Vol. 41 (2021)
No. 6

数 学 杂 志
J. of Math. (PRC)

SHARP ERROR ESTIMATE OF BDF2 SCHEME WITH VARIABLE TIME STEPS FOR LINEAR REACTION-DIFFUSION EQUATIONS

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: While the variable time-steps two-step backward differentiation formula (BDF2) is valuable and widely used to capture the multi-scale dynamics of model solutions, the stability and convergence of BDF2 with variable time steps still remain incomplete. In this work, we revisit BDF2 scheme for linear diffusion-reaction problem. By using the technique of the discrete orthogonal convolution (DOC) kernels developed in [11], and introducing the concept of the discrete complementary convolution (DCC) kernels, we present that BDF2 scheme is unconditionally stable under a adjacent time-step ratio condition: $0 < r_k := \tau_k/\tau_{k-1} \leq r_{\max} \approx 4.8645$. With the uses of DOC and DCC kernels, the second-order temporal convergence can be achieved under $0 < r_k \leq r_{\max} \approx 4.8645$. Our analysis shows that the second-order convergence is sharp and robust. The robustness means that the second-order convergence is sharp for any time step satisfying $0 < r_k \leq r_{\max} \approx 4.8645$, this is, it does not need extra restricted conditions on the time steps. In addition, our analysis also shows that the first level solution u^1 obtained by BDF1 (i.e. Euler scheme) does not cause the loss of global accuracy of second order with $0 < r_k \leq 4.8645$. Numerical examples are provided to demonstrate our theoretical analysis.

Keywords: BDF2; DOC; DCC; variable time steps; sharp error estimate

2010 MR Subject Classification: 65M06; 65M12

Document code: A **Article ID:** 0255-7797(2021)06-0471-18

1 Introduction

In this paper, we revisit two-step backward differentiation formula (BDF2) with variable time steps for solving the linear reaction-diffusion equation:

$$\begin{aligned} &u_t = \Delta u + \kappa u + f(x,t), && x \in \Omega, t \in (0,T], \\ &u(x,0) = u_0(x), && x \in \bar{\Omega}, \\ &u(x,t) = 0, && x \in \partial\Omega, t \in [0,T], \end{aligned} \tag{1.1}$$

where the reaction coefficient $\kappa \in \mathbb{R}$, and Ω is a bounded domain.

Set the generally nonuniform time levels $0 = t_0 < t_1 < t_2 < \cdots < t_N = T$ with the kth time-step size $\tau_k := t_k - t_{k-1}$, the maximum step size $\tau := \max_{1\leq k\leq N} \tau_k$, and the adjacent

* **Received date:** 2021-09-13 **Accepted date:** 2021-10-15

Foundation item: Supported by NSFC under grant Nos. 11771035 and NSAF U1930402, the Natural Science Foundation of Hubei Province No. 2019CFA007. The numerical simulations in this work have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

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

Corresponding author: Zhao Chengchao

time-step ratios

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

The BDF1 and BDF2 formulas with variable time steps are respectively defined by

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

where the difference operator $\nabla_\tau u^n := u^n - u^{n-1}$ for $1 \le n \le N$.

By taking $b_0^{(1)} := 1/\tau_1$, and for $n > 1$

$$b_0^{(n)} = \frac{1+2r_n}{\tau_n(1+r_n)}, \quad b_1^{(n)} = -\frac{r_n^2}{\tau_n(1+r_n)}, \quad \text{and} \quad b_j^{(n)} = 0 \ \ (\text{for } 2 \le j \le n-1), \tag{1.2}$$

the BDF1 and BDF2 can be written as a unified discrete convolution form

$$\mathcal{D}_2 u^n := \sum_{k=1}^{n} b_{n-k}^{(n)} \nabla_\tau u^k, \qquad n \ge 1. \tag{1.3}$$

For $n = 1$, we use BDF1 scheme to obtain the solution u^1, and for $n > 1$, we use BDF2 scheme. Based on the unified notation (1.3), the BDF2 scheme with variable time steps is given as

$$\mathcal{D}_2 u^n = \Delta u^n + \kappa u^n + f^n, \quad \text{for} \quad 1 \le n \le N. \tag{1.4}$$

The BDF2 with variable time steps is widely used to solve stiff or differential-algebraic problems [3, 4, 6, 15, 16, 17] as it has the nice property of the strong stability. One can refer to [1, 2, 3, 12, 16] for the details. While the practical use of BDF2 is well developed, the theoretical analysis seems to be difficult. Even so, many excellent mathematicians still make a big progress on the analysis of BDF2 scheme with variable time steps.

For the stability analysis of problem (1.1) with $\kappa = 0$, twenty years ago Becker [1] (one also refers to Thomée's classical book [16, Lemma 10.6]) presented the bound under the ratio condition $0 < r_k \le (2+\sqrt{13})/3 \approx 1.868$ that

$$\|u^n\| \le C\exp(C\Gamma_n)\big(\,\|u_0\| + \sum_{j=1}^{n}\tau_j\,\big\|f^j\big\|\,\big) \quad \text{for} \quad n \ge 1, \tag{1.5}$$

where $\Gamma_n := \sum_{k=2}^{n-2}\max\{0, r_k - r_{k+2}\}$ and $\|\cdot\|$ denotes the L_2-norm. As pointed out in [16] and [2], the magnitudes of Γ_n can be zero, bounded [16, pp. 175] and unbounded [2, Remark 4.1] by selecting certain step-ratio sequence and vanishing step sizes. After that, Emmrich [3] improves the Becker's constrained condition to $0 < r_k \le 1.91$, but still keeps the undesirable factor $\exp(C\Gamma_n)$ in the L_2-norm stability. Recently, Chen *et al.* [2] present the energy stability for the Cahn-Hilliard equation under a new ratio condition $0 \le r_k \le (3+\sqrt{17})/2 \approx 3.561$, and then Liao and Zhang [11] propose the technique of the discrete orthogonal convolution (DOC) kernels and present the stability estimate for the linear problem (1.1) under the same ratio condition. The new ratio condition improves Grigorieff's stability condition $0 \le r_k < 1+\sqrt{2}$ given nearly forty years ago [5]. One also refers to [13] and [6, Section III.5] a classical book by Hairer *et al.* In addition, Liao and Zhang [11] obtain the first-order convergence under $0 \le r_k \le 3.561$. If the second-order

convergence is expected in [11], they need an extra restriction condition $|\mathfrak{R}_p| \leq N_0 \ll N$ with the index set

$$\mathfrak{R}_p = \left\{ k \;\middle|\; 1+\sqrt{2} \leq r_k \leq (3+\sqrt{17})/2 \right\}. \tag{1.6}$$

As pointed out in [11], the condition $|\mathfrak{R}_p| \leq N_0 \ll N$ seems to be more theoretical rather than practical. For more practical applications, it is natural to ask if we can theoretically extend the restriction on adjacent time-step ratios and without any extra restriction condition like $|\mathfrak{R}_p| \leq N_0 \ll N$, meanwhile, keep the sharp error estimate.

The aim of this paper is to extend $0 \leq r_k \leq 3.561$ to a new ratio condition $0 < r_k < 4.8645$ and prove that the second-order convergence of BDF2 scheme is sharp and robust. To do so, we use the concept of DOC kernels originally developed in [11], and also introduce the concept of discrete complementary convolution (DCC) kernels developed for solving fractional PDEs [7, 8, 9, 10]. Based on DOC and DCC kernels, we first present the corresponding energy (H^1-norm) stability estimate with a new adjacent time-step ratio condition

$$\textbf{A1}\text{: } 0 < r_k \leq r_{\max} = \tfrac{1}{6}\left(\sqrt[3]{1196-12\sqrt{177}} + \sqrt[3]{1196+12\sqrt{177}} \right) + \tfrac{4}{3} \approx 4.8645, \quad \text{for } 2 \leq k \leq N.$$

Here $r_{\max}$ is the positive real solution of $x^3 = (2x+1)^2$, see the details in Lemma 2.1.

For the sharp and robust convergence, we further express the local truncated error by an error convolution structure (**ECS**) with the BDF2 kernel, see more details in Lemma 3.9. Using the definition of DOC kernels, the ECS can significantly circumvent the complex calculation of BDF2 and DOC kernels. Thus, we have the sharp and robust second-order convergence under the ratio condition **A1** (i.e. $0 < r_k \leq 4.8645$). The robustness means the error estimate only depends on the adjacent step ratio restriction **A1**, and does not suffer from other conditions on the mesh sizes, like the restricted condition $|\mathfrak{R}_p| \leq N_0 \ll N$ in [11]. In this sense, the second-order convergence is robust for variable time step sizes. On the other hand, our analysis also shows that the first-order BDF1 for the first level solution u^1 is enough to have the sharp second-order convergence. Thus, our analysis removes the doubt of the choice of the first level solution u^1 computed by BDF1, which further improves the nice results in [11, 14].

The organization of the paper is given as follows. In Section 2, we present the semi-positive definition of BDF2 kernels under condition **A1**, and the properties of DOC and DCC kernels. The stability analysis and second-order convergence of the BDF2 scheme (1.3) are given in Section 3. Numerical examples are provided to demonstrate our theoretical analysis.

2 The Properties of BDF2, DOC and DCC Kernels

In this section, we first consider the positive semi-definiteness of BDF2 convolution kernels and the properties of DOC and DCC kernels, which are useful for the analysis of stability and convergence of BDF2 scheme in section 3.

2.1 Positive Semi-Definiteness of BDF2 Convolution Kernels

We first consider the positive semi-definiteness of BDF2 convolution kernels $b_{n-k}^{(n)}$. It has been proven in [11] that the BDF2 convolution kernels $b_{n-k}^{(n)}$ are positive semi-definite under the ratio condition $0 < r_k \leq 3.561$. A natural question is whether the ratio condition can be relaxed. In this subsection, we will prove the positive semi-definiteness of BDF2 convolution kernels $b_{n-k}^{(n)}$ under a new ratio condition **A1** (i.e., $0 < r_k \leq 4.8645$). To this

end, we first present the following lemma which plays a key role in the proof of the positive semi-definiteness of $b_{n-k}^{(n)}$.

Lemma 2.1 There exist typical values $\epsilon_* > 0$ and $x_{\max} > 0$ such that

$$\mathfrak{F}(x, y, \epsilon_*) := \frac{2\epsilon_* + 4\epsilon_* x - \epsilon_*^2 x^2}{(1+x)} - \frac{y}{(1+y)} \geq 0, \qquad 0 < x, y \leq x_{\max}, \tag{2.7}$$

where $\epsilon_* = 1/\sqrt{x_{\max}}$ and $x_{\max}$ is the positive root of the equation $x^3 = (1+2x)^2$.

Proof We now present the details how to find ϵ_* and $x_{\max}$. Set $y = x$ in (2.7) and consider the quadratic function

$$\mathfrak{H}(x) = -\epsilon^2 x^2 + (4\epsilon - 1)x + 2\epsilon. \tag{2.8}$$

The positive root of $\mathfrak{H}(x) = 0$ is given by

$$x = \frac{4\epsilon - 1 + \sqrt{8\epsilon^3 + 16\epsilon^2 - 8\epsilon + 1}}{2\epsilon^2}. \tag{2.9}$$

Noting x is a function of ϵ. We can produce its maximum by searching ϵ_* such that $x' = 0$. To do so, we take the derivative of x with respect to ϵ as

$$x' = \frac{-2\epsilon^3 - 8\epsilon^2 + 6\epsilon - 1 - (2\epsilon - 1)\sqrt{8\epsilon^3 + 16\epsilon^2 - 8\epsilon + 1}}{\epsilon^3\sqrt{8\epsilon^3 + 16\epsilon^2 - 8\epsilon + 1}}.$$

To find an ϵ_* such that $x' = 0$, we only need to consider

$$\begin{aligned}
& -2\epsilon_*^3 - 8\epsilon_*^2 + 6\epsilon_* - 1 - (2\epsilon_* - 1)\sqrt{8\epsilon_*^3 + 16\epsilon_*^2 - 8\epsilon_* + 1} = 0, \\
\Rightarrow \quad & (2\epsilon_* - 1)^2(8\epsilon_*^3 + 16\epsilon_*^2 - 8\epsilon_* + 1) = (2\epsilon_*^3 + 8\epsilon_*^2 - 6\epsilon_* + 1)^2, \\
\Rightarrow \quad & (\epsilon_*^3 + 2\epsilon_* - 1)\epsilon_*^3 = 0, \\
(\epsilon_* \neq 0) \Rightarrow \quad & \epsilon_*^3 + 2\epsilon_* - 1 = 0.
\end{aligned}$$

Thus, we have the positive root of $\epsilon_*^3 + 2\epsilon_* - 1 = 0$ given as

$$\epsilon_* = \frac{\sqrt[3]{12}}{6}\left(\sqrt[3]{\sqrt{177} + 9} - \sqrt[3]{\sqrt{177} - 9}\right) \approx 0.4534.$$

Set $g(\epsilon) = \sqrt{8\epsilon^3 + 16\epsilon^2 - 8\epsilon + 1}$. From $x'(\epsilon_*) = 0$, we have

$$\begin{aligned}
& x'(\epsilon_*) = \frac{-2\epsilon_*^3 - 8\epsilon_*^2 + 6\epsilon_* - 1 - (2\epsilon_* - 1)g(\epsilon_*)}{\epsilon_*^3 g(\epsilon_*)} = 0, \\
\Rightarrow \quad & g(\epsilon_*) = \frac{-2\epsilon_*^3 - 8\epsilon_*^2 + 6\epsilon_* - 1}{2\epsilon_* - 1} = \frac{-8\epsilon_*^2 + 10\epsilon_* - 3}{2\epsilon_* - 1},
\end{aligned}$$

where we have used $\epsilon_*^3 = 1 - 2\epsilon_*$ in the last identity.

From (2.9), we have the maximum value $x_{\max}$ at ϵ_* as

$$\begin{aligned}
x_{\max} &= \frac{4\epsilon_* - 1 + g(\epsilon_*)}{2\epsilon_*^2} = \frac{4\epsilon_* - 1 + \frac{-8\epsilon_*^2 + 10\epsilon_* - 3}{2\epsilon_* - 1}}{2\epsilon_*^2} = \frac{1}{\epsilon_*^2} \\
&= \frac{1}{6}\left(\sqrt[3]{1196 - 12\sqrt{177}} + \sqrt[3]{1196 + 12\sqrt{177}}\right) + \frac{4}{3} \approx 4.8645.
\end{aligned} \tag{2.10}$$

From $\epsilon_*^3+2\epsilon_*-1=0$, we have $\epsilon_*^2(\epsilon_*^2+2)^2=1$ and substitute $\epsilon_*=\frac{1}{\sqrt{x_{\max}}}$ into the resulting. Then, the direct calculation shows that $x_{\max}$ satisfies the equation

$$x_{\max}^3=(1+2x_{\max})^2.$$

For the given value ϵ_* and $r_{\max}$, we now prove (2.7) holds. Considering the function

$$\mathfrak{G}(x)=(2\epsilon_*+4\epsilon_* x-\epsilon_*^2x^2)(1+x)^{-1},$$

its derivative $\mathfrak{G}'(x)$ is given by

$$\mathfrak{G}'(x)=-\frac{\big(x-(\sqrt{\epsilon_*(\epsilon_*+2)}-\epsilon_*)\big)\big(x+(\sqrt{\epsilon_*(\epsilon_*+2)}-\epsilon_*)\big)}{\epsilon_*^2(1+x)^2}.$$

When $0<x\le\sqrt{\epsilon_*(\epsilon_*+2)}-\epsilon_*$, we have $\mathfrak{G}'(x)\ge 0$. Hence, it holds that

$$\mathfrak{G}(x)\ge\mathfrak{G}(0)=2\epsilon_*(\approx 0.9068)>\frac{x_{\max}}{1+x_{\max}}(\approx 0.8295).$$

When $\sqrt{\epsilon_*(\epsilon_*+2)}-\epsilon_*<x\le x_{\max}$, we have $\mathfrak{G}'(x)\le 0$. Hence, it holds that

$$\mathfrak{G}(x)\ge\mathfrak{G}(x_{\max})=\frac{x_{\max}}{1+x_{\max}}.$$

Thus, we prove $\mathfrak{F}(x,x_{\max},\epsilon_*)\ge 0$ for $0<x\le x_{\max}$. Noting $\mathfrak{F}(x,y,\epsilon)$ is a decreasing function with respect to y, we have $\mathfrak{F}(x,y,\epsilon_*)\ge\mathfrak{F}(x,x_{\max},\epsilon_*)\ge 0$ for $0<x,y\le x_{\max}$. The proof is complete.

Lemma 2.2 Assume the time step ratio r_k satisfy **A1** (i.e., $0<r_k\le 4.8645$). For any real sequence $\{w_k\}_{k=1}^n$, it holds that

$$2w_k\sum_{j=1}^{k}b_{k-j}^{(k)}w_j\ge\frac{r_{k+1}\sqrt{r_{\max}}}{(1+r_{k+1})}\frac{w_k^2}{\tau_k}-\frac{r_k\sqrt{r_{\max}}}{(1+r_k)}\frac{w_{k-1}^2}{\tau_{k-1}},\quad k\ge 2,\tag{2.11}$$

$$2\sum_{k=1}^{n}w_k\sum_{j=1}^{k}b_{k-j}^{(k)}w_j\ge 0,\quad\text{for } n\ge 1.\tag{2.12}$$

Proof Noting $2ab\le\epsilon a^2+b^2/\epsilon$ $(\forall\epsilon>0)$ and $b_1^{(k)}<0$, we have for $k\ge 2$ that

$$\begin{aligned}2w_k\sum_{j=1}^{k}b_{k-j}^{(k)}w_j&=2b_0^{(k)}w_k^2+2b_1^{(k)}w_kw_{k-1}\ge(2b_0^{(k)}+\epsilon b_1^{(k)})w_k^2+\frac{b_1^{(k)}}{\epsilon}w_{k-1}^2\\&=\frac{2\epsilon+4\epsilon r_k-\epsilon^2r_k^2}{1+r_k}\frac{w_k^2}{\epsilon\tau_k}-\frac{r_k}{1+r_k}\frac{w_{k-1}^2}{\epsilon\tau_{k-1}}\\&=\frac{2\epsilon+4\epsilon r_k-\epsilon^2r_k^2}{1+r_k}\frac{w_k^2}{\epsilon\tau_k}-\frac{r_{k+1}}{1+r_{k+1}}\frac{w_k^2}{\epsilon\tau_k}+\frac{r_{k+1}}{1+r_{k+1}}\frac{w_k^2}{\epsilon\tau_k}-\frac{r_k}{1+r_k}\frac{w_{k-1}^2}{\epsilon\tau_{k-1}}\\&=\underbrace{\left(\frac{2\epsilon+4\epsilon r_k-\epsilon^2r_k^2}{1+r_k}-\frac{r_{k+1}}{1+r_{k+1}}\right)}_{=:\mathfrak{F}(r_k,r_{k+1},\epsilon)}\frac{w_k^2}{\epsilon\tau_k}+\frac{r_{k+1}}{1+r_{k+1}}\frac{w_k^2}{\epsilon\tau_k}-\frac{r_k}{1+r_k}\frac{w_{k-1}^2}{\epsilon\tau_{k-1}}\\&=\frac{r_{k+1}}{1+r_{k+1}}\frac{w_k^2}{\epsilon\tau_k}-\frac{r_k}{1+r_k}\frac{w_{k-1}^2}{\epsilon\tau_{k-1}}+\mathfrak{F}(r_k,r_{k+1},\epsilon)\frac{w_k^2}{\epsilon\tau_k}.\end{aligned}$$

Set $\epsilon = \epsilon_* = 1/\sqrt{r_{\max}}$, it follows from Lemma 2.1 that

$$\mathfrak{F}(r_k, r_{k+1}, \epsilon_*) \geq 0, \quad \forall 0 < r_k, r_{k+1} \leq r_{\max}. \tag{2.13}$$

Thus, the inequality (2.11) holds true.

From the inequality (2.11), the direct calculation yields

$$\begin{aligned} 2\sum_{k=1}^{n} w_k \sum_{j=1}^{k} b_{k-j}^{(k)} w_j &\geq \frac{2}{\tau_1} w_1^2 + \frac{r_{n+1}\sqrt{r_{\max}}}{(1+r_{n+1})} \frac{w_n^2}{\tau_n} - \frac{r_2\sqrt{r_{\max}}}{(1+r_2)} \frac{w_1^2}{\tau_1} \\ &= \frac{r_{n+1}\sqrt{r_{\max}}}{(1+r_{n+1})} \frac{w_n^2}{\tau_n} + \frac{2+(2-\sqrt{r_{\max}})r_2}{(1+r_2)} \frac{w_1^2}{\tau_1} \\ &\geq \frac{r_{n+1}\sqrt{r_{\max}}}{(1+r_{n+1})} \frac{w_n^2}{\tau_n} + \frac{2+(2-\sqrt{r_{\max}})r_{\max}}{(1+r_{\max})} \frac{w_1^2}{\tau_1} \\ &\geq \frac{r_{n+1}\sqrt{r_{\max}}}{(1+r_{n+1})} \frac{w_n^2}{\tau_n} + \frac{w_1^2}{(1+r_{\max})\tau_1} \geq 0, \quad n \geq 1, \end{aligned}$$

where the monotonicity of function $l(x) = \frac{x}{1+x}$ and the fact $2 + (2 - \sqrt{r_{\max}})r_{\max} = 1$ are used. The proof is complete.

2.2 The Relationship Between DOC and DCC Kernels

The DCC kernels $p_{n-j}^{(n)}$ are introduced in analogy of $\int_0^t v'(s)ds = v(t) - v(0)$ such that

$$\sum_{j=1}^{n} p_{n-j}^{(n)} \mathcal{D}_2 u^j = \sum_{j=1}^{n} p_{n-j}^{(n)} \sum_{l=1}^{j} b_{j-l}^{(j)} \nabla_\tau u^l = \sum_{l=1}^{n} \nabla_\tau u^l \sum_{j=l}^{n} p_{n-j}^{(n)} b_{j-l}^{(j)} = u^n - u^0, \quad \forall n \geq 1. \tag{2.14}$$

From the identity (2.14) holds for all $n \geq 1$, we define the DCC kernels by

$$\sum_{j=k}^{n} p_{n-j}^{(n)} b_{j-k}^{(j)} \equiv 1, \quad \forall 1 \leq k \leq n,\ 1 \leq n \leq N. \tag{2.15}$$

From (2.15), the DCC kernels $p_{n-j}^{(n)}$ can be explicitly expressed by the BDF2 kernels $b_{j-k}^{(j)}$, namely,

$$p_0^{(n)} = 1/b_0^{(n)},\ p_{n-j}^{(n)} = \frac{1}{b_0^{(j)}} \sum_{k=j+1}^{n} (b_{k-j-1}^{(n)} - b_{k-j}^{(n)}) p_{n-k}^{(n)} \quad (1 \leq j \leq n-1). \tag{2.16}$$

The discrete orthogonal convolution (DOC) kernels are given in [11] as

$$\sum_{j=k}^{n} \theta_{n-j}^{(n)} b_{j-k}^{(j)} = \delta_{nk}, \quad \text{for all } 1 \leq k \leq n, \tag{2.17}$$

where δ_{nk} represents the Kronecker delta symbol with $\delta_{nk} = 1$ if $n = k$ and $\delta_{nk} = 0$ if $n \neq k$. From the DOC kernels (2.17), we have

$$\sum_{j=1}^{n} \theta_{n-j}^{(n)} \mathcal{D}_2 u^j = \sum_{l=1}^{n} \nabla_\tau u^l \sum_{j=l}^{n} \theta_{n-j}^{(n)} b_{j-l}^{(j)} = u^n - u^{n-1}, \quad 1 \leq n \leq N. \tag{2.18}$$

The two kernels have the following intimate relationship.

Proposition 2.1 The DCC kernels defined by (2.15) and DOC kernels defined in (2.17) have the following relationships

$$p_{n-j}^{(n)} = \sum_{l=j}^{n} \theta_{l-j}^{(l)}, \quad \forall 1 \le j \le n, \tag{2.19}$$

$$\theta_{n-j}^{(n)} = p_{n-j}^{(n)} - p_{n-1-j}^{(n-1)}, \quad \forall 1 \le j \le n, \tag{2.20}$$

where $p_{-1}^{(n)} := 0$ $(\forall n \ge 0)$ are defined.

Proof Set $q_{n-j}^{(n)} = \sum_{l=j}^{n} \theta_{l-j}^{(l)}$ $(\forall 1 \le j \le n)$. Then from the definition (2.17), we have

$$\sum_{j=k}^{n} q_{n-j}^{(n)} b_{j-k}^{(j)} = \sum_{j=k}^{n} \sum_{l=j}^{n} \theta_{l-j}^{(l)} b_{j-k}^{(j)} = \sum_{l=k}^{n} \sum_{j=k}^{l} \theta_{l-j}^{(l)} b_{j-k}^{(j)} = \sum_{l=k}^{n} \delta_{lk} = 1.$$

Hence, $q_{n-j}^{(n)} = \sum_{l=j}^{n} \theta_{l-j}^{(l)}$ $(1 \le j \le n)$ are solutions to (2.15). Noting the DCC kernels uniquely exist due to the explicit expression (2.16). Thus, we have $p_{n-j}^{(n)} = q_{n-j}^{(n)} = \sum_{l=j}^{n} \theta_{l-j}^{(l)}$. The equality (2.20) can be directly yielded by (2.19) and the proof is complete.

2.3 Properties of DOC and DCC Kernels

We now rewrite the definitions of DCC kernel (2.15) and DOC kernel (2.17) by

$$\mathbf{BP} = \mathbf{I}, \qquad \mathbf{B\Theta} = \mathbf{I_0}, \tag{2.21}$$

where we denote

$$\mathbf{B} = \begin{bmatrix} b_0^{(n)} & 0 & 0 & \cdots & 0 & 0 & 0 \\ b_1^{(n)} & b_0^{(n-1)} & 0 & \cdots & 0 & 0 & 0 \\ 0 & b_1^{(n-1)} & b_0^{(n-2)} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b_1^{(3)} & b_0^{(2)} & 0 \\ 0 & 0 & 0 & \cdots & 0 & b_1^{(2)} & b_0^{(1)} \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} p_0^{(n)} \\ p_1^{(n)} \\ p_2^{(n)} \\ \vdots \\ \vdots \\ p_{n-2}^{(n)} \\ p_{n-1}^{(n)} \end{bmatrix}, \ \mathbf{\Theta} = \begin{bmatrix} \theta_0^{(n)} \\ \theta_1^{(n)} \\ \theta_2^{(n)} \\ \vdots \\ \vdots \\ \theta_{n-2}^{(n)} \\ \theta_{n-1}^{(n)} \end{bmatrix},$$

and $\mathbf{I} = [1, 1, 1, \cdots, 1, 1]^T$ and $\mathbf{I}_0 = [1, 0, 0, \cdots, 0, 0]^T$. It is easy to verify that each component of $\mathbf{P}$ and $\mathbf{\Theta}$ is positive by using mathematical induction.

The positive semi-definitiveness of the DOC kernel $\theta_{n-k}^{(n)}$ can be derived in [11] by the positive semi-definitiveness of $b_{n-k}^{(n)}$.

Lemma 2.3 ([11]) If the BDF2 kernels $b_{n-k}^{(n)}$ defined in (1.2) are positive semi-definite, then the DOC kernels $\theta_{n-k}^{(n)}$ defined in (2.17) are also positive semi-definite. This is, it holds for any real sequence $\{\omega_j\}_{j=1}^{n}$ 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 2.4 ([11]) The DOC kernels $\theta_{n-j}^{(n)}$ have the following properties:

$$\sum_{j=1}^{n} \theta_{n-j}^{(n)} = \tau_n, \qquad \text{for } n \geq 1, \tag{2.22}$$

$$\sum_{k=1}^{n} \sum_{j=1}^{k} \theta_{k-j}^{(k)} = t_n, \qquad \text{for } n \geq 1. \tag{2.23}$$

Lemma 2.5 ([11]) The DOC kernel $\theta_{n-j}^{(n)}$ can be explicitly represented by

$$\theta_{n-k}^{(n)} = \begin{cases} \frac{\tau_n(1+r_k)}{1+2r_k} \prod\limits_{i=k+1}^{n} \frac{r_i}{1+2r_i}, & \text{for } 2 \leq k \leq n. \\ \tau_n \prod\limits_{i=k+1}^{n} \frac{r_i}{1+2r_i}, & \text{for } k = 1. \end{cases} \tag{2.24}$$

We point out the results in Lemma 2.5 play an important role for the following convergence analysis and the bound of DCC kernels. We now consider the properties of DCC kernels.

Proposition 2.2 Let τ be the maximum time-step size and the time-step ratios satisfy $0 < r_k \leq r_*$, where r_* is any given positive constant. The DCC kernels $p_{n-k}^{(n)}$ defined in (2.15) satisfy

$$p_{n-j}^{(n)} = \frac{1+r_j}{1+2r_j} \sum_{k=j}^{n} \tau_k \prod_{i=j+1}^{k} \frac{r_i}{1+2r_i}, \quad 2 \leq j \leq n, \tag{2.25}$$

$$p_{n-1}^{(n)} = \sum_{k=1}^{n} \tau_k \prod_{i=2}^{k} \frac{r_i}{1+2r_i}, \tag{2.26}$$

$$\sum_{j=1}^{n} p_{n-j}^{(n)} = t_n, \tag{2.27}$$

$$p_{n-j}^{(n)} \leq \sum_{k=j}^{n} \tau_k \left(\frac{r_*}{1+2r_*} \right)^{k-j} \leq \sum_{k=j}^{n} \frac{\tau_k}{2^{k-j}} \leq 2\tau, \tag{2.28}$$

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

Proof It follows from (2.19) in Proposition 2.1 and Lemma 2.5 that, for $2 \leq j \leq n$,

$$p_{n-j}^{(n)} = \sum_{k=j}^{n} \theta_{k-j}^{(k)} = \frac{1+r_j}{1+2r_j} \sum_{k=j}^{n} \tau_k \prod_{i=j+1}^{k} \frac{r_i}{1+2r_i} \leq \sum_{k=j}^{n} \tau_k \left(\frac{r_*}{1+2r_*} \right)^{k-j},$$

and for $j = 1$,

$$p_{n-1}^{(n)} = \sum_{k=j}^{n} \theta_{k-1}^{(k)} = \sum_{k=1}^{n} \tau_k \prod_{i=2}^{k} \frac{r_i}{1+2r_i} \leq \sum_{k=j}^{n} \tau_k \left(\frac{r_*}{1+2r_*} \right)^{k-j},$$

where the monotonicity of function $h(x) = \frac{x}{1+2x}$ is used. The application of $\frac{r_*}{1+2r_*} \leq \frac{1}{2}$ for any $r_* \geq 0$ yields the last inequality in (2.28). The equality (2.27) can be derived directly by Proposition 2.1 and Lemma 2.4 since

$$\sum_{j=1}^{n} p_{n-j}^{(n)} = \sum_{j=1}^{n}\sum_{k=j}^{n} \theta_{k-j}^{(k)} = \sum_{k=1}^{n}\sum_{j=1}^{k} \theta_{k-j}^{(k)} = \sum_{k=1}^{n} \tau_k = t_n.$$

The proof is complete.

3 Stability and Convergence Analysis for BDF2 Scheme

3.1 Energy Stability

It is known that problem (1.1) with $\kappa \leq 0$ has the property of energy dissipation. We now present the corresponding energy stability for BDF2 scheme (1.4). To the end, we define a (modified) discrete energy E^k by

$$E^k := \frac{r_{k+1}\sqrt{r_{\max}}\tau_k}{1+r_{k+1}}\|\partial_\tau u^k\|^2 + |u^k|_1^2 - \kappa\|u^k\|^2, \quad \text{for } \kappa \leq 0 \text{ and } k \geq 1, \tag{3.29}$$

where the initial energy $E^0 := |u^0|_1^2 - \kappa\|u^0\|^2$ and $\partial_\tau u^k = \nabla_\tau u^k/\tau_k$. Here we remark that our discrete energy E^k defined by (3.29) differs from the one in [11] due to the different modified formula, i.e., the first term in (3.29).

Here and below $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ represent the inner product and norm in $L^2(\Omega)$ space.

Theorem 3.6 Assume the condition **A1** holds and $\kappa \leq 0$, then the discrete solution u^n to the BDF2 scheme (1.4) with variable time steps satisfies

$$\partial_\tau E^k \leq 2\langle f^k, \partial_\tau u^k\rangle, \quad \forall k \geq 1. \tag{3.30}$$

Furthermore, the discrete energy has the following estimate

$$\sqrt{E^n} \leq \sqrt{E^0} + 4C_\Omega\Big(\sum_{k=1}^{n}\|\nabla_\tau f^k\| + \|f^0\|\Big), \quad \forall n \geq 1. \tag{3.31}$$

Proof For $k \geq 2$, the weak form of (1.4) is given as

$$\langle \mathcal{D}_2 u^k, v\rangle = -\langle\nabla u^k, \nabla v\rangle + \kappa\langle u^k, v\rangle + \langle f^k, v\rangle, \quad \forall v \in H_0^1(\Omega),\ 2 \leq k \leq N. \tag{3.32}$$

Setting $\upsilon = 2\nabla_\tau u^k$ in the weak form (3.32), we have

$$2\langle \mathcal{D}_2 u^k, \nabla_\tau u^k\rangle + 2\langle\nabla u^k, \nabla_\tau\nabla u^k\rangle + 2\langle -\kappa u^k, \nabla_\tau u^k\rangle = 2\langle f^k, \nabla_\tau u^k\rangle, \quad k \geq 2. \tag{3.33}$$

It follows from Lemma 2.2 that

$$2\langle \mathcal{D}_2 u^k, \nabla_\tau u^k\rangle \geq \frac{r_{k+1}\sqrt{r_{\max}}\tau_k}{1+r_{k+1}}\|\partial_\tau u^k\|^2 - \frac{r_k\sqrt{r_{\max}}\tau_{k-1}}{1+r_k}\|\partial_\tau u^{k-1}\|^2. \tag{3.34}$$

Applying the inequality $2ab \leq a^2 + b^2$, one has

$$\begin{aligned} &2\langle\nabla u^k, \nabla_\tau\nabla u^k\rangle = 2\|\nabla u^k\|^2 - 2\|\nabla u^k\|\|\nabla u^{k-1}\| \geq \|\nabla u^k\|^2 - \|\nabla u^{k-1}\|^2, \\ &2\langle -\kappa u^k, \nabla_\tau u^k\rangle = -\kappa(2\|u^k\|^2 - 2\|u^k\|\|u^{k-1}\|) \geq -\kappa(\|u^k\|^2 - \|u^{k-1}\|^2). \end{aligned} \tag{3.35}$$

Inserting (3.34) and (3.35) into (3.33) and using the definition (3.29), we arrive at

$$\nabla_\tau E^k \le 2\langle f^k, \nabla_\tau u^k\rangle, \quad \text{for } k \ge 2.$$

We now consider $k = 1$. For $k = 1$, the direct calculation produces

$$2\langle \mathcal{D}_2 u^1, \nabla_\tau u^1\rangle = 2\tau_1\|\partial_\tau u^1\|^2 \ge \frac{r_{\max}^{3/2}}{1+r_{\max}}\tau_1\|\partial_\tau u^1\|^2 \ge \frac{r_2\sqrt{r_{\max}}}{1+r_2}\tau_1\|\partial_\tau u^1\|^2. \tag{3.36}$$

Noting the inequalities (3.35) also hold for $k = 1$, together with (3.36), we have

$$\frac{r_2\sqrt{r_{\max}}}{1+r_2}\tau_1\|\partial_\tau u^1\|^2 + \|\nabla u^1\|^2 - \kappa\|u^1\|^2 \le \|\nabla u^0\|^2 - \kappa\|u^0\|^2 + 2\langle f^1, \nabla_\tau u^1\rangle,$$

which implies

$$\nabla_\tau E^1 \le 2\langle f^1, \nabla_\tau u^1\rangle.$$

Thus, we prove the inequality (3.30).

Taking summation from 1 to n for (3.30), we have

$$\begin{aligned} E^n &\le E^0 + 2\sum_{k=1}^{n}\langle f^k, \nabla_\tau u^k\rangle \\ &= E^0 + 2\langle f^n, u^n\rangle - 2\sum_{k=2}^{n}\langle \nabla_\tau f^k, u^{k-1}\rangle - 2\langle f^1, u^0\rangle \\ &\le E^0 + 2\|f^n\|\|u^n\| + 2\sum_{k=2}^{n}\|u^{k-1}\|\|\nabla_\tau f^k\| + 2\|f^1\|\|u^0\|, \end{aligned} \tag{3.37}$$

where the Cauchy-Schwartz inequality is used to the last term in (3.37). On the other hand, noting the Poincaré inequality produces $\|u^n\| \le C_\Omega|u^n|_1$, we have $\|u^n\| \le C_\Omega\sqrt{E^n}$. Thus, from (3.37), we arrive at

$$E^n \le E^0 + 2C_\Omega(\|f^n\|\sqrt{E^n} + \sum_{k=2}^{n}\sqrt{E^{k-1}}\|\nabla_\tau f^k\| + \|f^1\|\sqrt{E^0}).$$

Choose an integer n_0 $(0 \le n_0 \le n)$ to satisfy $E^{n_0} = \max_{0\le k\le n} E^k$. Then

$$\begin{aligned} E^{n_0} &\le \sqrt{E^0}\sqrt{E^{n_0}} + 2C_\Omega\sqrt{E^{n_0}}(\|f^{n_0}\| + \sum_{k=2}^{n_0}\|\nabla_\tau f^k\| + \|f^1\|) \\ &\le \sqrt{E^0}\sqrt{E^{n_0}} + 4C_\Omega\sqrt{E^{n_0}}(\sum_{k=2}^{n_0}\|\nabla_\tau f^k\| + \|f^1\|), \end{aligned}$$

where the last inequality has used the fact that $f^{n_0} = f^1 + \sum_{k=2}^{n_0}\nabla_\tau f^k$. Noting that $\|f^1\| \le \|\nabla_\tau f^1\| + \|f^0\|$, we have

$$\begin{aligned} \sqrt{E^n} \le \sqrt{E^{n_0}} &\le \sqrt{E^0} + 4C_\Omega(\sum_{k=1}^{n_0}\|\nabla_\tau f^k\| + \|f^0\|) \\ &\le \sqrt{E^0} + 4C_\Omega(\sum_{k=1}^{n}\|\nabla_\tau f^k\| + \|f^0\|). \end{aligned}$$

The proof is complete.

3.2 Stability Analysis of the Discrete Scheme

We first introduce a discrete Grönwall inequality.

Lemma 3.7 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

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

The standard induction hypothesis can give the proof of lemma 3.7, which is omitted here.

Theorem 3.8 If the condition **A1** holds, the solution u^n of BDF2 scheme (1.4) is unconditionally stable in the L^2-norm. If $\kappa > 0$ and the maximum time-step size $\tau \le \frac{1}{4\kappa}$, it holds

$$\begin{aligned} \|u^n\| &\le 2\exp\big(4\kappa t_{n-1}\big)\Big(\big\|u^0\big\| + 2\sum_{k=1}^{n} \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j\| \Big) \\ &\le 2\exp\big(4\kappa t_{n-1}\big)\Big(\big\|u^0\big\| + 2\sum_{j=1}^{n} p_{n-j}^{(n)} \big\|f^j\big\| \Big), \quad \text{for } 1 \le n \le N. \end{aligned} \tag{3.38}$$

If $\kappa \le 0$, it holds

$$\|u^n\| \le \big\|u^0\big\| + 2\sum_{j=1}^{n} \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j\| \le \big\|u^0\big\| + 2\sum_{j=1}^{n} p_{n-j}^{(n)} \big\|f^j\big\|, \quad \text{for } 1 \le n \le N. \tag{3.39}$$

Proof Applying the property (2.18) of DOC kernels to scheme (1.4), we have

$$\nabla_\tau u^k = \sum_{j=1}^{k} \theta_{k-j}^{(k)} (\Delta u^j + \kappa u^j) + \sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j, \quad \text{for} \quad k \ge 1. \tag{3.40}$$

Noting the positive semi-definiteness of the DOC kernels in Lemma 2.3, we have

$$\sum_{k=1}^{n} \sum_{j=1}^{k} \langle u^k, \theta_{k-j}^{(k)} \Delta u^j \rangle = -\sum_{k=1}^{n} \sum_{j=1}^{k} \langle \nabla u^k, \theta_{k-j}^{(k)} \nabla u^j \rangle \le 0. \tag{3.41}$$

Taking the inner product with u^k on both sides of (3.40), summing the resulting from 1 to n and using (3.41), we have

$$\sum_{k=1}^{n} \langle u^k, \nabla_\tau u^k \rangle \le \sum_{k=1}^{n} \langle u^k, \sum_{j=1}^{k} \theta_{k-j}^{(k)} \kappa u^j \rangle + \sum_{k=1}^{n} \langle u^k, \sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j \rangle, \quad \text{for } 1 \le n \le N. \tag{3.42}$$

If $\kappa \leq 0$ in (3.42), we use $\|u^k\|^2 - \|u^{k-1}\|^2 \leq 2\langle u^k, \nabla_\tau u^k\rangle$, the Cauchy-Schwarz inequality and Lemma 2.3 to have

$$\|u^n\|^2 \leq \|u^0\|^2 + 2\sum_{k=1}^{n} \|u^k\| \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j\|, \quad \text{for } 1 \leq n \leq N. \tag{3.43}$$

Selecting an integer n_0 $(0 \leq n_0 \leq n)$ such that $\|u^{n_0}\| = \max_{0\leq k\leq n} \left\|u^k\right\|$. From (3.43), we have

$$\|u^{n_0}\|^2 \leq \left\|u^0\right\| \|u^{n_0}\| + 2\|u^{n_0}\| \sum_{k=1}^{n_0} \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j\|. \tag{3.44}$$

Eliminating a $\|u^{n_0}\|$ for both sides of (3.44) and noting $n_0 \leq n$, we have

$$\begin{aligned}
\|u^n\| \leq \|u^{n_0}\| &\leq \left\|u^0\right\| + 2\sum_{k=1}^{n_0} \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j\| \\
&\leq \left\|u^0\right\| + 2\sum_{k=1}^{n} \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j\| \\
&\leq \left\|u^0\right\| + 2\sum_{j=1}^{n} \|f^j\| \sum_{k=j}^{n} \theta_{k-j}^{(k)} \\
&= \left\|u^0\right\| + 2\sum_{j=1}^{n} p_{n-j}^{(n)} \|f^j\|,
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality, exchanged the order of summation and used the property (2.19).

If $\kappa > 0$ in (3.42), we apply the Cauchy-Schwarz inequality to have

$$\|u^n\|^2 \leq \|u^0\|^2 + 2\kappa \sum_{k=1}^{n} \|u^k\| \sum_{j=1}^{k} \theta_{k-j}^{(k)} \|u^j\| + 2\sum_{k=1}^{n} \|u^k\| \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j\|, \quad \text{for } 1 \leq n \leq N. \tag{3.45}$$

Similar to (3.44) by selecting n_0 $(0 \leq n_0 \leq n)$ such that $\|u^{n_0}\| = \max_{0\leq k\leq n} \left\|u^k\right\|$, one has

$$\|u^{n_0}\|^2 \leq \left\|u^0\right\| \|u^{n_0}\| + 2\kappa\|u^{n_0}\| \sum_{k=1}^{n_0} \|u^k\| \sum_{j=1}^{k} \theta_{k-j}^{(k)} + 2\|u^{n_0}\| \sum_{k=1}^{n_0} \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j\|. \tag{3.46}$$

Eliminating a $\|u^{n_0}\|$ for both sides of (3.46), we further have

$$\begin{aligned}
\|u^n\| \leq \|u^{n_0}\| &\leq \left\|u^0\right\| + 2\kappa\tau_k \sum_{k=1}^{n_0} \left\|u^k\right\| + 2\sum_{k=1}^{n_0} \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j\| \\
&\leq \left\|u^0\right\| + 2\kappa\tau_k \sum_{k=1}^{n} \left\|u^k\right\| + 2\sum_{k=1}^{n} \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j\|,
\end{aligned}$$

where we use the facts $n_0 \le n$ and $\sum\limits_{j=1}^{k} \theta_{k-j}^{(k)} = \tau_k$. Taking the maximum time-step size $\tau \le \frac{1}{4\kappa}$ in the above inequality, we finally arrive at

$$\begin{aligned}\|u^n\| &\le 2\left\|u^0\right\| + 4\kappa\tau_k \sum_{k=1}^{n-1} \left\|u^k\right\| + 4\sum_{k=1}^{n} \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j\| \\ &\le 2\left\|u^0\right\| + 4\kappa\tau_k \sum_{k=1}^{n-1} \left\|u^k\right\| + 4\sum_{j=1}^{n} \|f^j\| \sum_{k=j}^{n} \theta_{k-j}^{(k)} \\ &= 2\left\|u^0\right\| + 4\kappa\tau_k \sum_{k=1}^{n-1} \left\|u^k\right\| + 4\sum_{j=1}^{n} p_{n-j}^{(n)} \|f^j\|.\end{aligned}$$

The Grönwall inequality in Lemma 3.7 directly produces the result (3.38). The proof is complete.

3.3 Convergence Analysis of the Discrete Scheme

Set $e^n := u(t_n, x) - u^n(x)$ $(n \ge 1)$. From (1.4), the error function is governed by

$$\mathcal{D}_2 e^n = \Delta e^n + \kappa e^n + \eta^n, \quad \text{for} \quad 1 \le n \le N, \tag{3.47}$$

where $\eta^n := \mathcal{D}_2 u(t_n) - u_t(t_n) (1 \le n \le N)$ denotes the truncation error.

Lemma 3.9 Denote

$$\begin{aligned} G^l &:= -\frac{1}{2}\int_{t_{l-1}}^{t_l} (t - t_{l-1})^2 u_{ttt}\,\mathrm{d}t, \qquad 1 \le l \le N, \\ R^j &:= -\frac{1}{2} b_1^{(j)} \tau_{j-1} \int_{t_{j-1}}^{t_j} (2(t - t_{j-1}) + \tau_{j-1}) u_{ttt}\,\mathrm{d}t, \qquad 2 \le j \le N, \\ R^1 &:= \frac{1}{2\tau_1}\int_0^{t_1} t^2 u_{ttt}\,\mathrm{d}t - \frac{1}{\tau_1}\int_0^{t_1} t u_{tt}\,\mathrm{d}t. \end{aligned} \tag{3.48}$$

The truncation error $\eta^j := \mathcal{D}_2 u(t_j) - u_t(t_j)$ $(1 \le j \le N)$ can be expressed by the following form

$$\eta^j = \sum_{l=1}^{j} b_{j-l}^{(j)} G^l + R^j, \quad 1 \le j \le N. \tag{3.49}$$

Moreover, we have the follwing estimate

$$\begin{aligned} &2\sum_{k=1}^{n} \|\sum_{j=1}^{k} \theta_{k-j}^{(k)} \eta^j\| \le 2\sum_{k=1}^{n} \|G^k\| + 2\sum_{k=1}^{n} p_{n-k}^{(k)} \|R^k\| \\ \le &4\tau \int_0^{t_1} \|u_{tt}\|\,\mathrm{d}t + \sum_{k=1}^{n} \tau_k^2 \int_{t_{k-1}}^{t_k} \|u_{ttt}\|\,\mathrm{d}t + 2t_n \max_{1\le k\le n} \tau_k \int_{t_{k-1}}^{t_k} \|u_{ttt}\|\,\mathrm{d}t. \end{aligned} \tag{3.50}$$

Proof By using the Taylor's expansion (see [16]), one has

$$\begin{aligned}\eta^j &= \frac{1}{2}b_0^{(j)}G^j + \frac{1}{2}b_1^{(j)}G^{j-1} - \frac{1}{2}b_1^{(j)}\tau_{j-1}\int_{t_{j-1}}^{t_j}(2(t-t_{j-1})+\tau_{j-1})u_{ttt}\,\mathrm{d}t\\ &= \sum_{l=1}^{j} b_{j-l}^{(j)}G^l - \frac{1}{2}b_1^{(j)}\tau_{j-1}\int_{t_{j-1}}^{t_j}(2(t-t_{j-1})+\tau_{j-1})u_{ttt}\,\mathrm{d}t\\ &= \sum_{l=1}^{j} b_{j-l}^{(j)}G^l + R^j, \quad 2\le j\le N,\end{aligned}$$

where the property of BDF2 kernels (1.2) that $b_k^{(j)}=0$ for $k\ge 2$ is used. For $j=1$, using the Taylor's expansion again, one has

$$\begin{aligned}\eta^1 &= \frac{u(t_1)-u(0)}{\tau_1} - u_t(t_1) = -b_0^{(1)}\int_0^{t_1} tu_{tt}\,\mathrm{d}t\\ &= b_0^{(1)}G^1 + \frac{1}{2\tau_1}\int_0^{t_1} t^2u_{ttt}\,\mathrm{d}t - \frac{1}{\tau_1}\int_0^{t_1} tu_{tt}\,\mathrm{d}t\\ &= b_0^{(1)}G^1 + R^1.\end{aligned}$$

Hence, the equality (3.49) holds.

From (3.49), we now estimate

$$\begin{aligned}2\sum_{k=1}^{n}\|\sum_{j=1}^{k}\theta_{k-j}^{(k)}\eta^j\| &\le 2\sum_{k=1}^{n}\|G^k\| + 2\sum_{k=1}^{n}\|\sum_{j=1}^{k}\theta_{k-j}^{(k)}R^j\|\\ &\le 2\sum_{k=1}^{n}\|G^k\| + 2\sum_{k=1}^{n}p_{n-k}^{(k)}\|R^k\|,\end{aligned}\tag{3.51}$$

where the last inequality uses (2.19). Note that G^l, R^j can be bounded by

$$\|G^l\| \le \frac{\tau_l^2}{2}\int_{t_{l-1}}^{t_l}\|u_{ttt}\|\,\mathrm{d}t, \quad 1\le l\le n\ ,\tag{3.52}$$

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

Noting $\sum_{k=1}^{n}p_{n-k}^{(n)} = t_n$ in (2.27) and $p_{n-1}^{(n)}\le 2\tau$ in (2.28), we have

$$\begin{aligned}\sum_{k=1}^{n}p_{n-k}^{(n)}\|R^k\| &= \sum_{k=2}^{n}p_{n-k}^{(n)}\|R^k\| + p_{n-1}^{(n)}\|R^1\|\\ &\le \sum_{k=1}^{n}p_{n-k}^{(n)}\tau_j\int_{t_{j-1}}^{t_j}\|u_{ttt}\|\,\mathrm{d}t + p_{n-1}^{(n)}\int_0^{t_1}\|u_{tt}\|\,\mathrm{d}t\\ &\le t_n\max_{1\le k\le n}\tau_k\int_{t_{k-1}}^{t_k}\|u_{ttt}\|\,\mathrm{d}t + 2\tau\int_0^{t_1}\|u_{tt}\|\,\mathrm{d}t.\end{aligned}\tag{3.53}$$

Inserting (3.52) and (3.53) into (3.51), we have the inequality (3.54). The proof is complete.

Theorem 3.10 Let $u(t,x)$ be the exact solution to problem (1.1). If the condition **A1** holds, then the discrete solution u^n to BDF2 scheme (1.4) has the second-order convergence in the L^2-norm. If $\kappa > 0$ and the maximum time-step size $\tau < 1/(4\kappa)$, it holds

$$\|u(t_n) - u^n\| \leq 2\exp\big(4\kappa t_{n-1}\big)\bigg(\big\|u(0) - u^0\big\| + 4\tau\int_0^{t_1}\|u_{tt}\|\,\mathrm{d}t + \sum_{k=1}^{n}\tau_k^2\int_{t_{k-1}}^{t_k}\|u_{ttt}\|\,\mathrm{d}t + 2t_n\max_{1\leq k\leq n}\tau_k\int_{t_{k-1}}^{t_k}\|u_{ttt}\|\,\mathrm{d}t\bigg), \quad \text{for } 1\leq n\leq N. \tag{3.54}$$

If $\kappa \leq 0$, it holds

$$\|u(t_n) - u^n\| \leq \big\|u(0) - u^0\big\| + 4\tau\int_0^{t_1}\|u_{tt}\|\,\mathrm{d}t + \sum_{k=1}^{n}\tau_k^2\int_{t_{k-1}}^{t_k}\|u_{ttt}\|\,\mathrm{d}t + 2t_n\max_{1\leq k\leq n}\tau_k\int_{t_{k-1}}^{t_k}\|u_{ttt}\|\,\mathrm{d}t, \quad \text{for } 1\leq n\leq N. \tag{3.55}$$

Proof From Theorem 3.8 that if $\kappa > 0$ and the maximum time step $\tau \leq \frac{1}{4\kappa}$, it holds

$$\|u(t_n) - u^n\| \leq 2\exp\big(4\kappa t_{n-1}\big)\Big(\big\|u(0) - u^0\big\| + 2\sum_{k=1}^{n}\|\sum_{j=1}^{k}\theta_{k-j}^{(k)}f^j\|\Big) \quad \text{for } 1\leq n\leq N. \tag{3.56}$$

If $\kappa \leq 0$, it holds

$$\|u(t_n) - u^n\| \leq \big\|u(0) - u^0\big\| + 2\sum_{j=1}^{n}\|\sum_{j=1}^{k}\theta_{k-j}^{(k)}f^j\|. \quad \text{for } 1\leq n\leq N. \tag{3.57}$$

The direct application of Lemma 3.9 to (3.56) and (3.57) produces the error estimates (3.54) and (3.55), respectively. The proof is complete.

Remark 1 For the error estimate of problem (1.1) with $\kappa = 0$, Becker [1] gives an estimate for $0 < r_k < \frac{2+\sqrt{13}}{3} \approx 1.868$, which is improved to $0 < r_k < 1.91$ in [3] later. By choosing different r_k, the fact Γ_n in (1.5) can be bounded [16, pp. 175], unbounded [2, Remark 4.1] or zero. Recently, Liao and Zhang [11] give an improved estimate

$$\|u(t_n) - u^n\| \leq 2\exp\big(4\kappa t_{n-1}\big)\big(\big\|u_0 - u^0\big\| + 2t_n\int_0^{t_1}\|u_{tt}\|\,\mathrm{d}t + 3t_n\max_{1\leq k\leq n}\tau_k\int_{t_{k-1}}^{t_k}\|u_{ttt}\|\,\mathrm{d}t\big) \tag{3.58}$$

with $0 \leq r_k \leq 3.561$. One can see that the right-hand-side second term in (3.58) has the first-order convergence when t_n is large. If they expect to have the second-order convergence, they need an extra restriction condition $|\mathfrak{R}_p| \leq N_0 \ll N$ with the index set defined by (1.6). A similar error estimate is given in [18] with $0 < r_k < \sqrt{2} + 1$.

Our result in Theorem 3.10 shows the sharp second-order convergence under $0 < r_k \leq 4.8645$. And the second-order convergence is robust, which means the convergence order remains valid for any time step satisfying the ratio $0 < r_k \leq r_{\max} \approx 4.864$. As far as we know, it is a pioneer paper to clarify the robust and sharp second-order convergence under the new ratio $0 < r_k \leq 4.864$.

Table 1 Numerical accuracy on random time mesh for $\kappa = 0$

N	$e(N)$	Order	τ	max r_k
32	1.5345e-04	–	5.9985 e-02	16.754
64	3.7947e-05	2.0157	2.9585e-02	42.059
128	9.4821e-06	2.0007	1.4825e-02	86.0224
256	2.3648e-06	2.0035	7.4163e-03	167.412

Table 2 Numerical accuracy on random time mesh for $\kappa = 4$

N	$e(N)$	Order	τ	max r_k
32	1.9048e-04	–	6.2473e-02	104.606
64	4.7853e-05	1.9930	2.7617e-02	10.0333
128	1.1964e-05	1.9999	1.4900e-02	48.278
256	3.0186e-06	1.9867	7.9698e-03	430.559

4 Numerical Experiment

We now report two examples to investigate the convergence order of BDF2 scheme (1.4) with variable time-steps. In the simulations, we set the computational domain $\Omega = (0, 2)^2$, final time $T = 1$, the number of spatial mesh M chosen by $M = N$. By taking

$$f = (\pi^2/2 - \kappa - 1)\exp(-t)\sin(\pi x/2)\sin(\pi y/2),$$

we can construct an exact solution to problem (1.1) as a benchmark solution in the form of

$$u = \exp(-t)\sin(\pi x/2)\sin(\pi y/2).$$

The time meshes are constructed by the random time-steps $\tau_k = T\chi_k/C$, where $C = \sum_{k=1}^{N}\chi_k$ and χ_k is randomly drawn from the uniform distribution on $(0, 1)$. In each run, the discrete L^2-norm at the final time $T = 1$

$$e(N) = h\sqrt{\sum_{1\le i,j\le M}(u(x_i, y_j, T) - u_h^N(x_i, y_j))^2}$$

is recorded in Tables 1 and 2, in which we also list the maximum time-step τ and maximum adjacent time-step ratio. The numerical rate of convergence is calculated by

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

From the current data and more tests not listed here, we see that the BDF2 scheme is robustly stable and convergent in the second order, which is consistent with our theoretical analysis. Due to the time step is randomly chosen without any constrain condition, one can see the first-step BDF1 does not bring the loss of accuracy, which again implies the effectiveness of our analysis.

5 Conclusion

With the applications of DCC and DOC kernels, we present the stability and convergence analysis of BDF2 scheme with variable time-steps under condition **A1**. We extend

the adjacent time step to a new ratio $r_k := \tau_k/\tau_{k-1} \leq r_{\max} = 4.8645$, and obtain the robust and sharp second-order convergence without the extra constrained condition on ratios in [11]. Our convergence results shows that the BDF1 scheme for first step is enough to have the globally optimal second-order convergence. This conclusion removes the doubt of the choice of the first level solution with first-order accuracy. Numerical results are provided to demonstrate the theoretical analysis.

The technique developed in this work can be extended to a family of multi-step schemes with variable time-steps for the stability and convergence analysis. The main challenge is how to explore the useful properties of DOC and DCC kernels, and multi-step schemes's kernels.

Acknowledgements

The authors would like to thank professor Tao Tang, professor Zhimin Zhang and Dr. Honglin Liao for their valuable discussions on this topic.

References

[1] Becker J. A second order backward difference method with variable steps for a parabolic problem[J]. BIT, 1998, 38(4), 644–662.

[2] 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.

[3] Emmrich E. Stability and error of the variable two-step BDF for semilinear parabolic problems[J]. J. Appl. Math. Comput., 2005, 19(1-2): 33–55.

[4] Gear C, Tu K. The effect of variable mesh size on the stability of multistep methods[J]. SIAM J. Num. Anal., 1974, 11(5): 1025–1043.

[5] Grigorieff R. Stability of multistep-methods on variable grids[J]. Numer. Math., 1983, 42(3): 359–377.

[6] Hairer E, Nørsett S, Wanner G. Solving ordinary differential equations I: Nonstiff problems[M]. New York: Springer-Verlag, 1993.

[7] 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(1): 86–103.

[8] Li D, Wang J, Zhang J. Unconditionally convergent L1-Galerkin FEMs for nonlinear time-fractional Schrödinger equations[J]. SIAM J. Sci. Comp., 2017, 39(6): A3067–A3088.

[9] 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(2): 1112–1133.

[10] Liao H, McLean W, Zhang J. A discrete Grönwall inequality with application to numerical schemes for fractional reaction-subdiffusion problems[J]. SIAM J. Numer. Anal., 2019, 57(1): 218–237.

[11] Liao H, Zhang Z. Analysis of adaptive BDF2 scheme for diffusion equations[J]. Math. Comput, 2021, 90(3): 1207–1226.

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

[13] Crouzeix M, Lisbona F. The convergence of variable-stepsize, variable formula, multistep methods[J]. SIAM J. Numer. Anal., 1984, 21(3): 512–534.

[14] Nishikawa H. On large start-up error of BDF2[J]. J. Comput. Phys., 2019, 392: 456–461.

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

[16] Thomée V. Galerkin finite element methods for parabolic problems[J]. Mathematics of Computation, 2006, 17(2): 186–187.

[17] 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(3): 1289–1317.

[18] Wang W, Mao M, Wang Z. Stability and error estimates for the variable step-size BDF2 method for linear and semilinear parabolic equations[J]. Adv. Comput. Math., https://doi.org/10.1007/s10444-020-09839-2

线性反应扩散方程的时间变步长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: 虽然时间变步长的两步向后差分公式(BDF2)在模拟多尺度动力学具有重要的价值和广泛的应用, 但其稳定性和收敛性分析仍不完整. 在本工作中, 我们重新讨论了线性反应扩散问题的BDF2格式. 利用[11]中离散正交卷积(DOC)核的技巧, 引入离散互补卷积(DCC)核的概念, 我们证明了在相邻时间步长比条件$0 < r_k := \tau_k/\tau_{k-1} \leq r_{\max} \approx 4.8645$下, BDF2格式是无条件稳定的且具有二阶收敛率. 我们的分析表明, 二阶收敛性是最优且鲁棒的. 鲁棒性指对于任意满足$0 < r_k := \tau_k/\tau_{k-1} \leq r_{\max} \approx 4.8645$ 的时间步长, BDF2格式仍保持二阶收敛性, 并不需要额外的时间步长比限制条件. 此外, 我们的分析还表明, 当$0 < rk \leq 4.8645$时, 用BDF1(即Euler格式)计算第一步数值解u^1 不会导致全局二阶精度的损失. 最后, 我们给出了数值例子来佐证本文理论分析.

关键词: BDF2; DOC; DCC; 时间变步长; 最优误差估计

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