\frac{1}{8} \leq \lambda_1, \lambda_2 \leq 1$, which implies $\|Q_{\boldsymbol{f}}\|_2 \leq 1$. By using the property of matrix norm that $|A\boldsymbol{x}| \leq \|A\|_2 |\boldsymbol{x}|$, one has

$$|\boldsymbol{f}(\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{w})| \leq \|\nabla \boldsymbol{f}(\theta \boldsymbol{v} + (1-\theta)\boldsymbol{w})\|_2 |\boldsymbol{v} - \boldsymbol{w}| \leq |\boldsymbol{v} - \boldsymbol{w}|.$$

The proof is complete.

The Proof of Lemma 4.7:

From Lemma 4.6, there exist symmetric matrix sequences $Q_f^j \in \mathbb{R}^{2\times 2}$ such that

$$\boldsymbol{f}(\boldsymbol{v}^j + \boldsymbol{w}^j) - \boldsymbol{f}(\boldsymbol{v}^j) = Q_f^j \boldsymbol{w}^j, \quad 1 \leq j \leq n,$$

where the eigenvalues μ_1^j, μ_2^j of Q_f^j satisfy $-\frac{1}{8} \leq \mu_1^j, \mu_2^j \leq 1$.

Define the symmetric matrix

$$Q := \mathrm{diag}(Q_f^1, \cdots, Q_f^n) \in \mathbb{R}^{2n\times 2n}.$$

The spectral radius of the symmetric matrix Q satisfies $\rho(Q) \leq 1$, which implies that $\|Q\|_2 \leq 1$. According to Lemma 4.5, one has

$$\begin{aligned} \sum_{k=1}^n \sum_{j=1}^k \theta_{k-j}^{(k)} (\boldsymbol{z}^k)^T [\boldsymbol{f}(\boldsymbol{v}^j + \boldsymbol{w}^j) - \boldsymbol{f}(\boldsymbol{v}^j)] &= \sum_{k=1}^n \sum_{j=1}^k \theta_{k-j}^{(k)} (\boldsymbol{z}^k)^T Q_f^j \boldsymbol{w}^j \\ &\leq \frac{\epsilon}{2} \boldsymbol{z}^T \widehat{\boldsymbol{\Theta}} \boldsymbol{z} + \frac{1}{2\epsilon} \boldsymbol{w}^T Q^T \widehat{\boldsymbol{B}}^{-1} Q \boldsymbol{w} \\ &= \epsilon \sum_{k=1}^n \sum_{j=1}^k \theta_{k-j}^{(k)} (\boldsymbol{z}^k)^T \boldsymbol{z}^j + \frac{1}{2\epsilon} \boldsymbol{w}^T Q^T \widehat{\boldsymbol{B}}^{-1} Q \boldsymbol{w}, \end{aligned} \tag{7.55}$$

where $\boldsymbol{z} = ((\boldsymbol{z}^1)^T, \cdots, (\boldsymbol{z}^n)^T)^T$ and $\boldsymbol{w} = ((\boldsymbol{w}^1)^T, \cdots, (\boldsymbol{w}^n)^T)^T$. It follows from Lemmas 4.3 and 4.5 that

$$\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{B}}_2^{-1})^T \widehat{\boldsymbol{B}} \widehat{\boldsymbol{B}}_2^{-1} = (\widehat{\boldsymbol{L}} \widehat{\boldsymbol{\Lambda}}_\tau^{-1} \widehat{\boldsymbol{B}}_2^{-1})^T \widehat{\boldsymbol{L}} \widehat{\boldsymbol{\Lambda}}_\tau^{-1} \widehat{\boldsymbol{B}}_2^{-1},$$

which implies

$$\boldsymbol{w}^T \widehat{\boldsymbol{\Theta}} \boldsymbol{w} = \| \widehat{\boldsymbol{L}} \widehat{\boldsymbol{\Lambda}}_\tau^{-1} \widehat{\boldsymbol{B}}_2^{-1} \boldsymbol{w} \|_2^2.$$

Then the right term of the last identity in (7.55) can be estimated by

$$\begin{aligned}
\boldsymbol{w}^T \widehat{\boldsymbol{Q}}^T \widehat{\boldsymbol{B}}^{-1} Q \boldsymbol{w} &= \| (\widehat{\boldsymbol{L}}^{-1})^T \widehat{\boldsymbol{\Lambda}}_\tau \widehat{\boldsymbol{Q}} \boldsymbol{w} \|_2^2 \\
&= \| (\widehat{\boldsymbol{L}}^{-1})^T \widehat{\boldsymbol{\Lambda}}_\tau Q \widehat{\boldsymbol{B}}_2 \widehat{\boldsymbol{\Lambda}}_\tau \widehat{\boldsymbol{L}}^{-1} \widehat{\boldsymbol{L}} \widehat{\boldsymbol{\Lambda}}_\tau^{-1} \widehat{\boldsymbol{B}}_2^{-1} \boldsymbol{w} \|_2^2 \\
&\le \| (\widehat{\boldsymbol{L}}^{-1})^T \widehat{\boldsymbol{\Lambda}}_\tau Q \widehat{\boldsymbol{B}}_2 \widehat{\boldsymbol{\Lambda}}_\tau \widehat{\boldsymbol{L}}^{-1} \|_2^2 \boldsymbol{w}^T \widehat{\boldsymbol{\Theta}} \boldsymbol{w} \\
&\le \| Q \|_2^2 \| \widetilde{\boldsymbol{B}}_2 \|_2^2 \| \widehat{\boldsymbol{L}}^{-1} \|_2^4 \boldsymbol{w}^T \widehat{\boldsymbol{\Theta}} \boldsymbol{w} \le 2 \mathcal{Q}_\delta \sum_{k=1}^{n} \sum_{j=1}^{k} \theta_{k-j}^{(k)} (\boldsymbol{w}^j)^T \boldsymbol{w}^k,
\end{aligned}$$

where the commutativity $\widehat{\boldsymbol{\Lambda}}_\tau Q = Q \widehat{\boldsymbol{\Lambda}}_\tau$ is used and $\mathcal{Q}_\delta$ is defined in (4.37). Inserting the above estimate to (7.55), one can obtain the first claim. The second claim can be proved by taking $\epsilon = \sqrt{\mathcal{Q}_\delta}$ in the first claim. The proof is complete.

References

[1] Becker J. A second order backward difference method with variable steps for a parabolic problem[J]. BIT, 1998, 38: 644–662.

[2] Chen W, Conde S, Wang C, Wang X, Wise S. A linear energy stable scheme for a thin film model without slope selection[J]. J. Sci. Comput., 2012, 52: 546–562.

[3] Chen W, Wang X, Yan Y, Zhang Z. A second order BDF numerical scheme with variable steps for the Cahn-Hilliard equation[J]. SIAM J. Numer. Anal., 2019, 57(1): 495–525.

[4] Crouzeix M, Lisbona F. The convergence of variable-stepsize, variable formula, multistep methods[J]. SIAM J. Numer. Anal., 1984, 21: 512–534.

[5] Emmrich E. Stability and error of the variable two-step BDF for semilinear parabolic problems[J]. J. Appl. Math. Comput., 2005, 19: 33–55.

[6] Evans J, Thiel P, Bartelt M C. Morphological evolution during epitaxial thin film growth: formation of 2D islans and 3D mounds[J]. Surf. Sci. Rep., 2006, 61: 1–128.

[7] Evans J, Thiel P. A little chemistry helps the big get Bigger[J]. Science, 2010, 330: 599–600.

[8] Grigorieff R. Stability of multistep-methods on variable grids[J]. Numer. Math., 1983, 42: 359–377.

[9] Golubović L. Interfacial coarsening in epitaxial growth models without slope selection[J]. Phys. Rev. Lett., 1997, 78: 90–93.

[10] Gear C, Tu K. The effect of variable mesh size on the stability of multistep methods[J]. SIAM J. Num. Anal., 1974, 11: 1025–1043.

[11] Hairer E, NØrsett S, Wanner G. Solving ordinary differential equations I: Nonstiff problems[M]. Berlin: Springer-Verlag, 1987.

[12] Ju L, Li X, Qiao Z, Zhang H. Energy stability and error estimates of exponential time differencing schemes for the epitaxial growth model without slope selection[J]. Math. Comp., 2018, 87: 1859–1885.

[13] Leroux M. Variable step size multistep methods for parabolic problems[J]. SIAM J. Numer. Anal., 1982, 19(4): 725–741.

[14] Liao H, Li D, Zhang J. Sharp error estimate of the nonuniform L1 formula for reaction-subdiffusion equations[J]. SIAM J. Numer. Anal., 2018, 56: 1112–1133.

[15] Li D, Liao H, Wang J, Sun W, Zhang J. Analysis of L1-Galerkin FEMs for time-fractional nonlinear parabolic problems[J]. Commun. Comput. Phys., 2018, 24: 86–103.

[16] Li B, Liu J. Thin film epitaxy with or without slope selection[J]. European J. Appl. Math., 2003, 14: 713–743.

[17] Li D, Wang J, Zhang J. Unconditionally convergent L1-Galerkin FEMs for nonlinear time-fractional Schrödinger equations[J]. SIAM J. Sci. Comp., 2017, 39: 3067–3088.

[18] Liao H, Ji B, Zhang L. An adaptive BDF2 implicit time-stepping method for the phase field crystal model[J]. IMA J. Numer. Anal., 2022, 42(1): 649-679.

[19] Liao H, Song X, Tang T, Zhou T. Analysis of the second order BDF scheme with variable steps for the molecular beam epitaxial model without slope selection[J]. Sci. China Math., 2021, 64(5): 887-902.

[20] Liao H, Zhang Z. Analysis of adaptive BDF2 scheme for diffusion equations[J]. Math. Comput., 2020, 90: 1207–1226.

[21] Qiao Z, Sun Z, Zhang Z. Stability and convergence of second-order schemes for the nonlinear epitaxial growth model without slope selection[J]. Math Comp., 2015, 84: 653–674.

[22] Ratke L, Voorhees P. Growth and coarsening[M]. Berlin: Springer-Verlag, 2002.

[23] Shampine L, Reichelt M. The MATLAB ODE suite[J]. SIAM J. Sci. Comput., 1997, 18: 1–22.

[24] Shen J, Wang C, Wang X, Wise S. Second-order convex splitting schemes for gradient flows with Ehrlich-Schwoebel type energy: application to thin film epitaxy[J]. SIAM J. Numer. Anal., 2012, 50: 105–125.

[25] Thomée V. Galerkin finite element methods for parabolic problems, Second Edition[M]. Springer-Verlag, 2006.

[26] Wang W, Chen Y, Fang H. On the variable two-step IMEX BDF method for parabolic integro-differential equations with nonsmooth initial data arising in finance[J]. SIAM J. Numer. Anal., 2019, 57: 1289–1317.

[27] Xu J, Li Y, Wu S, Bousequet A. On the stability and accuracy of partially and fully implicit schemes for phase field modeling[J]. Comput. Methods Appl. Mech. Engrg., 2019, 345: 826–853.

[28] Xu C, Tang T. Stability analysis of large time-stepping methods for epitaxial growth models[J]. SIAM J. Numer. Anal., 2006, 44: 1759–1779.

[29] Zhang J, Zhao C. Sharp error estimate of BDF2 scheme with variable time steps for linear reaction-diffusion equations[J]. J. Math., 2020, 41: 1–19.

无倾斜选择的分子束外延模型变步长BDF2格式的最优误差估计

张继伟[1], 赵成超[2]

(1. 武汉大学数学与统计学院; 计算科学湖北省重点实验室, 湖北武汉430072)

(2. 北京计算科学研究中心应用与计算数学部, 北京100193)

摘要: 对于没有斜率选择的分子束外延模型，具有可变时间步长的两步向后微分公式（BDF2）的稳定性和收敛性仍未被完全解决。在本文中，我们首先证明了该BDF2格式在新的相邻时间步长比条件下保持修正的能量耗散定律：$r_k := \tau_k/\tau_{k-1} \leq 4.8645 - \delta$，其中$\delta > 0$是给定的任意小常数。然后，我们介绍了最近发展的离散正交卷积（DOC）和离散互补卷积（DCC）核技巧，并在新的比率条件$r_k \leq 4.8645 - \delta$下给出了BDF2格式的鲁棒且最优的二阶收敛性。鲁棒性意味着，除了$r_k \leq 4.8645 - \delta$以外，收敛性不需要其他时间步长上的约束条件。此外，我们的分析表明，使用一阶BDF1格式计算第一步数值解足以确保全局最优收敛阶。也就是说，选择BDF1格式计算起始步的数值解不会导致全局二阶收敛的损失。数值算例验证了我们的理论分析。

关键词: 变步长BDF2; 离散正交卷积(DOC)核;离散互补卷积(DCC)核; 误差卷积结构(ECS); 最优误差估计; 分子束外延(MBE)模型

MR(2010)**主题分类号**: 65M06; 65M12 **中图分类号**: O241.1; O241.82