Analysis of adaptive BDF2 scheme for diffusion equations

Hong-lin Liao∗ Zhimin Zhang†

December 25, 2019

Abstract

The variable two-step backward differentiation formula (BDF2) is revisited via a new theoretical framework using the positive semi-definiteness of BDF2 convolution kernels and a class of orthogonal convolution kernels. We prove that, if the adjacent time-step ratios
\[ r_k := \frac{\tau_k}{\tau_{k-1}} \leq \frac{(3 + \sqrt{17})}{2} \approx 3.561, \]
the adaptive BDF2 time-stepping scheme for linear reaction-diffusion equations is unconditionally stable and (maybe, first-order) convergent in the \( L^2 \) norm. The second-order temporal convergence can be recovered if almost all of time-step ratios \( r_k \leq 1 + \sqrt{2} \) or some high-order starting scheme is used. Specially, for linear dissipative diffusion problems, the stable BDF2 method preserves both the energy dissipation law (in the \( H^1 \) seminorm) and the \( L^2 \) norm monotonicity at the discrete levels. An example is included to support our analysis.

Keywords: linear diffusion equations, adaptive BDF2 scheme, orthogonal convolution kernels, positive semi-definiteness, stability and convergence

AMS subject classifications: 65M06, 65M12

1 Introduction

Adaptive time-stepping strategies are practically useful in capturing the multi-scale behaviors in many time-dependent differential equations (PDEs). They always require theoretically reliable (stable) time-stepping methods on arbitrary time meshes, or on general setting of time step-size variations. For linear and nonlinear parabolic problems, the rigorous numerical analysis of one-step methods, such as the backward Euler and Crank-Nicolson schemes, may be relatively easy because they involve only one degree (the current step size \( \tau_n \)) of freedom. However, the stability and convergence of multi-step time-stepping approaches with unequal time-steps would be challenging difficult because they always involve multiple degrees (including the current step \( \tau_n \), the previous step \( \tau_{n-1} \) and so on) of freedom.

Due to its strong stability, the variable two-step backward differentiation formula (BDF2) is practically valuable for stiff or differential-algebraic problems \([4,5,7,15,16]\). But the stability and convergence theory remains incomplete so far, see \([1,2,4,9,16]\), even for the simplest linear heat conduction equation \( \partial_t u = \Delta u + f \). In this report, we revisit the BDF2 method from a

∗ORCID 0000-0003-0777-6832; Department of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P. R. China. E-mails: liaohl@nuaa.edu.cn and liaohl@csrc.ac.cn. He is supported by a grant 1008-56S.YAH18037 from NUAA Scientific Research Starting Fund of Introduced Talent.
†Beijing Computational Science Research Center, Beijing 100193, P. R. China; and Department of Mathematics, Wayne State University, Detroit, MI 48202, USA. E-mails: zmzhang@csrc.ac.cn and ag7761@wayne.edu. He is supported partly by the NSFC grant 11871092 and NSAF grant U1530401.
new point of view by using the positive semi-definiteness of BDF2 convolution kernels and a novel concept, namely, discrete orthogonal convolution kernels. Typically, consider the linear reaction-diffusion problem in a bounded convex domain \( \Omega \),

\[
\partial_t u - \varepsilon \Delta u = \kappa(x)u + f(t, x) \quad \text{for } x \in \Omega \text{ and } 0 < t < T,
\]

subject to the Dirichlet boundary condition \( u = 0 \) on the smooth boundary \( \partial \Omega \), and the initial data \( u(0, x) = u_0 \) for \( x \in \Omega \). Assume that the diffusive coefficient \( \varepsilon > 0 \) is a constant and the reaction coefficient \( \kappa(x) \) is smooth but bounded by \( \kappa^* > 0 \).

Choose the (possibly nonuniform) time levels \( 0 = t_0 < t_1 < t_2 < \cdots < t_N = T \) with the time-step \( \tau_k := t_k - t_{k-1} \) for \( 1 \leq k \leq N \), and the maximum step size \( \tau := \max_{1 \leq k \leq N} \tau_k \). For any time sequence \( \{v^n\}_{n=0}^N \), denote \( \nabla_r v^n := v^n - v^{n-1} \) and \( \partial_r v^n := \nabla_r v^n / \tau_n \). For \( k = 1, 2, \) let \( \Pi_{n,k} v \) denote the interpolating polynomial of a function \( v \) over \( k + 1 \) nodes \( t_{n-k}, \ldots, t_{n-1} \) and \( t_n \). Taking \( v^n = v(t_n) \), one can find (for instance, by using the Lagrange interpolation) that the BDF1 formula \( D_1 v^n := (\Pi_{n,1} v)'(t_n) = \nabla_r v^n / \tau_n \) for \( n \geq 1 \), and the BDF2 formula

\[
D_2 v^n := (\Pi_{n,2} v)'(t_n) = \frac{1 + 2r_n}{\tau_n(1 + r_n)} \nabla_r v^n - \frac{r_n^2}{\tau_n(1 + r_n)} \nabla_r v^{n-1} \quad \text{for } n \geq 2,
\]

where the adjacent time-step ratios

\[
r_k := \frac{\tau_k}{\tau_{k-1}} \quad \text{for } 2 \leq k \leq N.
\]

Always, one can use the BDF1 scheme, by defining \( D_2 v^1 := D_1 v^1 \), to compute the first-level solution \( u^1 \) because the two-step BDF2 formula needs two starting values and the BDF1 scheme generates a second-order accurate solution at the first time level. Without losing the generality, we consider only a time-discrete solution, \( u^n(x) \approx u(t_n, x) \) for \( x \in \Omega \), is defined by the following adaptive BDF2 time-stepping scheme

\[
D_2 u^n = \varepsilon \Delta u^n + \kappa u^n + f^n, \quad \text{for } 1 \leq n \leq N
\]

where the initial data \( u^0 = u_0 \) and the exterior force \( f^n(x) = f(t_n, x) \). The weak form of the time-discrete problem (1.3) reads

\[
\langle D_2 u^n, w \rangle + \varepsilon \langle \nabla u^n, \nabla w \rangle = \langle \kappa u^n, w \rangle + \langle f^n, w \rangle, \quad \text{for } \forall w \in H^1_0(\Omega) \text{ and } k \geq 1,
\]

where \( \langle u, w \rangle \) denotes the usual inner product in the space \( L^2(\Omega) \). Correspondingly, \( \|\cdot\| \) denotes the associated \( L^2 \) norm and \( |\cdot|_1 \) is the \( H^1 \) seminorm. There exists a positive constant \( C_\Omega \), dependent on the domain \( \Omega \) such that \( \|w\| \leq C_\Omega |w|_1 \) for any \( w \in L^2(\Omega) \cap H^1_0(\Omega) \).

Our numerical analysis begins with a new perspective, that is, the BDF2 formula (1.2) is regarded as a discrete convolution summation,

\[
D_2 v^n = \sum_{k=1}^{n} b_{n-k}^{(n)} \nabla_r v^k \quad \text{for } n \geq 1
\]

where the discrete convolution kernels \( b_{n-k}^{(n)} \) are defined by \( b_0^{(1)} := 1 / \tau_1 \), and when \( n \geq 2 \),

\[
b_0^{(n)} := \frac{1 + 2r_n}{\tau_n(1 + r_n)}, \quad b_1^{(n)} := -\frac{r_n^2}{\tau_n(1 + r_n)} \quad \text{and} \quad b_j^{(n)} := 0 \quad \text{for } 2 \leq j \leq n - 1.
\]
To establish the $L^2$ norm stability and convergence, we introduce a new concept, namely, discrete orthogonal convolution (DOC) kernels $\{\theta^{(n)}_{n-k}\}_{k=1}^n$, by a recursive procedure

$$
\theta^{(n)}_0 := \frac{1}{b^{(n)}_0} \quad \text{and} \quad \theta^{(n)}_{n-k} := -\frac{1}{b^{(k)}_0} \sum_{j=k+1}^n \theta^{(n)}_{n-j} b^{(j)}_{j-k} \quad \text{for } 1 \leq k \leq n-1.
$$

(1.7)

Here and hereafter, assume the summation $\sum_{k=i}^n b^{(k)}_0$ to be zero and the product $\prod_{k=i}^n b^{(k)}_0$ to be one if the index $i > j$. Obviously, the DOC kernels $\theta^{(n)}_{n-j}$ satisfies the discrete orthogonal identity

$$
\sum_{j=k}^n \theta^{(n)}_{n-j} b^{(j)}_{j-k} \equiv \delta_{nk} \quad \text{for } \forall 1 \leq k \leq n,
$$

(1.8)

where $\delta_{nk}$ is the Kronecker delta symbol with $\delta_{nk} = 0$ if $k \neq n$. It is to note that the positive semi-definiteness of BDF2 convolution kernels $b^{(n)}_{n-k}$ and the corresponding DOC kernels $\theta^{(n)}_{n-k}$, see Lemmas 2.1 and 2.2, plays an important role in our numerical analysis.

To make our arguments more clearly, we consider firstly a simple case. Next section focuses on the linear dissipative parabolic equations (1.1) with the reaction coefficient $\kappa(x) \leq 0$. In the numerical analysis of (1.3) with $\kappa = 0$, Becker [1] proved that, if $0 < r_k \leq \frac{2+\sqrt{5}}{4} \approx 1.686$, the discrete solution fulfills, also see the Thomée’s classical book [16, Lemma 10.6],

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

where the quantity $\Gamma_n := \sum_{k=2}^{n-2} \max \{0, r_k - r_{k+2}\}$ and $C > 0$ are dependent on the sequence of step size ratios $r_k$. To our knowledge, this type (but more restrictive) quantity was firstly introduced by Le Roux [3] nearly forty years ago; but the quantity $\Gamma_n$ may takes the values of zero, bounded [16, p.175] or unbounded [2, Remark 4.1] at vanishing step sizes by choosing certain step-ratio sequences $\{r_k\}$. The $L^2$ norm stability estimate by Emmrich [4] continues to retain the undesirable prefactor $\exp \left(\Gamma_n t\right)$ but improves slightly the Becker’s restriction to $0 < r_k \leq 1.91$. More recently, with the help of a generalized discrete Grönwall inequality, Chen et al [2] improved the Becker’s estimate by replacing the prefactor $\exp \left(\Gamma_n t\right)$ with $\exp \left(\Gamma_n t_{\text{b}}\right)$ but introduced a stronger step-ratio restriction $0 < r_k \leq 1.53$. This estimate avoids the worst case of $\Gamma_n$ being unbounded, but may lose some other approximately ideal situations with $\Gamma_n = 0$.

Note that, the solution of (1.1) with $\kappa(x) = 0$ satisfies an energy dissipation law

$$
\frac{d}{dt} \left( \varepsilon \|u(t)\|^2_1 + \langle -\kappa u(t), u(t) \rangle \right) \leq 2 \langle f(t), \partial_t u(t) \rangle \quad \text{for } \kappa \leq 0 \text{ and } t > 0,
$$

(1.9)

and the following $L^2$ norm estimate

$$
\|u(t)\| \leq \|u_0\| + 2 \int_0^t \|f(s)\| \, ds \quad \text{for } t \geq 0.
$$

(1.10)

To the best of our knowledge, the existing $L^2$ norm estimates for the adaptive BDF2 scheme [1.3] always require certain discrete Grönwall-Bellman type inequalities. In section 2, we show in Theorem 2.1 that the variable BDF2 method has a discrete energy dissipation law, simulating (1.9) at the discrete time levels, when the adjacent time-step ratios $r_k$ satisfy a sufficient condition.
S1. $0 < r_k \leq (3 + \sqrt{17})/2 \approx 3.561$ for $2 \leq k \leq N$.

Then, with this condition S1, Theorem 2.2 establishes a novel $L^2$ norm estimate

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

which perfectly simulates the continuous estimate (1.10). So the adaptive BDF2 time-stepping method (1.3) is monotonicity-preserving (in the sense of [8]), unconditionally stable and (maybe, first-order) convergent in the $L^2$ norm. It is interesting to mention that, the step-ratio condition S1 ensures the A-stability and L-stability of adaptive BDF2 method considering the linear ODE model equation $y' = \lambda y$ with $\Re(\lambda) \leq 0$, see Remark 3.

In Section 3, the stability and convergence of adaptive BDF2 method (1.3) is established for the linear diffusion equations (1.1) with a bounded coefficient $\kappa(x)$. Theorems 3.1 gives

$$
\|u^n\| \leq 2 \exp(4\kappa^*t_{n-1}) \left(\|u^0\| + 2 \sum_{k=1}^{n} \sum_{j=1}^{k} \theta_{k-j}^{(k)}\|f^j\|\right) \quad \text{for} \quad 1 \leq n \leq N,
$$

and thus the numerical solution is unconditionally stable with respect to the $L^2$ norm under the condition S1. As a by-product of the proof for Theorems 3.1 we find that the condition S1 also ensures the zero-stability, see Remark 5 by considering the nonlinear ODE problem $y' = g(t, y)$ with the Lipschitz-continuous perturbations. Note that, our stability condition S1 updates the classical stability restriction, $0 < r_k < 1 + \sqrt{2}$, given by Grigorieff [6] nearly forty years ago, also see [3] and the classical book [7, Section III.5] by Hairer et al.

When the solution varies slowly, we can use the uniform and quasi-uniform meshes to capture the numerical behavior, and the adjacent step ratios $r_k$ are always close to 1. The restriction of adjacent step ratios often takes its effect in the fast-varying (high gradient) domains, and in the transition regions between the slow-varying and fast-varying domains. In the “slow-to-fast” transition regions and fast-varying domains, we need a reduction of time-steps with the adjacent step ratios $r_k \in (0, 1)$, which is covered by the condition S1. Actually, the restriction S1 limits only the amplification of time-steps in the “fast-to-slow” transition regions. S1 says that one can use a series of increasing time-steps with the amplification factor up to 3.561. Nonetheless, very large time-steps always lead to a loss of numerical precisions and thus large amplification factors would be rarely used continuously in practical simulations. So it is reasonable to assume that the size of $\Re_p$ is very small, where $\Re_p$ is an index set

$$
\Re_p := \left\{ k \mid 1 + \sqrt{2} \leq r_k \leq (3 + \sqrt{17})/2 \right\}.
$$

Then we prove in Theorem 3.2 that the adaptive BDF2 method (1.3) is second-order convergent in the $L^2$ norm under the following step-ratio condition

S2. The step ratios $r_k$ are contained in S1, but $|\Re_p| = N_0 \ll N$.

This condition seems more theoretically rather than practically. Always, potential users are suggested to choose all of time-step ratios $r_k \in (0, 1 + \sqrt{2})$, the Grigorieff’s stability restriction, with $N_0 = 0$ for the second-order accurate computations; however, the condition S2 provides certain redundancy for practical choices of time-steps in self-adaptive numerical simulations. Numerical tests using random time meshes are presented in Section 4 to support our theoretical results. We end this article by presenting some concluding remarks in the last section.
2 Stability analysis for dissipative diffusion problem

2.1 Positive semi-definiteness and energy stability

We describe firstly a sufficient condition on the adjacent time-step ratios $r_k$ so that the discrete convolution kernels $b^{(n)}_{n-k}$ are positive semi-definite, which will be essential to the stability and convergence of BDF2 time-stepping scheme. We consider only certain restriction of each step ratio here. However, in the adaptive time-stepping process, one can choose the next time-step size $\tau_{m+1}$ (or the step ratio $r_{m+1}$) properly according to the information from previous time-step ratios $\{r_k\}_{k=2}^m$, see more comments in Remark 1.

Lemma 2.1. Assume that the adjacent step ratios $r_k$ satisfy S1. For any real sequence $\{w_k\}_{k=1}^n$ with $n$ entries, it holds that

$$2w_k \sum_{j=1}^k b^{(k)}_{k-j} w_j \geq \frac{r_{k+1} w^2_k}{1 + r_{k+1} \tau_k} - \frac{r_k w^2}{1 + r_k \tau_k}$$

for $k \geq 2$.

So the discrete convolution kernels $b^{(n)}_{n-k}$ defined in [10] are positive semi-definite,

$$\sum_{k=1}^n w_k \sum_{j=1}^k b^{(k)}_{k-j} w_j \geq 0 \quad \text{for } n \geq 1.$$

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

$$2w_k \sum_{j=1}^k b^{(k)}_{k-j} w_j = 2b^{(k)}_0 w^2_k + 2b^{(k)}_1 w_k w_{k-1} \geq \left(2b^{(k)}_0 + b^{(k)}_1\right) w^2_k + b^{(k)}_1 w^2_{k-1}$$

$$= \frac{2 + 4r_k - r^2_k}{\tau_k (1 + r_k)} w^2_k - \frac{r^2_k}{\tau_k (1 + r_k)} w^2_{k-1} = \frac{2 + 4r_k - r^2_k}{1 + r_k} \frac{w^2_k}{\tau_k} - \frac{r_k}{1 + r_k} \frac{w^2_{k-1}}{\tau_{k-1}}$$

$$= \frac{r_{k+1} w^2_k}{1 + r_{k+1} \tau_k} - \frac{r_k}{1 + r_k} \frac{w^2_{k-1}}{\tau_{k-1}} + \left(\frac{2 + 4r_k - r^2_k}{1 + r_k} - \frac{r_{k+1}}{1 + r_{k+1}}\right) \frac{w^2_k}{\tau_k}$$

for $k \geq 2$. If the time-step ratios $0 < r_k \leq r_s$, where $r_s = \frac{3 + \sqrt{17}}{2}$ is the positive root of the equation $2 + 3r_s - r^2_s = 0$, then the following inequality holds

$$\frac{2 + 4r_k - r^2_k}{1 + r_k} \geq \frac{r_s}{1 + r_k} \geq \frac{r_{k+1}}{1 + r_{k+1}} \quad \text{for } k \geq 2.$$

Actually, denote $h(x) := \frac{2 + 4x - x^2}{1 + x}$ and then $h'(x) = (1 + x)^{-2}(x + 1 + \sqrt{3})(\sqrt{3} - 1 - x)$. Consider two cases: (i) If $0 < x \leq \sqrt{3} - 1$, then $h'(x) \geq 0$. So $h(r_k) \geq h(0) = 2 > \frac{r_s}{1 + r_s}$. (ii) If $\sqrt{3} - 1 < x \leq r_s$, then $h'(x) \leq 0$. So $h(r_k) \geq h(r_s) = \frac{r_s}{1 + r_s}$. Thus it follows that

$$2w_k \sum_{j=1}^k b^{(k)}_{k-j} w_j \geq \frac{r_{k+1} w^2_k}{1 + r_{k+1} \tau_k} - \frac{r_k}{1 + r_k} \frac{w^2_{k-1}}{\tau_{k-1}} \quad \text{for } k \geq 2.$$
Therefore, we have
\[ 2 \sum_{k=1}^{n} w_k \sum_{j=1}^{k} b_{k-j}^{(n)} w_j \geq \frac{2}{\tau_1} w_1^2 + \frac{r_{n+1}}{1 + r_{n+1} \tau_n} w_n^2 - \frac{r_2}{1 + r_2 \tau_1} w_1^2 \]
\[ \geq \frac{r_{n+1}}{1 + r_{n+1} \tau_n} w_n^2 + \frac{2 + r_2 w_1^2}{1 + r_2 \tau_1} \geq 0 \text{ for } n \geq 1. \]

It completes the proof. \[ \square \]

**Remark 1.** Numerical tests on random time meshes in Section 4 suggest that the step ratio condition \( S_1 \) is not necessary. While, in mathematical manner, the condition \( S_1 \) is also not necessary since the positive semi-definiteness of discrete convolution kernels \( b_{n-k}^{(n)} \) in [1.6] should be determined by the eigenvalues of the following tridiagonal symmetric matrix
\[
B_2 := \begin{pmatrix}
2b_0^{(1)} & b_1^{(2)} & & & \\
b_1^{(2)} & 2b_0^{(2)} & b_1^{(3)} & & \\
 & \ddots & \ddots & \ddots & \\
 & & b_1^{(n-1)} & 2b_0^{(n-1)} & b_1^{(n)} \\
 & & & b_1^{(n)} & 2b_0^{(n)}
\end{pmatrix}.
\]

It is seen that, some weaker (maybe, sufficient and necessary) condition for the positive semi-definiteness of the matrix \( B_2 \) would be a certain combination involving all time-step ratios; however, it is open to us up to now.

We now consider the energy (\( H^1 \) seminorm) stability of BDF2 scheme (1.3) by defining a (modified) discrete energy \( E^k \),
\[
E^k := \frac{r_{k+1}}{1 + r_{k+1}} \| \partial_\tau u^k \|^2 + \varepsilon |u^k|^2 + \langle -\kappa u^k, u^k \rangle \text{ for } \kappa \leq 0 \text{ and } k \geq 1,
\]
(2.1)
together with the initial energy \( E^0 := \varepsilon |u^0|^2 + \langle -\kappa u^0, u^0 \rangle \).

**Theorem 2.1.** Under the condition \( S_1 \), the discrete solution \( u^n \) of the BDF2 time-stepping scheme (1.3) with \( \kappa \leq 0 \) satisfies
\[
\partial_\tau E^k \leq 2 \langle f^k, \partial_\tau u^k \rangle, \text{ for } k \geq 1,
\]
(2.2)
which simulates the energy dissipation law (1.9) numerically. So the discrete solution is unconditionally stable in the energy norm,
\[
\sqrt{E^n} \leq \sqrt{E^0} + 4\varepsilon^{-\frac{1}{2}} C_\Omega \left( \| f^1 \| + \sum_{k=2}^{n} \| \nabla_\tau f^k \| \right) \text{ for } n \geq 1.
\]

**Proof.** Taking \( w = 2\nabla_\tau u^k \) in the weak form (1.4) for \( k \geq 2 \), we have
\[
2 \langle D_2 u^k, \nabla_\tau u^k \rangle + 2\varepsilon \langle \nabla u^k, \nabla_\tau \nabla u^k \rangle + 2\langle -\kappa u^k, \nabla_\tau u^k \rangle = 2 \langle f^k, \nabla_\tau u^k \rangle, \text{ for } k \geq 2.
\]
Lemma [2.1] gives
\[
2 \langle D_2 u^k, \nabla_E u^k \rangle \geq \frac{r_{k+1}}{1 + r_{k+1}} \tau_k \| \partial_T u^k \|^2 - \frac{r_k}{1 + r_k} \tau_{k-1} \| \partial_T u^{k-1} \|^2 \quad \text{for } k \geq 2.
\]

With the help of the inequality \(2a(a - b) \geq a^2 - b^2\), it is easy to obtain that
\[
\nabla_E E^k \leq 2 \langle f^k, \nabla_E u^k \rangle, \quad \text{for } k \geq 2.
\]

(2.3)

Also, by taking \(w = 2 \nabla_E u^1\) in (1.3) for the case \(k = 1\), we get
\[
2 r_1 \| \partial_T u^1 \|^2 + \varepsilon |u^1|^2 + \langle - \kappa u^1, u^1 \rangle \leq \varepsilon |u^0|^2 + \langle - \kappa u^0, u^0 \rangle + 2 \langle f^1, \nabla_E u^1 \rangle,
\]

which implies that
\[
\nabla_E E^1 \leq 2 \langle f^1, \nabla_E u^1 \rangle.
\]

Combining it with the general case (2.3), one gets the discrete energy dissipation law (2.2). Summing the inequality (2.2) from \(k = 1\) to \(n\), we have
\[
E^n \leq E^0 + 2 \sum_{k=1}^n \langle f^k, \nabla_E u^k \rangle \quad \text{for } n \geq 1.
\]

(2.4)

By applying the technique of time summation by parts (cf. [10 Lemma 2.6]) and the Cauchy-Schwarz inequality, we obtain
\[
\sum_{k=1}^n \langle f^k, \nabla_E u^k \rangle = \langle f^n, u^n \rangle - \sum_{k=2}^n \langle \nabla_E f^k, u^{k-1} \rangle - \langle f^1, u^0 \rangle
\]
\[
\leq \| u^n \| \| f^n \| + \sum_{k=2}^n \| u^{k-1} \| \| \nabla_E f^k \| + \| u^0 \| \| f^1 \|
\]
\[
\leq \varepsilon^{-\frac{1}{2}} C_\Omega \left( \sqrt{E^n} \| f^n \| + \sum_{k=2}^n \sqrt{E^{k-1}} \| \nabla_E f^k \| + \sqrt{E^0} \| f^1 \| \right) \quad \text{for } n \geq 1,
\]

where the Poincaré inequality has been used. It follows from (2.4) that
\[
E^n \leq E^0 + 2 \varepsilon^{-\frac{1}{2}} C_\Omega \left( \sqrt{E^n} \| f^n \| + \sum_{k=2}^n \sqrt{E^{k-1}} \| \nabla_E f^k \| + \sqrt{E^0} \| f^1 \| \right) \quad \text{for } n \geq 1.
\]

For any finite \(n\), choose \(n_0 \) (\(0 \leq n_0 \leq n\)) so that \(E^{n_0} = \max_{0 \leq j \leq n} E^j\). One can take \(n = n_0\) in the above inequality and apply the triangle inequality to obtain
\[
E^{n_0} \leq \sqrt{E^n} \sqrt{E^{n_0}} + 2 \varepsilon^{-\frac{1}{2}} C_\Omega \sqrt{E^{n_0}} \left( \| f^{n_0} \| + \sum_{k=2}^{n_0} \| \nabla_E f^k \| + \| f^1 \| \right)
\]
\[
\leq \sqrt{E^n} \sqrt{E^{n_0}} + 4 \varepsilon^{-\frac{1}{2}} C_\Omega \sqrt{E^{n_0}} \left( \| f^1 \| + \sum_{k=2}^{n_0} \| \nabla_E f^k \| \right)
\]

7
because $f^{n_0} = f^1 + \sum_{k=2}^{n_0} \nabla_f f^k$. Thus it follows that
\[
\sqrt{E^n} \leq \sqrt{E^{n_0}} \leq \sqrt{E^0} + 4e^{-\frac{n}{2}} C_\Omega \left( \|f^1\| + \sum_{k=2}^{n_0} \|\nabla_f f^k\| \right) \\
\leq \sqrt{E^0} + 4e^{-\frac{n}{2}} C_\Omega \left( \|f^1\| + \sum_{k=2}^{n} \|\nabla_f f^k\| \right) \quad \text{for } n \geq 1.
\]

It yields the claimed estimate and completes the proof.

If the exterior force $f(x, t)$ is zero-valued, the discrete energy law (2.2) gives
\[
E^k \leq E^{k-1} \quad \text{for } k \geq 1,
\]
so that the variable-step BDF2 scheme (1.3) preserves the energy dissipation law at the discrete levels. This property would be important in simulating the gradient flow problems, cf. [2, 12, 13] and the references therein. However, the energy estimate in Theorem 2.1 always leads to a suboptimal $H^1$ seminorm error estimate with respect to both the temporal accuracy and the dependence of the diffusive coefficient $\varepsilon$ (especially when $\varepsilon$ is small).

**Remark 2.** From the computational view of point, the discrete energy form $E^k$ in (2.1) suggests that small time-steps (with small step ratios) are necessary to capture the solution behaviors when $\|\partial_t u\|$ becomes large, and large time-steps (with some big step ratios) are acceptable to accelerate the time integration when $\|\partial_t u\|$ is small.

### 2.2 Orthogonal convolution kernels and $L^2$ norm stability

**Lemma 2.2.** If the BDF2 kernels $b^{(n)}_{n-k}$ in (1.6) are positive semi-definite, the DOC kernels $\theta^{(n)}_{n-j}$ defined in (1.7) are also positive semi-definite. For any real sequence $\{w_j\}_{j=1}^{n}$, it holds that
\[
\sum_{k=1}^{n} w_k \sum_{j=1}^{k} \theta^{(k)}_{k-j} w_j \geq 0 \quad \text{for } n \geq 1.
\]

**Proof.** Given any real sequence $\{w_j\}_{j=1}^{n}$, one applies the discrete kernels in (1.6) to define another sequence $\{V_j\}_{k=1}^{n}$ by
\[
V_j = -\frac{1}{b_0^{(j)}} \sum_{\ell=1}^{j-1} b^{(j)}_{j-\ell} V_\ell + \frac{w_j}{b_0^{(j)}} \quad \text{for } j \geq 1,
\]
or
\[
w_j = \sum_{\ell=1}^{j} b^{(j)}_{j-\ell} V_\ell \quad \text{for } j \geq 1. \tag{2.5}
\]

Multiplying both sides of the above equality (2.5) by the DOC kernels $\theta^{(k)}_{k-j}$, and summing $j$ from 1 to $k$, we find
\[
\sum_{j=1}^{k} \theta^{(k)}_{k-j} w_j = \sum_{j=1}^{k} \theta^{(k)}_{k-j} \sum_{\ell=1}^{j} b^{(j)}_{j-\ell} V_\ell = \sum_{\ell=1}^{k} V_\ell \sum_{j=\ell}^{k} \theta^{(k)}_{k-j} b^{(j)}_{j-\ell} = V_k \quad \text{for } k \geq 1, \tag{2.6}
\]
where the summation order has been exchanged in the second equality and the orthogonal identity (1.8) was used in the third one. Thus it follows from (2.5)-(2.6) that
\[
\sum_{k=1}^{n} w_k \sum_{j=1}^{k} \theta_{k-j}^{(k)} w_j = \sum_{k=1}^{n} V_k \sum_{\ell=1}^{k} b_{k-\ell}^{(k)} V_\ell \geq 0 \quad \text{for } n \geq 1,
\]
because the BDF2 kernels \(b_{n-k}^{(n)}\) are positive semi-definite. The proof is completed. \(\Box\)

**Corollary 2.1.** The DOC kernels \(\theta_{n-j}^{(n)}\) in (1.7) fulfill
\[
\sum_{j=1}^{n} \theta_{n-j}^{(n)} \equiv \tau_n \quad \text{such that} \quad \sum_{k=1}^{n} \sum_{j=1}^{k} \theta_{k-j}^{(k)} \equiv t_n \quad \text{for } n \geq 1.
\]

**Proof.** This proof is similar to that of Lemma 2.2. Taking \(v^n = t_n\) in (1.5), one can find that
\[
1 \equiv \sum_{\ell=1}^{j} b_{j-\ell}^{(j)} \tau_\ell \quad \text{for } j \geq 1.
\]
Multiplying both sides of the above equality by the DOC kernels \(\theta_{n-j}^{(n)}\), and summing \(j\) from 1 to \(n\), we apply the orthogonal identity (1.8) to find
\[
\sum_{j=1}^{n} \theta_{n-j}^{(n)} \equiv \sum_{j=1}^{n} \theta_{n-j}^{(n)} \sum_{\ell=1}^{j} b_{j-\ell}^{(j)} \tau_\ell = \sum_{\ell=1}^{n} \tau_\ell \sum_{j=1}^{n} \theta_{n-j}^{(n)} b_{j-\ell}^{(j)} = \tau_n \quad \text{for } n \geq 1,
\]
as desired. The proof is complete. \(\Box\)

**Lemma 2.3.** The DOC kernels \(\theta_{n-j}^{(n)}\) in (1.7) have an explicit formula
\[
\theta_{n-k}^{(n)} = \frac{1}{b_0^{(k)}} \prod_{i=k+1}^{n} \frac{r_i^2}{1 + 2r_i} = \frac{\tau_n}{b_0^{(k)} \tau_k} \prod_{i=k+1}^{n} \frac{r_i}{1 + 2r_i} \quad \text{for } 1 \leq k \leq n.
\]

**Proof.** Denote \(\hat{\theta}_{n-k}^{(n)} := \theta_{n-k}^{(n)} b_0^{(k)}\) for \(1 \leq k \leq n\). For \(n = 1\), the definition (1.7) yields \(\hat{\theta}_{0}^{(1)} = 1\). For the index \(n \geq 2\), we use the definition (1.7) and the BDF2 convolution kernels in (1.6) to find the following recursive procedure
\[
\hat{\theta}_{0}^{(n)} = 1 \quad \text{and} \quad \hat{\theta}_{n-k}^{(n)} = \frac{b_{k+1}^{(k+1)}}{b_0^{(k+1)}} \hat{\theta}_{n-k-1}^{(n)} \quad \text{for } 1 \leq k \leq n - 1.
\]

Thus a simple induction yields
\[
\hat{\theta}_{n-k}^{(n)} = \prod_{i=k+1}^{n} \frac{r_i^2}{1 + 2r_i} > 0 \quad \text{for } 1 \leq k \leq n.
\]
It yields the claimed formula and completes the proof. \(\Box\)

Now we establish the \(L^2\) norm stability of the BDF2 scheme (1.3) for the case \(\kappa \leq 0\).
Theorem 2.2. If the BDF2 kernels \( b_{n-k}(n) \) in (1.3) are positive semi-definite (or the sufficient condition S1 holds), the discrete solution \( u^n \) of the adaptive BDF2 scheme (1.3) with the reaction coefficient \( \kappa \leq 0 \) is unconditionally stable in the \( L^2 \) norm,

\[
\|u^n\| \leq \|u^0\| + 2 \sum_{k=1}^{n} \sum_{j=1}^{k} \theta_{k-j}^{(k)} \|f^j\| \leq \|u^0\| + 2t_n \max_{1 \leq j \leq n} \|f^j\| \quad \text{for } n \geq 1.
\]

Thus the BDF2 scheme (1.3) is monotonicity-preserving (taking \( f \equiv 0 \)) according to [8].

Proof. Multiplying both sides of the equation (1.3) by the DOC kernels \( \theta_{k-n}^{(k)} \), and summing \( n \) from 1 to \( k \), we find

\[
\sum_{j=1}^{k} \theta_{k-j}^{(k)} D_2 u^j = \sum_{j=1}^{k} \theta_{k-j}^{(k)} (\varepsilon \Delta + \kappa) u^j + \sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j \quad \text{for } k \geq 1.
\]

Applying the orthogonal identity (1.8), one has

\[
\sum_{j=1}^{k} \theta_{k-j}^{(k)} D_2 u^j = \sum_{j=1}^{k} \theta_{k-j}^{(k)} (\varepsilon \Delta + \kappa) u^j + \sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j \quad \text{for } k \geq 1,
\]

where the summation order has been exchanged in the second equality. So we have

\[
\nabla \tau u^k = \sum_{j=1}^{k} \theta_{k-j}^{(k)} (\varepsilon \Delta + \kappa) u^j + \sum_{j=1}^{k} \theta_{k-j}^{(k)} f^j \quad \text{for } k \geq 1. \tag{2.9}
\]

Making the inner product of the equation (2.9) with \( u^k \), and summing the resulting equality from \( k = 1 \) to \( n \), one has

\[
\sum_{k=1}^{n} \langle u^k, \nabla \tau u^k \rangle = \sum_{k=1}^{n} \sum_{j=1}^{k} \langle u^k, \theta_{k-j}^{(k)} (\varepsilon \Delta + \kappa) u^j \rangle + \sum_{k=1}^{n} \sum_{j=1}^{k} \langle u^k, \theta_{k-j}^{(k)} f^j \rangle \leq \sum_{k=1}^{n} \sum_{j=1}^{k} \langle u^k, \theta_{k-j}^{(k)} f^j \rangle \quad \text{for } n \geq 1,
\]

where the following inequality derived by Lemma 2.2 has been used,

\[
\sum_{k=1}^{n} \sum_{j=1}^{k} \langle u^k, \theta_{k-j}^{(k)} (\varepsilon \Delta + \kappa) u^j \rangle = \sum_{k=1}^{n} \sum_{j=1}^{k} \left[ - \varepsilon \langle \nabla u^k, \theta_{k-j}^{(k)} \nabla u^j \rangle + \langle \kappa u^k, \theta_{k-j}^{(k)} u^j \rangle \right] \leq 0.
\]

Note that, Lemma 2.3 shows that \( \theta_{k-j}^{(k)} > 0 \). Then the Cauchy-Schwarz inequality yields

\[
\|u^n\|^2 + \sum_{k=1}^{n} \|\nabla \tau u^k\|^2 \leq \|u^0\|^2 + 2 \sum_{k=1}^{n} \|u^k\| \sum_{j=1}^{k} \theta_{k-j}^{(k)} \|f^j\| \quad \text{for } n \geq 1.
\]
Taking some integer \( n_1 \) \((0 \leq n_1 \leq n)\) such that \( \|u^{n_1}\| = \max_{0 \leq k \leq n} \|u^k\| \). Taking \( n := n_1 \) in the above inequality, one gets

\[
\|u^{n_1}\|^2 \leq \|u^0\| \|u^{n_1}\| + 2 \|u^{n_1}\| \sum_{k=1}^{n_1} \sum_{j=1}^k \theta_{k-j} \|f^j\|,
\]

and thus

\[
\|u^n\| \leq \|u^0\| + 2 \sum_{k=1}^{n_1} \sum_{j=1}^k \theta_{k-j} \|f^j\| \leq \|u^0\| + 2 \sum_{k=1}^{n} \sum_{j=1}^k \theta_{k-j} \|f^j\| \quad \text{for} \quad n \geq 1.
\]

The claimed second estimate follows from Corollary 2.4 immediately.

On the uniform mesh with \( r_k \equiv 1 \), Lemma 2.3 yields

\[
\theta_{k-1} = \frac{\tau}{3k-1} \quad \text{and} \quad \theta_{k-j} = \frac{2\tau}{3k-j+1} \quad \text{for} \quad 2 \leq j \leq k.
\]

So Theorem 2.2 gives (by exchanging the summation order)

\[
\|u^n\| \leq \|u^0\| + 2 \|f^1\| \sum_{k=1}^n \theta_{k-1} + 2 \sum_{j=2}^n \|f^j\| \sum_{k=j}^n \theta_{k-j}
\]

\[
= \|u^0\| + 3\tau \left(1 - \frac{1}{3n}\right) \|f^1\| + 2\tau \sum_{j=2}^n \left(1 - \frac{1}{3n-j+1}\right) \|f^j\| \quad \text{for} \quad n \geq 1,
\]

which recovers our previous result in [11] Lemma 3.2] with a slightly different constant for the term \( \|f^1\| \). Also, it directly leads to the estimate (1.61) in [16] Theorem 1.7. On the other hand, Theorem 2.2 recovers the solution estimate (1.10) under a mild restriction \( S1 \), and essentially improves the existing \( L^2 \) norm estimates including the classical one [16] Lemma 10.6]. Also, no any discrete Grönwall inequalities have been used in our \( L^2 \) norm estimate and no any restrictions of maximum time-step size are required.

**Remark 3.** Consider the BDF2 scheme \( D_\tau y^n = \lambda y^n \) for solving the ODE model \( y' = \lambda y \) with \( \Re(\lambda) \leq 0 \). Reminding the inequality \( 2\Re(y^k \nabla r y^k) \geq |y^k|^2 - |y^{k-1}|^2 \), one can follow the proof of Theorem 2.2 to obtain \( \|y^n\| \leq \|y^0\| \) for \( n \geq 1 \). So the adaptive BDF2 scheme is A-stable under the step ratio condition \( S1 \). Obviously, it is also L-stable considering the limit \( \lambda \tau_n \to -\infty \).

**Remark 4.** Our analysis is fit for any other starting schemes although the BDF1 scheme is applied here to compute the first-level solution. To see more clear, multiplying the equation (1.3) by the DOC kernels \( \theta_{k-n}^{(k)} \) and summing \( n \) from 2 to \( k \), we have

\[
\sum_{j=2}^k \theta_{k-j}^{(k)} D_\tau u^j = \sum_{j=2}^k \theta_{k-j}^{(k)} (\varepsilon \Delta + \kappa) u^j + \sum_{j=2}^k \theta_{k-j}^{(k)} f^j \quad \text{for} \quad k \geq 2.
\]
Applying the orthogonal identity (1.8), one has
\[ \sum_{j=2}^{k} \theta_{k-j}^{(k)} b_{2}^{(2)} \sum_{j=2}^{k} \theta_{k-j}^{(k)} b_{1}^{(2)} \nabla_{\tau} u^{1} = \nabla_{\tau} u^{k} + \theta_{k-2}^{(k)} b_{1}^{(2)} \nabla_{\tau} u^{1} \text{ for } k \geq 2, \]
where the summation order has been exchanged in the second equality. So we have
\[ \nabla_{\tau} u^{k} = -\theta_{k-2}^{(k)} b_{1}^{(2)} \nabla_{\tau} u^{1} + \sum_{j=2}^{k} \theta_{k-j}^{(k)} (\varepsilon \Delta + \kappa) u^{j} + \sum_{j=2}^{k} \theta_{k-j}^{(k)} f^{j} \text{ for } k \geq 2. \]
Making the inner product of this equation with \( u^{k} \), and summing the resulting equality from \( k = 2 \) to \( n \), one applies Lemma 2.2 and the Cauchy-Schwarz inequality to find
\[ \| u^{n} \|^{2} - \| u^{1} \|^{2} \leq -2 \sum_{k=2}^{n} \theta_{k-2}^{(k)} b_{1}^{(2)} \langle u^{k}, \nabla_{\tau} u^{1} \rangle + 2 \sum_{k=2}^{n} \sum_{j=2}^{k} \langle u^{k}, \theta_{k-j}^{(k)} f^{j} \rangle \]
\[ \leq -2 b_{1}^{(2)} \| \nabla_{\tau} u^{1} \| \sum_{k=2}^{n} \theta_{k-2}^{(k)} \| u^{k} \| + 2 \sum_{k=2}^{n} \| u^{k} \| \sum_{j=2}^{k} \theta_{k-j}^{(k)} \| f^{j} \| \text{ for } n \geq 2. \]
By taking \( \| u^{n} \| = \max_{1 \leq k \leq n} \| u^{k} \| \), it is easy to get
\[ \| u^{n} \| \leq \| u^{n} \| \leq \| u^{1} \| - 2 b_{1}^{(2)} \| \nabla_{\tau} u^{1} \| \sum_{k=2}^{n} \theta_{k-2}^{(k)} + 2 \sum_{k=2}^{n} \sum_{j=2}^{k} \theta_{k-j}^{(k)} \| f^{j} \| \text{ for } n \geq 2. \]
From the recursive relationship (2.7), with the auxiliary discrete kernels \( \hat{\theta}_{k-1}^{(k)} \) in (2.8), one can obtain that \(-b_{1}^{(2)} \hat{\theta}_{k-2}^{(k)} = \frac{r_{1}^{i}}{1+2r_{1}} \hat{\theta}_{k-2}^{(k)} = \hat{\theta}_{k-1}^{(k)} \) and
\[ \tau_{1} \sum_{k=1}^{n} \hat{\theta}_{k-1}^{(k)} = \tau_{1} + \sum_{k=2}^{n} \tau_{1} \prod_{i=2}^{k} \frac{r_{1}^{i}}{1+2r_{1}} = \tau_{1} + \sum_{k=2}^{n} \tau_{k} \prod_{i=2}^{k} \frac{r_{1}^{i}}{1+2r_{1}} \leq t_{n} \text{ for } 1 \leq n \leq N. \]
Thus by Corollary 2.2, we arrive at the following corollary.

**Corollary 2.2.** If the BDF2 kernels \( b_{n-k}^{(n)} \) in (1.4) are positive semi-definite (or the sufficient condition \( S1 \) holds), the discrete solution \( u^{n} \) of the adaptive BDF2 scheme (1.3) with the reaction coefficient \( \kappa \leq 0 \) is unconditionally stable in the \( L^{2} \) norm,
\[ \| u^{n} \| \leq \| u^{1} \| + 2 \| \partial_{\tau} u^{1} \| \tau_{1} \sum_{k=2}^{n} \hat{\theta}_{k-1}^{(k)} + 2 \sum_{k=2}^{n} \sum_{j=2}^{k} \theta_{k-j}^{(k)} \| f^{j} \| \]
\[ \leq \| u^{1} \| + 2 t_{n} \| \partial_{\tau} u^{1} \| + 2 t_{n} \max_{2 \leq j \leq n} \| f^{j} \| \text{ for } n \geq 2. \]
Obviously, once the first-level solution \( u^{1} \) and the discrete time derivative \( \partial_{\tau} u^{1} \) are second-order accurate, this estimate yields the second-order accuracy of the adaptive BDF2 scheme under the step-ratio condition \( S1 \), cf. the consistency analysis in subsection 3.2.
3 $L^2$ norm convergence for linear diffusion problems

3.1 Priori estimate

Now consider the $L^2$ norm priori estimate of the adaptive BDF2 scheme (1.3) for a general case $|\kappa(x)| \leq \kappa^*$. This situation always needs a discrete Grönwall inequality.

Lemma 3.1. Let $\lambda \geq 0$, the time sequences $\{\xi_k\}_{k=0}^N$ and $\{V_k\}_{k=1}^N$ be nonnegative. If

$$V_n \leq \lambda \sum_{j=1}^{n-1} \tau_j V_j + \sum_{j=0}^{n} \xi_j \quad \text{for } 1 \leq n \leq N,$$

then it holds that

$$V_n \leq \exp(\lambda t_{n-1}) \sum_{j=0}^{n} \xi_j \quad \text{for } 1 \leq n \leq N.$$

Proof. The proof is standard. Under the induction hypothesis $V_j \leq \exp(\lambda t_{j-1}) \sum_{k=0}^{j} \xi_k$ for $1 \leq j \leq n-1$, the desired inequality for the index $n$ follows directly from

$$\lambda \sum_{j=1}^{n-1} \tau_j \exp(\lambda t_{j-1}) \leq \lambda \int_{0}^{t_{n-1}} \exp(\lambda t) \, dt = \exp(\lambda t_{n-1}) - 1.$$

The principle of induction completes the proof.

Theorem 3.1. If the BDF2 kernels $b_{n-k}^{(a)}$ in (1.6) are positive semi-definite (or the sufficient condition $S1$ holds) and the maximum time-step size $\tau \leq 1/(4\kappa^*)$, the discrete solution $u^n$ of the BDF2 scheme (1.3) is unconditionally stable in the $L^2$ norm,

$$\|u^n\| \leq 2 \exp(4\kappa^* t_{n-1}) \left( \|u^0\| + 2 \sum_{k=1}^{n} \sum_{j=1}^{k} \theta_{k-j}^{(k)} \|f^j\| \right) \quad \text{for } 1 \leq n \leq N. \quad (3.1)$$

Proof. We can start from (2.9) in the proof of Theorem 2.2. Lemma 2.2 implies that

$$\sum_{k=1}^{n} \sum_{j=1}^{k} \langle u^k, \theta_{k-j}^{(k)} \Delta w^j \rangle \leq 0.$$

Then, making the inner product of (2.9) by $u^k$, and summing up the resulting equality from $k = 1$ to $n$, one derives that

$$\sum_{k=1}^{n} \langle u^k, \nabla \varphi \rangle = \sum_{k=1}^{n} \sum_{j=1}^{k} \langle u^k, \theta_{k-j}^{(k)} (\varepsilon \Delta + \kappa) u^j \rangle + \sum_{k=1}^{n} \sum_{j=1}^{k} \langle u^k, \theta_{k-j}^{(k)} f^j \rangle \leq \sum_{k=1}^{n} \sum_{j=1}^{k} \langle u^k, \theta_{k-j}^{(k)} \kappa u^j \rangle + \sum_{k=1}^{n} \sum_{j=1}^{k} \langle u^k, \theta_{k-j}^{(k)} f^j \rangle \quad \text{for } 1 \leq n \leq N.$$
Lemma 2.3 implies that the DOC kernels \( \theta_{k-j}^{(k)} > 0 \). So one applies the Cauchy-Schwarz inequality to find
\[
\| u^n \|^2 \leq \| u^0 \|^2 + 2\kappa^* \sum_{k=1}^{n} \| u^k \| \sum_{j=1}^{k} \theta_{k-j}^{(k)} \| u^j \| + 2 \sum_{k=1}^{n} \| u^k \| \sum_{j=1}^{k} \theta_{k-j}^{(k)} \| f^j \| \quad \text{for } 1 \leq n \leq N.
\]

Choosing some integer \( n_2 (0 \leq n_2 \leq n) \) such that \( \| u^{n_2} \| = \max_{0 \leq k \leq n} \| u^k \| \). Then, taking \( n := n_2 \) in the above inequality, one gets
\[
\| u^{n_2} \|^2 \leq \| u^0 \| \| u^{n_2} \| + 2\kappa^* \| u^{n_2} \| \sum_{k=1}^{n_2} \| u^k \| \sum_{j=1}^{k} \theta_{k-j}^{(k)} + 2 \| u^{n_2} \| \sum_{k=1}^{n_2} \sum_{j=1}^{k} \theta_{k-j}^{(k)} \| f^j \|.
\]

With the help of Corollary 2.1, it follows that
\[
\| u^n \| \leq \| u^{n_2} \| \leq \| u^0 \| + 2\kappa^* \sum_{k=1}^{n_2} \| u^k \| \sum_{j=1}^{k} \theta_{k-j}^{(k)} + 2 \sum_{k=1}^{n_2} \sum_{j=1}^{k} \theta_{k-j}^{(k)} \| f^j \| \leq \| u^0 \| + 2\kappa^* \sum_{k=1}^{n} \| u^k \| \sum_{j=1}^{k} \theta_{k-j}^{(k)} + 2 \sum_{k=1}^{n} \sum_{j=1}^{k} \theta_{k-j}^{(k)} \| f^j \| \quad \text{for } 1 \leq n \leq N.
\]

Setting the maximum time-step size \( \tau \leq 1/(4\kappa^*) \), one has
\[
\| u^n \| \leq 2\kappa^* \sum_{k=1}^{n-1} \| u^k \| + 4 \sum_{k=1}^{n} \sum_{j=1}^{k} \theta_{k-j}^{(k)} \| f^j \| \quad \text{for } 1 \leq n \leq N.
\]

Lemma 3.1 gives the desired estimate (3.1) and completes the proof.

**Remark 5.** Let \( g(t, y) \) be a Lipschitz-continuous nonlinear function with the Lipschitz constant \( L_g > 0 \). Apply the BDF2 time-stepping scheme \( D_2y^n = g(t_n, y^n) \) to the nonlinear ODE model \( y' = g(t, y) \) for \( 0 < t \leq T \). Assuming the perturbed solution \( \tilde{y}^n \) solves \( D_2\tilde{y}^n = g(t_n, \tilde{y}^n) + \varepsilon^n \), we can follow the proof of Theorem 3.1 to obtain
\[
\| y^n - \tilde{y}^n \| \leq 2 \exp (4L_g t_{n-1}) \left( \| y^0 - \tilde{y}^0 \| + 2t_n \max_{1 \leq j \leq n} \| \varepsilon^j \| \right) \quad \text{for } 1 \leq n \leq N.
\]

So the BDF2 scheme is zero-stable under the step-ratio condition S1. It updates the Grigorieff’s stability restriction given in [8], also see the classical book [7, Section III.5] by Hairer et al.

### 3.2 Consistency and convergence

By using Corollary 2.1, the priori estimate (3.1) gives the following estimate
\[
\| u^n \| \leq 2 \exp (4\kappa^* t_{n-1}) \left( \| u^0 \| + 2t_n \max_{1 \leq j \leq n} \| f^j \| \right) \quad \text{for } 1 \leq n \leq N.
\]

This estimate would lead to a loss of time accuracy in error analysis, cf. Theorem 3.3 below, because the BDF1 scheme for the first-level solution \( u^1 \) is only first-order consistent. Also, the step ratio condition S1 requires \( \tau_1/\tau_2 \geq \sqrt{3} \approx 0.281 \), and prevents our use of very small initial
step size, like \( \tau_1 = O(\tau^2) \), to recover the second-order accuracy. In this sense, the analysis and numerical evidences in the note [14] are inadequate, although the suggested second-order singly-diagonal implicit Runge-Kutta method would be reliable to compute the first-level solution.

The loss of accuracy is attributed to the unequal time steps and the associated step ratios because, as well-known, the uniform BDF2 scheme is globally second-order order accurate. Actually, in next lemma, the global convolution term combined with the DOC kernels \( \theta_{k-j}^{(k)} \) in the estimate (3.1) is evaluated carefully. To a certain degree, it reveals the error behavior of BDF2 time-stepping with respect to the unequal time-step sizes.

**Lemma 3.2.** For the consistency error \( \eta^j : = D_2 u(t_j) - \partial_t u(t_j) \) at \( t = t_j \), it holds that

\[
\sum_{k=1}^n \sum_{j=1}^k \theta_{k-j}^{(k)} \eta^j \leq \tau_1 \sum_{k=1}^n \theta_{k-1}^{(k)} \int_0^{t_1} \| \partial_{tt} u \| \, dt + \frac{3}{2} \sum_{j=1}^n \tau_j^2 \sum_{k=j}^n \theta_{k-j}^{(k)} \int_{t_{j-1}}^{t_j} \| \partial_{tt} u \| \, dt
\]

for \( 1 \leq n \leq N \), where the discrete kernels \( \hat{\theta}_{k-j}^{(k)} \) is given by (2.8).

**Proof.** For simplicity, denote

\[
G_{12}^j = \int_{t_{j-1}}^{t_j} \| \partial_{tt} u \| \, dt \quad \text{and} \quad G_{13}^j = \int_{t_{j-1}}^{t_j} \| \partial_{tt} u \| \, dt \quad \text{for} \quad j \geq 1.
\]

For the case of \( j = 1 \), the consistency error is bounded by

\[
\| \eta^1 \| \leq \frac{1}{\tau_1} \int_0^{t_1} \| \partial_t u(t) - \partial_t u(t_1) \| \, dt \leq \frac{1}{\tau_1} \int_0^{t_1} \int_s^{t_1} \| \partial_{tt} u \| \, dt \, ds \leq b_0^{(1)} \tau_1 G_{12}^1.
\]

Then Lemma [2.3] together with the discrete kernels \( \hat{\theta}_{k-j}^{(k)} \) in (2.8) yields

\[
\| \eta^1 \| \sum_{k=1}^n \theta_{k-1}^{(k)} \leq \tau_1 G_{12}^1 b_0^{(1)} \sum_{k=1}^n \theta_{k-1}^{(k)} \leq \tau_1 G_{12}^1 \sum_{k=1}^n \hat{\theta}_{k-1}^{(k)}.
\]

By using the Taylor’s expansion formula, one can derive that, also see [16, Theorem 10.5],

\[
\eta^j = \frac{1 + r_j}{2\tau_j} \int_{t_{j-1}}^{t_j} (t - t_{j-1})^2 \partial_{ttt} u \, dt + \frac{r_j^2}{2(1 + r_j)\tau_j} \int_{t_{j-2}}^{t_{j-1}} (t - t_{j-2})^2 \partial_{ttt} u \, dt
\]

\[
= -\frac{1}{2} (b_0^{(j)} - b_1^{(j)}) \int_{t_{j-1}}^{t_j} (t - t_{j-1})^2 \partial_{ttt} u \, dt - \frac{1}{2} b_1^{(j)} \int_{t_{j-2}}^{t_{j-1}} (t - t_{j-2})^2 \partial_{ttt} u \, dt
\]

\[
- \frac{1}{2} b_1^{(j)} \int_{t_{j-1}}^{t_j} (t - t_{j-1} + \tau_{j-1})^2 \partial_{ttt} u \, dt
\]

\[
= -\frac{1}{2} b_0^{(j)} \int_{t_{j-1}}^{t_j} (t - t_{j-1})^2 \partial_{ttt} u \, dt - \frac{1}{2} b_1^{(j)} \int_{t_{j-2}}^{t_{j-1}} (s - t_{j-2})^2 \partial_{ttt} u \, dt
\]

\[
- \frac{1}{2} b_1^{(j)} \tau_{j-1} \int_{t_{j-1}}^{t_j} (2(t - t_{j-1}) + \tau_{j-1}) \partial_{ttt} u \, dt \quad \text{for} \quad j \geq 2,
\]
where the BDF2 convolution kernels \( \text{(1.6)} \) with \( b_0^{(j)} - b_1^{(j)} = (1 + r_j)/\tau_j \) have been used. So we apply the equality 
\(-b_1^{(j)}/b_0^{(j)} = r_1^{2}/1 + 2r_j \) to obtain

\[
\|\eta^j\| \leq \frac{1}{2} b_0^{(j)} \tau_j^2 G_{t3}^{j} \frac{1}{2} b_1^{(j)} (2r_j + \tau_{j-1}) \tau_{j-1} G_{t3}^{j-1} - \frac{1}{2} b_1^{(j)} r_j^2 \tau_{j-1} G_{t3}^{j-1} \\
= b_0^{(j)} \left[ \tau_j^2 G_{t3}^{j} - \frac{b_1^{(j)}}{b_0^{(j)}} (1 + 2r_j) \tau_{j-1} G_{t3}^{j} - \frac{b_1^{(j)}}{b_0^{(j)}} r_j^2 \tau_{j-1} G_{t3}^{j-1} \right] \\
= b_0^{(j)} \tau_j^2 G_{t3}^{j} + \frac{r_j^2 \tau_{j-1}}{2(1 + 2r_j)} b_0^{(j)} G_{t3}^{j-1} \quad \text{for } j \geq 2.
\]

By using Lemma \( \text{(2.8)} \) and the recursive formula \( \text{(2.7)} \), we derive that

\[
\sum_{j=2}^{n} \|\eta^j\| \sum_{k=j}^{n} \theta_{k-j}^{(k)} \leq \sum_{j=2}^{n} \tau_j^2 G_{t3}^{j} \sum_{k=j}^{n} \theta_{k-j}^{(k)} + \frac{n}{2} \sum_{j=2}^{n} \tau_j^2 \tau_{j-1} G_{t3}^{j} \sum_{k=j}^{n} \theta_{k-j}^{(k)} \\
= \sum_{j=2}^{n} \tau_j^2 G_{t3}^{j} \sum_{k=j}^{n} \theta_{k-j}^{(k)} + \frac{n}{2} \sum_{j=2}^{n} \tau_j^2 \tau_{j-1} G_{t3}^{j} \sum_{k=j}^{n} \theta_{k-j}^{(k)} \\
= \sum_{j=2}^{n} \tau_j^2 G_{t3}^{j} \sum_{k=j}^{n} \theta_{k-j}^{(k)} + \frac{n}{2} \sum_{j=2}^{n} \tau_j^2 \tau_{j-1} G_{t3}^{j} \sum_{k=j}^{n} \theta_{k-j}^{(k)} \\
\leq \frac{3}{2} \sum_{j=1}^{n} \tau_j^2 G_{t3}^{j} \sum_{k=j}^{n} \theta_{k-j}^{(k)} \quad \text{for } 2 \leq n \leq N. \quad \text{(3.3)}
\]

Then the claimed estimate follows from the following equality

\[
\sum_{k=1}^{n} \sum_{j=1}^{k} \theta_{k-j}^{(k)} \|\eta^j\| = \|\eta^1\| \sum_{k=1}^{n} \theta_{k-1}^{(k)} + \sum_{j=2}^{n} \|\eta^j\| \sum_{k=j}^{n} \theta_{k-j}^{(k)} ,
\]

and the above two estimates \( \text{(3.2)}-\text{(3.3)} \). It completes the proof. \( \square \)

To process the error analysis, we apply the formula \( \text{(2.8)} \) to bound the terms \( \sum_{k=j}^{n} \hat{\theta}_{k-j}^{(k)} \) in Lemma \( \text{(3.2)} \) for \( 1 \leq j \leq n \). If all step ratios \( r_k \) fulfill the Grigorieff’s condition, \( 0 < r_k < 1 + \sqrt{2} \), we have \( \frac{r_j^2}{1 + 2r_j} < 1 \) for \( k \geq 2 \) and thus

\[
\sum_{k=j}^{n} \hat{\theta}_{k-j}^{(k)} = \sum_{k=j}^{n} \prod_{i=j+1}^{k} \frac{r_i^2}{1 + 2r_i} \leq \sum_{k=j}^{n} \left( \frac{r_c^2}{1 + 2r_c} \right)^{k-j} \leq \frac{1 + 2r_c}{1 + 2r_c - r_c^2} \quad \text{for } 1 \leq j \leq n,
\]

where \( r_c \) takes the maximum value of all step ratios \( r_k \). One has the following extension.

**Lemma 3.3.** Consider the discrete kernels \( \hat{\theta}_{k-j}^{(k)} \) in \( \text{(2.8)} \). If the step ratios satisfy \( S2 \), then

\[
\sum_{k=j}^{n} \hat{\theta}_{k-j}^{(k)} \leq C_r := \left( \frac{r_c^2}{1 + 2r_c} \right)^{N_0} \frac{1 + 2r_c}{1 + 2r_c - r_c^2} \quad \text{for } 1 \leq j \leq n, \quad \text{(3.4)}
\]

where \( r_c \) takes the maximum value of all step ratios \( r_k \in (0, 1 + \sqrt{2}) \) and \( \hat{r}_c \) takes the maximum value of those step ratios \( r_k \in [1 + \sqrt{2}, 3 + \sqrt{2}] \) for \( 2 \leq k \leq N \).
Let \( \tilde{u}^n := u(t_n, x) - u^n(x) \) for \( n \geq 0 \). Then the error equation of (1.3) reads
\[
D_2 \tilde{u}^n = \varepsilon \Delta \tilde{u}^n + \kappa \tilde{u}^n + \eta^n, \quad \text{for } 1 \leq n \leq N \tag{3.5}
\]
where the local consistency error \( \eta^j = D_2 u(t_j) - \partial_t u(t_j) \) for \( j \geq 1 \). If the step ratios satisfy \( S_1 \) with the maximum time-step size \( \tau \leq 1/(4\kappa^*) \), the priori estimate (3.1) in Theorem 3.1 yields
\[
\| \tilde{u}^n \| \leq 2 \exp (4\kappa^* t_{n-1}) \left( \| \tilde{u}^0 \| + 2 \sum_{k=1}^n \sum_{j=1}^k \theta^{(k)}_{k-j} \| \eta^j \| \right) \quad \text{for } 1 \leq n \leq N. \tag{3.6}
\]

With the help of Lemmas 3.2–3.3, it is easy to obtain the following result.

**Theorem 3.2.** Let \( u(t_n, x) \) and \( u^n(x) \) be the solutions of the diffusion problem (1.1) and the BDF2 scheme (1.3), respectively. If the step ratio condition \( S_2 \) holds with the maximum time-step size \( \tau \leq 1/(4\kappa^*) \), then the time-discrete solution \( u^n \) is convergent in the \( L^2 \) norm,
\[
\| u(t_n) - u^n \| \leq \frac{2C_r \exp (4\kappa^* t_{n-1})}{C_r} \left( \| u_0 - u^0 \| + 2\tau \int_0^{t_1} \| \partial_t u \| \, dt + 3 \sum_{j=1}^n \tau_j^2 \int_{t_{j-1}}^{t_j} \| \partial_{tt} u \| \, dt \right)
\]
for \( 1 \leq n \leq N \), where the mesh-dependent constant \( C_r = C_r (N_0, r_c, \hat{r}_c) \) is defined in (3.4).

Although large step ratios are allowed in the condition \( S_2 \), the users are suggested to choose the Grigorieff’s step-ratio restriction \( r_k \in (0, 1 + \sqrt{2}) \). In such case, \( N_0 = 0 \) and \( C_r = \frac{1+2r_c}{1+2r_c-r_c^2} \).

Generally, when the time-step ratios \( r_k \) are chosen so that the BDF2 kernels \( b_{n-k}^{(n)} \) are positive semi-definite (the condition \( S_1 \) is sufficient), the series \( \sum_{k=1}^n \hat{\theta}^{(k)}_{k-1} \) in (3.6) would be unbounded as the step sizes vanish. On the other hand, the solution remains the first-order convergence
\[
\sum_{j=1}^n \sum_{k=j}^n \tau_j \hat{\theta}^{(k)}_{k-j} = \sum_{j=1}^n \sum_{k=j}^n \hat{b}^{(j)}_{r_c} \tau_j \theta^{(k)}_{k-j} \leq 2t_n \quad \text{for } 1 \leq n \leq N.
\]

Then the error estimate (3.6) gives the following theorem.

**Theorem 3.3.** If the BDF2 kernels \( b_{n-k}^{(n)} \) in (1.6) are positive semi-definite (or the sufficient condition \( S_1 \) holds) and the maximum time-step size \( \tau \leq 1/(4\kappa^*) \), then the solution \( u^n \) of BDF2 scheme (1.3) is convergent in the \( L^2 \) norm in the sense that
\[
\| u(t_n) - u^n \| \leq \frac{2 \exp (4\kappa^* t_{n-1})}{C_r} \left( \| u_0 - u^0 \| + 2t_n \int_0^{t_1} \| \partial_t u \| \, dt + 3 \max_{1 \leq j \leq n} \tau_j \int_{t_{j-1}}^{t_j} \| \partial_{tt} u \| \, dt \right),
\]
for \( 1 \leq n \leq N \). If the BDF1 scheme in (1.3) is replaced by some high-order starting scheme, one can follow the proof of Corollary 2.2 to derive that
\[
\| u(t_n) - u^n \| \leq \frac{2 \exp (4\kappa^* t_{n-1})}{C_r} \left( \| u(t_1) - u^1 \| + 2t_n \| \partial_{c}(u(t_1) - u^1) \| \right)
\]
\[
+ 3t_n \max_{1 \leq j \leq n} \tau_j \int_{t_{j-1}}^{t_j} \| \partial_{tt} u \| \, dt \quad \text{for } 2 \leq n \leq N.
\]
4 Numerical example

The nonuniform BDF2 method together with the Fourier pseudo-spectral in space is applied to solve the heat equation $\partial_t u = \varepsilon \Delta u + f$ on the space-time domain $(0, 2)^2 \times (0, 1]$. The exterior force $f$ is chosen so that the equation admits an exact solution $u = e^{-t}\sin 2\pi x \cos 2\pi y$.

| $N$  | $e(N)$  | $\tau$  | Order | $\max r_k$ | $N_1$ |
|------|---------|---------|-------|------------|-------|
| 64   | 1.56e-02| 1.12e-01| –     | 13.74      | 3     |
| 128  | 3.24e-03| 6.09e-02| 2.27  | 15.26      | 8     |
| 256  | 8.66e-04| 3.27e-02| 1.90  | 32.15      | 13    |
| 512  | 1.67e-04| 1.59e-02| 2.38  | 395.6      | 26    |
| 1024 | 4.45e-05| 7.26e-03| 1.91  | 60.13      | 40    |

| $N$  | $e(N)$  | $\tau$  | Order | $\max r_k$ | $N_1$ |
|------|---------|---------|-------|------------|-------|
| 64   | 9.79e-02| 1.32e-01| –     | 10.94      | 2     |
| 128  | 2.13e-02| 6.36e-02| 2.20  | 13.62      | 7     |
| 256  | 6.06e-03| 3.02e-02| 1.82  | 81.12      | 10    |
| 512  | 1.40e-03| 1.49e-02| 2.11  | 604.0      | 24    |
| 1024 | 3.61e-04| 7.59e-03| 1.96  | 448.2      | 53    |

We consider the arbitrary mesh with random time-steps $\tau_k = T \epsilon_k / S$ for $1 \leq k \leq N$, where $S = \sum_{k=1}^{N} \epsilon_k$ and $\epsilon_k \in (0, 1)$ are random numbers subject to the uniform distribution. No any special treatments have been used to adjust the time-steps so that some large step ratios appear in our experiments, see the fifth column in Tables 4.1-4.2. In each run, the $L^2$ norm error $e(N) := \|u(T) - u^N\|$ at the final time $T = 1$ is recorded in Tables 4.1-4.2, in which we also list the maximum step-size $\tau$, the maximum step ratio and the number (denote $N_1$ in tables) of time levels with the step ratio $r_k \geq (3 + \sqrt{17})/2$. The experimental rate of convergence is estimated by $\text{Order} \approx \log_2 (e(N)/e(2N))$. From the current data and more tests not listed here, we see that the adaptive BDF2 time-stepping is robustly stable and second-order convergent, at least when the frequency of large step ratios is very low ($N_1/N \approx 5\%$ in our tests).

5 Concluding remarks

Consider some multi-step scheme having the discrete kernels $\{B^{(n)}_{n-k}\}_{k=1}^{n}$ for parabolic equations,

$$\sum_{k=1}^{n} B^{(n)}_{n-k} \nabla^2 u^k = \varepsilon \Delta u^n + f^n \quad \text{for } n \geq 1 \text{ and } B^{(n)}_0 \neq 0.$$
We present a novel framework for the numerical analysis by introducing a new class of DOC kernels \( \{ \Theta^{(n)}_{n-k} \}_{k=1}^{n} \) defined via the orthogonal identity

\[
\sum_{j=k}^{n} \Theta^{(n)}_{n-j} B^{(j)}_{j-k} \equiv \delta_{nk} \quad \text{for } 1 \leq k \leq n.
\]

Taking the advantage of orthogonality, one has the following alternative form

\[
\nabla \tau u^n = \varepsilon \sum_{j=1}^{n} \Theta^{(n)}_{n-j} \Delta u^j + \sum_{j=1}^{n} \Theta^{(n)}_{n-j} f^j \quad \text{for } n \geq 1.
\]

If the discrete kernels \( B^{(n)}_{n-k} \) are positive semi-definite, then the orthogonality implies the positive semi-definiteness of DOC kernels \( \Theta^{(n)}_{n-k} \). So one has the following \( L^2 \) norm priori estimate

\[
\| u^n \|^2 \leq \| u^0 \|^2 + 2 \sum_{k=1}^{n} \sum_{j=1}^{k} \langle u^k, \Theta^{(k)}_{k-j} f^j \rangle \quad \text{for } n \geq 1.
\]

For the adaptive BDF2 method applied to linear reaction-diffusion equations, the above approach provides a concise stability and convergence theory, which seems quite similar to that of the most robust BDF1 scheme. Some applications will be reported in subsequent articles for the numerical analysis of nonuniform BDF2 time-stepping scheme in simulating the gradient flows [12, 13], which always permit multiply time scales in approaching the steady state. We expect that the novel theoretical framework will be useful to establish the optimal \( L^2 \) norm error estimate for some other nonlocal time approximations having a discrete convolution form.

**Acknowledgements**

The authors would like to thank Dr. Buyang Li for his valuable discussions and fruitful suggestions.

**References**

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

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

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

[4] E. Emmrich, *Stability and error of the variable two-step BDF for semilinear parabolic problems*, J. Appl. Math. & Computing, 19 (2005), pp. 33–55.

[5] C.W. Gear and K.W. Tu, *The effect of variable mesh size on the stability of multistep methods*, SIAM J. Num. Anal., 11 (1974), pp.1025–1043.
[6] R.D. Grigorieff, Stability of multistep-methods on variable grids, Numer. Math., 42 (1983), pp. 359–377.
[7] E. Hairer, S.P. Norsett and G. Wanner, Solving Ordinary Differential Equations I: Nonstiff Problems, Volume 8 of Springer Series in Computational Mathematics, Second Edition, Springer-Verlag, 1992.
[8] W. Hundsdorfer, S. J. Ruuth and R. J. Spiteri, Monotonicity-preserving linear multistep methods, SIAM J. Numer. Anal., 41(2) (2004), pp. 605-623.
[9] M.-N. Le Roux, Variable step size multistep methods for parabolic problems, SIAM J. Numer. Anal., 19 (4) (1982), pp. 725–741.
[10] H.-L. Liao and Z.Z. Sun, Maximum norm error bounds of ADI and compact ADI methods for solving parabolic equations, Numer. Methods Partial Differ. Eq., 26 (2010), pp. 37-60.
[11] H.-L. Liao, P. Lyu and S. Vong, Second-order BDF time approximation for Riesz space-fractional diffusion equations, Inter. J. Comput. Math., 95 (2018), pp. 144–158.
[12] H.-L. Liao, B. Ji and L. Zhang, An adaptive BDF2 implicit time-stepping method for the phase field crystal model, 2019, preprint.
[13] H.-L. Liao, L. Wang and Z. Zhang, The analysis using discrete orthogonal convolution kernels for BDF2 time-stepping to simulate the Allen-Cahn equation, 2019, preprint.
[14] H. Nishikawa, On large start-up error of BDF2, J. Comput. Phys., 392 (2019), pp. 456–461.
[15] L.F. Shampine and M.W. Reichelt, The MATLAB ODE suite, SIAM J. Sci. Comput., 18(1) (1997), pp. 1–22.
[16] V. Thomée, Galerkin Finite Element Methods for Parabolic Problems, Second Edition, Springer-Verlag, 2006.