Composite backward differentiation formula: an extension of the TR-BDF2 scheme

Wenjun Ying

Department of Mathematics & Institute of Natural Sciences, Shanghai Jiao Tong University, Minhang, Shanghai 200240 (`wying@sjtu.edu.cn`*).*

Craig S. Henriquez

Departments of Biomedical Engineering and Computer Science, Duke University, P. O. Box 90281, Durham, NC 27708 (`ch@duke.edu`*).*

Donald J. Rose

Department of Computer Science, Duke University, P. O. Box 90129, Durham, NC 27708 (`djr@cs.duke.edu`*).*

Abstract

This paper presents a class of one-step multi-stage backward differentiation formulas (BDFs), called composite BDFs or C-BDFs, as a generalization of the TR-BDF2 composite scheme. These schemes are equivalent to singly diagonally implicit Runge-Kutta (SDIRK) methods of special type, and regarded as the one-step analogs of the multi-step Gear's methods, the so-called backward differentiation formulas (BDFs). Unlike the standard BDFs, however, the C-BDFs do not need external startup calculation while maintaining the full accuracy of the scheme. The C-BDF can be easily implemented as it is made up of the simplest Forward Euler and Backward Euler processes, which are interleaved with interpolation and extrapolation operations on intermediate solutions.

Key words: stiff problems, multiscale problems, implicit ODE solver, TR-BDF2 Scheme, IM-BDF2 Scheme, composite backward differentiation formula, C-BDF Scheme
PACS: 65L05, 65L06, 65L20

1 Introduction

Many problems in biology can be described by systems of ordinary differential equations [16, 20, 21]. These problems are often multiscale and stiff and hence are amenable to implicit solvers, given their better stability properties. While a variety of implicit solvers have been developed [7, 8, 13, 14, 23, 24, 26], many computational biologists avoid their use, as they appear algorithmically complex.

One can view standard time integration methods as falling into one of three families, assuming that those based on Richardson extrapolation [10] and deferred correction techniques [11, 18], including the classic Picard iteration method, are classified as advanced or non-standard methods. The first family of methods are the one-step multi-stage Runge-Kutta methods [6], which are based on numerical integration of derivative functions. The linearly implicit methods, named Rosenbrock-Wanner methods [13], can also be interpreted as one step, multi-stage Runge-Kutta methods. The second family of methods are the multi-step Adams methods [13,17], which were originally derived from numerical integration of derivative-interpolating polynomials. The Adams methods can be treated as the multi-step analogs of the one-step Runge-Kutta methods and vice versa. Unlike the Runge-Kutta and Adams methods, the third family of methods, popularly called as multi-step Gear's methods or backward differentiation formulas (BDFs) [12,25], are based on numerical differentiation of solution-interpolating polynomials. All the methods have high order extensions.

While this classification is somewhat arbitrary, there are currently no standard methods that can be viewed as high-order one-step analog of the multi-step Gear's methods (see Table 1). In this paper, we show that if properly constructed, composite implicit methods, such as TR-BDF2 [4,15], can be viewed as one-step, multi-stage methods. For example, by replacing the trapezoidal part in the TR-BDF2 scheme with the second-order implicit midpoint rule, another composite scheme, which we call IM-BDF2, can be derived with the same A- and L-stabilities as TR-BDF2. IM-BDF2 is a one-step, two-stage method. We show that IM-BDF2 can be extended to a one-step multi-stage scheme, where each stage uses a backward differentiation formula, BDFq. Here, the BDFq scheme is analogous to the standard BDF with q-steps. We therefore call the one-step multi-stage BDF scheme the composite BDF or C-BDF method.

In this paper, we present IM-BDF2 scheme for the first time and relate it to TR-BDF2. The stability properties of the two methods are analyzed and compared with a second-order singly diagonally implicit Runge-Kutta (SDIRK2) scheme. We then derive the generalized one-step multi-stage composite back-

Table 1
Time integration methods for initial value problems

	derivative-interpolating	solution-interpolating
one-step methods	Runge-Kutta	?
multi-step methods	Adams	Gear

ward differentiation formulas and show that the C-BDFs are equivalent to a special type of one step, multi-stage Runge-Kutta schemes. Finally, we discuss how the C-BDFs provide a computational elegance since certain elements of the algorithm can be reused and only the solutions and not the function derivatives at the intermediate time steps need to be stored.

The remainder of the paper is organized as follows. In section 2, we briefly summarize the TR-BDF2 scheme and the IM-BDF2 scheme, the latter of which has never been seen in the same form in the literature to the best of our knowledge. In section 3, two variants of the TR-BDF2 and IM-BDF2 schemes and their stability properties are analyzed and compared with a second-order singly diagonally implicit Runge-Kutta (SDIRK2) scheme. In section 4, we present the generalized one-step multi-stage composite backward differentiation formulas. In section 5, we show that the C-BDFs are equivalent to the special type Runge-Kutta schemes.

2 The Second-order TR-BDF2 and IM-BDF2 Schemes

Assume $f(u)$ is a sufficiently smooth nonlinear function of u. We consider numerically integrating the following autonomous ordinary differential equation (ODE),

$$\frac{du}{dt} = f(u), \tag{1}$$

from time t^n to $t^{n+1} = t^n + \Delta t$. The extension of the work to non-autonomous ODEs is standard and straightforward.

Let $\gamma > 0$ be a characteristic value, called as a stability constant, and let

$$\gamma_2 \equiv \frac{1-2\gamma}{2(1-\gamma)}.$$

Our discussion starts with the following two composite schemes:

- *The trapezoidal and BDF2 composite scheme (TR-BDF2),*

$$u^{n+2\gamma} - \gamma\Delta t\, f(u^{n+2\gamma}) = u^n + \gamma\Delta t\, f(u^n), \tag{2a}$$

$$u^{n+1} - \gamma_2\Delta t\, f(u^{n+1}) = \frac{1-\gamma_2}{2\gamma}u^{n+2\gamma} + \left(1 - \frac{1-\gamma_2}{2\gamma}\right)u^n. \tag{2b}$$

- *The implicit midpoint and BDF2 composite scheme (IM-BDF2),*

$$u^{n+\gamma} - \gamma\Delta t\, f(u^{n+\gamma}) = u^n, \tag{3a}$$

$$u^{n+1} - \gamma_2\Delta t\, f(u^{n+1}) = \frac{1-\gamma_2}{\gamma}u^{n+\gamma} + \left(1 - \frac{1-\gamma_2}{\gamma}\right)u^n. \tag{3b}$$

The TR-BDF2 scheme (2) was originally derived as a composite method of the trapezoidal rule and the backward-differentiation-formula of second-order (BDF2) [4]. The IM-BDF2 scheme (3) results from the replacement of the trapezoidal stage (2a) in the TR-BDF scheme with the implicit midpoint rule. Both the TR-BDF2 scheme (2) and the IM-BDF2 scheme (3) are of second-order in accuracy for any $\gamma \neq 0, 1$.

For the special value $\gamma = 1/2$, the TR-BDF2 scheme (2) degenerates to the trapezoidal rule; the IM-BDF2 scheme (3) becomes

$$\begin{aligned} u^{n+1/2} - \frac{1}{2}\Delta t f(u^{n+1/2}) &= u^n, \\ u^{n+1} &= 2u^{n+1/2} - u^n, \end{aligned}$$

which is exactly the implicit midpoint rule,

$$u^{n+1} - \Delta t f(\frac{u^n + u^{n+1}}{2}) = u^n.$$

The leading order term of the local truncation error of the TR-BDF2 scheme (2) is given by

$$\begin{aligned} E_{\text{TR-BDF2}} &= \left\{\frac{1}{6} - \frac{2\gamma^2 - 2\gamma + 1}{4(1-\gamma)}\right\}(f'^2 f + f^2 f'')\,\Delta t^3 \\ &= \left\{\frac{1}{6} - \frac{2\gamma^2 - 2\gamma + 1}{4(1-\gamma)}\right\}u'''(t^n)\,\Delta t^3. \end{aligned} \tag{5}$$

The leading order term of the local truncation error of the IM-BDF2 scheme (3) is given by

$$E_{\text{IM-BDF2}} = \left\{\frac{1}{6} - \frac{2\gamma^2 - 2\gamma + 1}{4(1-\gamma)}\right\} f'^2 f\,\Delta t^3 + \left\{\frac{1}{6} - \frac{1-\gamma}{4}\right\} f^2 f''\Delta t^3 \tag{6}$$

Let $C(\gamma)$ be the coefficient of the leading-order error $E_{\text{TR-BDF2}}$ of the TR-BDF2 scheme. It is easy to find that

$$C(\gamma) = \frac{2}{3} - \frac{1}{4}\left\{2(1-\gamma) + \frac{1}{1-\gamma}\right\} \in \left(-\infty, \frac{2}{3} - \frac{\sqrt{2}}{2}\right] \cup \left[\frac{2}{3} + \frac{\sqrt{2}}{2}, \infty\right). \quad (7)$$

The specific value

$$\gamma = 1 - \frac{\sqrt{2}}{2}$$

minimizes $|C(\gamma)|$. Later we will see that this value also makes identical the Jacobian matrices from both the trapezoidal/implicit midpoint rule and the BDF2 stages in (2)/(3) and the schemes be both A- and L-stable.

When the characteristic value γ takes the value that minimizes $|C(\gamma)|$, the leading-order errors from the TR-BDF2 and IM-BDF2 schemes are respectively

$$E_{\text{TR-BDF2}} \approx -0.0404\,(f'^2 f + f^2 f'')\Delta t^3 \quad (8)$$

and

$$E_{\text{IM-BDF2}} \approx -0.0404\,(f'^2 f + \frac{1}{4} f^2 f'')\Delta t^3. \quad (9)$$

3 Two Variants of The TR-BDF2 and IM-BDF2 Schemes

In this section, we will restrict ourselves to two variants of the two-stage TR-BDF2 and IM-BDF2 schemes, (2) and (3). In each variant, both stages use the identical time integration parameter γ and so involve the same Jacobian matrices.

- *A variant of the TR-BDF2 scheme (2) reads:*

$$u^{n+2\gamma} - \gamma\Delta t\, f(u^{n+2\gamma}) = u^n + \gamma\Delta t\, f(u^n), \quad (10a)$$

$$u^{n+1} - \gamma\Delta t\, f(u^{n+1}) = \left(\frac{1}{2\gamma} - \frac{1}{2}\right)u^{n+2\gamma} + \left(\frac{3}{2} - \frac{1}{2\gamma}\right)u^n. \quad (10b)$$

- *A variant of the IM-BDF2 scheme (3) reads:*

$$u^{n+\gamma} - \gamma\Delta t\, f(u^{n+\gamma}) = u^n, \quad (11a)$$

$$u^{n+1} - \gamma\Delta t\, f(u^{n+1}) = \left(\frac{1}{\gamma} - 1\right)u^{n+\gamma} + \left(2 - \frac{1}{\gamma}\right)u^n. \quad (11b)$$

Both of the modified TR-BDF2 scheme (10) and the modified IM-BDF2 scheme (11) above are of second-order in accuracy if and only if $\gamma = 1 \pm \sqrt{2}/2$, the zeros of the quadratic polynomial

$$P_2(\gamma) \equiv \gamma^2 - 2\gamma + \frac{1}{2}. \tag{12}$$

The TR-BDF2 and IM-BDF2 schemes, (10) and (11), have the same stability function, for general $\gamma > 0$,

$$S_\gamma(z) = \frac{1 + (1-2\gamma)z}{(1-\gamma z)^2} \quad \text{for } z \in \mathbb{C}. \tag{13}$$

It can be shown [27] that the modified schemes above are both A-Stable and L-Stable for $\gamma = 1 \pm \sqrt{2}/2$.

Remark 1 *As a comparison with the second-order TR-BDF2 and IM-BDF2 schemes, (10) and (11), we describe here a singly diagonally implicit Runge-Kutta (SDIRK2) scheme [9, 13],*

$$u^{n+1} = u^n + \frac{\Delta t}{2}\Big(k_1 + k_2\Big) \tag{14}$$

with

$$\begin{aligned} k_1 &= f(u^n + \gamma\Delta t\, k_1), \\ k_2 &= f(u^n + (1-2\gamma)\Delta t\, k_1 + \gamma\Delta t\, k_2). \end{aligned}$$

The corresponding stability function $S_\gamma(z)$ of the SDIRK2 scheme (14) is given by

$$S_\gamma(z) = \frac{1 + (1-2\gamma)z + (\gamma^2 - 2\gamma + 1/2)z^2}{(1-\gamma z)^2}.$$

The leading order term in the local truncation error for the SDIRK2 scheme (14) is given by

$$E_{SDIRK2} = \left\{\frac{1}{6} - \gamma(1-\gamma)\right\} f'^2 f\, \Delta t^3 + \left\{\frac{1}{6} - \frac{2\gamma^2 - 2\gamma + 1}{4}\right\} f^2 f''\, \Delta t^3.$$

The SDIRK2 scheme is both A- and L-stable if and only if

$$\gamma = 1 \pm \frac{\sqrt{2}}{2}. \tag{15}$$

When the characteristic value $\gamma = 1 - \sqrt{2}/2$, the leading-order error from the SDIRK2 scheme is

$$E_{SDIRK2} \approx -0.0404\,(f'^2 f - \frac{1}{2} f^2 f'')\Delta t^3. \tag{16}$$

4 The Composite Backward Differentiation Formulas

Now we present the generalization of the second-order composite schemes, (10) and (11), and collectively call the generalized schemes as composite backward differentiation formulas (C-BDFs, also mnemonic for the "counterpart of BDFs" or "closed BDFs" since they are BDFs but do not need external startup calculation).

Let $q > 0$ be the number of stages, $\gamma > 0$ be the characteristic constant, $\{\theta_i\}$ be the temporal parameters and the coefficients $\{\beta_{i,j}\}$ to be determined. Assume $w_0 = u^n \approx u(t^n)$, $w_i \approx u(t^n + \theta_i \Delta t)$ (for $i = 1, 2, \cdots, q$) and $u^{n+1} = w_q$.

- *A q-stage TR-BDFs scheme is in the following form:*

$$w_1 - \gamma \Delta t\, f(w_1) = w_0 + \gamma \Delta t\, f(w_0), \tag{17a}$$

$$w_i - \gamma \Delta t\, f(w_i) = \sum_{j=0}^{i-1} \beta_{i,j} w_j \quad \textit{for } i = 2, 3, \cdots, q. \tag{17b}$$

- *A q-stage IM-BDFs scheme is in the following form:*

$$w_1 - \gamma \Delta t\, f(w_1) = w_0, \tag{18a}$$

$$w_i - \gamma \Delta t\, f(w_i) = \sum_{j=0}^{i-1} \beta_{i,j} w_j \quad \textit{for } i = 2, 3, \cdots, q. \tag{18b}$$

For simplicity, we only restrict our discussion to the IM-BDFs scheme (18), which also reads

$$w_i - \gamma \Delta t\, f(w_i) = \sum_{j=0}^{i-1} \beta_{i,j} w_j \quad \text{for } i = 1, 2, \cdots, q. \tag{19}$$

We could determine the values of γ, $\{\theta_i\}$ and $\{\beta_{i,j}\}$ by the method of undetermined coefficients as usual through imposing consistency condition, which yields

$$\sum_{j=0}^{i-1} \beta_{i,j} = 1, \tag{20}$$

and order conditions.

Let $v_i = w_i - w_0$ $(i = 1, 2, \cdots, p)$. We have the following Taylor expansions around w_0:

$$f(w_i) = \sum_{k=0}^{p-1} \frac{1}{k!} f^{(k)}(w_0) v_i^k + O(v_i^p),$$

and

$$v_1 = \gamma\Delta t \sum_{k=0}^{p-1} \frac{1}{k!} f^{(k)}(w_0) v_1^k + O(\gamma\Delta t\, v_1^p), \tag{21a}$$

$$v_i = \sum_{j=1}^{i-1} \beta_{i,j} v_j + \gamma\Delta t \sum_{k=0}^{p-1} \frac{1}{k!} f^{(k)}(w_0) v_i^k + O(\gamma\Delta t\, v_i^p) \quad \text{for } i > 1. \tag{21b}$$

In general, it becomes extremely involved to explicitly write down order conditions in terms of the characteristic value γ and the coefficients $\{\beta_{i,j}\}$ as the order number p increases. We postpone the computation of the coefficients for general $p > 3$ until the next section. In the rest of this section, we only find the coefficients $\beta_{i,j}$ for the third-order scheme ($p = 3$). For conciseness, denote the k^{th}-order derivative $f^{(k)}(w_0)$ of $f(w)$ at w_0 simply by $f^{(k)}$; and $f' = f^{(1)}$, $f'' = f^{(2)}$, $f''' = f^{(3)}$.

First of all, note that

$$v_1 = \gamma\, \Delta t\, f + \gamma v_1\, \Delta t f' + \frac{1}{2}\gamma v_1^2\, \Delta t f'' + \cdots, \tag{22}$$

$$v_2 = \gamma(1 + \beta_{2,1})\, \Delta t f + \gamma(v_2 + \beta_{2,1} v_1)\, \Delta t f' + \frac{1}{2}\gamma(v_2^2 + \beta_{2,1} v_1^2)\, \Delta t f'' + \cdots, \tag{23}$$

and

$$\begin{aligned} v_3 = {} & \gamma\Big[1 + \beta_{3,1} + \beta_{3,2}(1 + \beta_{2,1})\Big]\, \Delta t f \\ & + \gamma\Big[v_3 + \beta_{3,1} v_1 + \beta_{3,2}(v_2 + \beta_{2,1} v_1)\Big]\, \Delta t f' \\ & + \frac{1}{2}\gamma\Big[v_3^2 + \beta_{3,1} v_1^2 + \beta_{3,2}(v_2^2 + \beta_{2,1} v_1^2)\Big]\, \Delta t f'' + \cdots. \end{aligned} \tag{24}$$

Equating the coefficients of low-order terms of v_3 with those in a Taylor expansion of the exact solution $u(t)$ around t^n gives us the following identities:

$$\Delta t f: \quad 1 = \gamma\Big[1 + \beta_{3,1} + \beta_{3,2}(1 + \beta_{2,1})\Big], \tag{25a}$$

$$\Delta t^2 f f': \quad \frac{1}{2} = \gamma\Big\{1 + \beta_{3,1}\gamma + \beta_{3,2}\Big[(1 + \beta_{2,1})\gamma + \beta_{2,1}\gamma\Big]\Big\}, \tag{25b}$$

$$\Delta t^3 f (f')^2: \quad \frac{1}{6} = \gamma\Big\{\frac{1}{2} + \beta_{3,1}\gamma^2 + \beta_{3,2}\Big[\gamma\Big[\gamma(1 + \beta_{2,1}) + \beta_{2,1}\gamma\Big] + \beta_{2,1}\gamma^2\Big]\Big\}, \tag{25c}$$

$$\Delta t^3 f^2 f'': \quad \frac{1}{6} = \frac{1}{2}\gamma\Big\{1 + \beta_{3,1}\gamma^2 + \beta_{3,2}\Big[(1 + \beta_{2,1})^2\gamma^2 + \beta_{2,1}\gamma^2\Big]\Big\}. \tag{25d}$$

From (25a)-(25b), we get

$$\gamma + \gamma(1-\gamma) + \beta_{3,2}\beta_{2,1}\gamma^2 = \frac{1}{2}. \tag{26}$$

From (25c)-(25d), we get

$$\frac{1}{2} + \beta_{3,2}\beta_{2,1}\beta_{2,1}\gamma^2 = \frac{1}{6\gamma}. \tag{27}$$

Let

$$\zeta = \beta_{3,2}\beta_{2,1}. \tag{28}$$

Equation (25a) can be rewritten as

$$(\beta_{3,1} + \beta_{3,2}) + \zeta = \frac{1}{\gamma} - 1. \tag{29}$$

Equation (25c) can be written as

$$(\beta_{3,1} + \beta_{3,2}) + 3\zeta = \frac{1}{\gamma^2}(\frac{1}{6\gamma} - \frac{1}{2}) \tag{30}$$

Subtracting (29) from (30) gives us

$$\zeta = \frac{1}{2}\left[\frac{1}{\gamma^2}(\frac{1}{6\gamma} - \frac{1}{2}) + 1 - \frac{1}{\gamma}\right]. \tag{31}$$

Note, from (26), we have another expression for ζ,

$$\zeta = \frac{1}{\gamma^2}\left[\frac{1}{2} - 2\gamma + \gamma^2\right]. \tag{32}$$

Combining (31) and (32) yields a nonlinear equation for γ,

$$\frac{1}{2}\left[\frac{1}{\gamma^2}(\frac{1}{6\gamma} - \frac{1}{2}) + 1 - \frac{1}{\gamma}\right] = \frac{1}{\gamma^2}\left[\frac{1}{2} - 2\gamma + \gamma^2\right], \tag{33}$$

or the cubic equation

$$P_3(\gamma) \equiv \gamma^3 - 3\gamma^2 + \frac{3}{2}\gamma - \frac{1}{6} = 0, \tag{34}$$

which has three distinct real roots:

$$\gamma_0 = 1 + \sqrt{2}\cos\frac{\varphi}{3},$$
$$\gamma_1 = 1 + \sqrt{2}\cos\frac{\varphi + 2\pi}{3},$$
$$\gamma_2 = 1 + \sqrt{2}\cos\frac{\varphi + 4\pi}{3},$$

Table 2
Coefficients for the third-order three-stage IM-BDFs scheme (19).

Coefficient	Value
γ	0.158983899989
$\beta_{2,0}$	- 1.644972541469
$\beta_{2,1}$	2.644972541469
$\beta_{3,0}$	3.911869324372
$\beta_{3,1}$	- 6.012776528027
$\beta_{3,2}$	3.100907203655

Coefficient	Value
γ	0.435866521508
$\beta_{2,0}$	0.352859819861
$\beta_{2,1}$	0.647140180139
$\beta_{3,0}$	-1.250979895058
$\beta_{3,1}$	3.729329662446
$\beta_{3,2}$	-1.478349767388

with

$$\varphi = \arctan \frac{\sqrt{2}}{4}.$$

After γ has been determined, back-substituting it into (32) and (27) gives us the values of ζ,

$$\beta_{2,1} = \frac{1}{\zeta\gamma^2}\left(\frac{1}{6\gamma} - \frac{1}{2}\right), \tag{35}$$

$\beta_{3,2} = \zeta/\beta_{2,1}$, $\beta_{3,1} = 1/\gamma - 1 - \zeta - \beta_{3,2}$, and so on. See Table 2 for two sets of parameters, which correspond to the cases when $\gamma < 1$.

Before closing this section, we will work on the stabilities of the general C-BDF schemes and the existence of the corresponding characteristic constant γ.

Let $\mathbf{w} = (w_1, w_2, \cdots, w_q)^T \in \mathbb{R}^q$ and

$$\mathbf{L} = \begin{bmatrix} 0 & & & & & \\ \beta_{2,1} & 0 & & & & \\ \beta_{3,1} & \beta_{3,2} & 0 & & & \\ \beta_{4,1} & \beta_{4,2} & \beta_{4,3} & \ddots & & \\ \vdots & \vdots & \vdots & \ddots & 0 & \\ \beta_{q,1} & \beta_{q,2} & \beta_{q,3} & \cdots & \beta_{q,q-1} & 0 \end{bmatrix} \in \mathbb{R}^{q\times q}. \tag{36}$$

Let $\mathbf{e}_q = (0, \cdots, 0, 1)^T \in \mathbb{R}^q$ and $\mathbf{e} = (1, 1, \cdots, 1)^q \in \mathbb{R}^q$ and $\mathbf{I} \in \mathbb{R}^{q\times q}$ be the identity matrix. When applied to time integration of the simple linear ODE,

$$\frac{du(t)}{dt} = \lambda\, u(t) \quad \text{with } \lambda \in \mathbb{C} \text{ and } \mathrm{Re}(\lambda) < 0 \tag{37}$$

from t^n to $t^{n+1} = t^n + \Delta t$, the q-stage IM-BDFs scheme (19) leads to the following linear system

$$(1-\gamma z)\mathbf{w} = (\mathbf{I}-\mathbf{L})\mathbf{e}\, w_0 + \mathbf{L}\mathbf{w}, \tag{38}$$

or

$$\mathbf{w} = [(1-\gamma z)\mathbf{I}-\mathbf{L}]^{-1}(\mathbf{I}-\mathbf{L})\,\mathbf{e}\, w_0, \tag{39}$$

with $z = \lambda\Delta t$. Note that $w_0 = u^n \approx u(t^n)$ and $u^{n+1} = w_q \approx u(t^{n+1})$. So,

$$w_q = \mathbf{e}_q^T\mathbf{w} = \mathbf{e}_q^T\,[(1-\gamma z)\mathbf{I}-\mathbf{L}]^{-1}(\mathbf{I}-\mathbf{L})\mathbf{e}\, w_0. \tag{40}$$

The stability function $S(z)$ of the q-stage IM-BDFs scheme (19) is given by

$$S_\gamma(z) = \mathbf{e}_q^T\,[(1-\gamma z)\mathbf{I}-\mathbf{L}]^{-1}(\mathbf{I}-\mathbf{L})\,\mathbf{e}. \tag{41}$$

Note that the determinant of the matrix $[(1-\gamma z)\mathbf{I}-\mathbf{L}]$ is equal to $(1-\gamma z)^q$. By Cramer's rule, the stability function $S(z)$ is a rational polynomial in the form

$$S_\gamma(z) = \frac{Q(\gamma z)}{(1-\gamma z)^q} = c_0 + c_1(\gamma z) + \cdots + c_p(\gamma z)^p + O((\gamma z)^{p+1}), \tag{42}$$

where the polynomial $Q(\gamma z)$ in the numerator has a degree not greater than $(q-1)$. As $\gamma \neq 0$, it is obvious that

$$S_\gamma(\infty) = 0. \tag{43}$$

So, if the scheme is A-stable, it must be L-stable.

If the IM-BDFs scheme (19) is of p^{th}-order in accuracy, the characteristic value γ and the coefficients $\beta_{i,j}$ must satisfy

$$S_\gamma(z) - e^z = c_0 + c_1(\gamma z) + \cdots + c_p(\gamma z)^p + O((\gamma z)^{p+1}) - e^z = O(z^{p+1}), \tag{44}$$

i.e.,

$$c_0 + c_1(\gamma z) + c_2(\gamma z)^2 + \cdots + c_p(\gamma z)^p - \sum_{k=0}^{p}\frac{1}{k!}z^k = 0, \tag{45}$$

with constants $\{c_k\}_{k=0}^p$ dependent of $\{\beta_{i,j}\}$. The identity (45) is true for any $z \in \mathbb{C}$. So we must have

$$c_k\gamma^k = \frac{1}{k!} \quad \text{for } k = 0, 1, \cdots, p. \tag{46}$$

Next we assume the polynomial $Q(\gamma z)$ in the numerator of (42) is in the form

$$Q(\gamma z) = a_0 + a_1(\gamma z) + a_2(\gamma z)^2 + \cdots + a_{q-1}(\gamma z)^{q-1}. \tag{47}$$

As

$$
\begin{aligned}
Q(\gamma z) &= (1-\gamma z)^p \, S_\gamma(z) \\
&= (1-\gamma z)^p \left\{ \sum_{k=0}^{p} c_k(\gamma z)^k + O((\gamma z)^{p+1}) \right\} \\
&= \sum_{k=0}^{p} \binom{p}{k} (-\gamma z)^k \left\{ \sum_{k=0}^{p} c_k(\gamma z)^k + O((\gamma z)^{p+1}) \right\},
\end{aligned}
\tag{48}
$$

the coefficient a_i of the i^{th}-order term in the polynomial $Q(\gamma z)$ can be computed by

$$
a_i = \gamma^{-i} \sum_{k=0}^{i} \binom{p}{i-k} (-\gamma)^{i-k} c_k \gamma^k = \sum_{k=0}^{i} \binom{p}{i-k} (-1)^{i-k} \gamma^{-k} \frac{1}{k!}
$$

for $i = 0, 1, \cdots$. Recall that the polynomial $Q(\gamma z)$ has a degree less than $q = p$, i.e., $a_p = 0$. This implies that the characteristic value γ must satisfy

$$
0 = a_p = \sum_{k=0}^{p} \binom{p}{p-k} (-1)^{p-k} \gamma^{-k} \frac{1}{k!}. \tag{49}
$$

or γ is a root of the following polynomial of degree $n = p$,

$$
P_n(\gamma) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{1}{k!} \gamma^{n-k}. \tag{50}
$$

The first a few polynomials ($p \leq 4$) are given by

$$
\begin{aligned}
P_1(\gamma) &= \gamma - 1, \\
P_2(\gamma) &= \gamma^2 - 2\gamma + \frac{1}{2}, \\
P_3(\gamma) &= \gamma^3 - 3\gamma^2 + \frac{3}{2}\gamma - \frac{1}{6}, \\
P_4(\gamma) &= \gamma^4 - 4\gamma^3 + 3\gamma^2 - \frac{2}{3}\gamma + \frac{1}{24}.
\end{aligned}
$$

Note that the polynomial defined by

$$
L_n(\xi) = \xi^n P_n(\frac{1}{\xi}) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{1}{k!} \xi^k \tag{52}
$$

is a Laguerre polynomial of degree n. The Laguerre polynomials $\{L_n(\xi)\}_{n=0}^{\infty}$ are orthogonal with weight $e^{-\xi}$ on the interval $[0, \infty)$. The Laguerre polynomial of degree n has n distinct positive real roots [19]. So, the polynomial (50) must have n distinct roots too.

5 Equivalence between C-BDFs and SDIRK Schemes

In this section, we will see that the IM-BDFs scheme is equivalent to a Radau-type singly diagonally implicit Runge-Kutta (SDIRK) scheme even for general nonlinear ODEs [2, 15, 22]. In the Butcher tableau for a Radau-type SDIRK scheme (see Table 3), the vector $\mathbf{b}^T$ is equal to the last row of the coefficient matrix $\mathbf{A}$ [6].

Table 3
A Butcher tableau

$\mathbf{c}$	$\mathbf{A}$
	$\mathbf{b}^{\mathrm{T}}$

We consider the following special SDIRK scheme, which is of Radau-type:

$$w_i = w_0 + \Delta t \sum_{j=1}^{i} \alpha_{i,j} f(w_j) \quad \text{for } i = 1, 2, \cdots, q, \tag{53}$$

with $\mathbf{A} = (\alpha_{i,j})_{q\times q}$, $\alpha_{1,1} = \alpha_{2,2} = \cdots = \alpha_{q,q} = \gamma$. and $\alpha_{i,j} = 0$ if $j > i$. We assume that $w_i \approx u(t^n + \theta_i \Delta t)$ with $0 = \theta_0 < \theta_1 < \theta_2 < \cdots < \theta_q = 1$. Let $\mathbf{w} = (w_1, w_2, \cdots, w_q)^T$, and let $f_j \equiv f(w_j)$ and $\mathbf{f} = (f_1, f_2, \cdots, f_q)^T$. The SDIRK scheme (53) can be rewritten as

$$\mathbf{w} = w_0\, \mathbf{e} + \Delta t \mathbf{A} \mathbf{f}. \tag{54}$$

So,

$$\mathbf{f} = \frac{1}{\Delta t} \mathbf{A}^{-1} (\mathbf{w} - w_0\, \mathbf{e}). \tag{55}$$

For simplicity, the $(i,j)^{th}$ entry in the inverse $\mathbf{A}^{-1}$ is denoted by $\frac{1}{\gamma}\alpha^*_{i,j}$. Then we have

$$f_i = \frac{1}{\gamma \Delta t} \sum_{j=1}^{i} \alpha^*_{i,j} (w_j - w_0) \quad \text{for } i = 1, 2, \cdots, q. \tag{56}$$

The SDIRK scheme (53) can be reformulated as

$$w_i - \gamma \Delta t f(w_i) = w_0 + \Delta t \sum_{j=1}^{i-1} f_j = w_0 + \frac{1}{\gamma} \sum_{j=1}^{i-1} \alpha_{i,j} \sum_{k=1}^{j} \alpha^*_{j,k} (w_k - w_0) = \sum_{j=0}^{i-1} \beta_{i,j} w_j,$$

with

$$\beta_{i,j} = \sum_{\ell=j}^{i} \alpha_{i,\ell}\, \alpha^*_{\ell,j}.$$

Similarly, we can show that the multi-stage TR-BDFs scheme (17) is equivalent to a (semi-explicit) Lobatto-type SDIRK scheme, where the last row of the coefficient matrix $\mathbf{A}$ in its Butcher tableau (see Table 3) is the same as the vector $\mathbf{b}^T$ and the first row of $\mathbf{A}$ is zero. The equivalence between the original second-order TR-BDF2 scheme (2) and a SDIRK scheme was first reported by Hosea and Shampine [15]. Based on the equivalence between the C-BDF schemes and the special type-SDIRK schemes, the existence and stabilities of the C-BDF schemes are guaranteed as long as the corresponding Radau-type or Lobatto-type SDIRK schemes exist and are A-, B-, L- or S-stable, etc. [1–3, 5]. The order conditions for the C-BDFs can also be derived from those for the Runge-Kutta schemes. For example, for the three-stage ($q = 3$) TR-BDFs scheme (17), we have the following order conditions:

$$\theta^k - \gamma\, k\, \theta^{k-1} = \mathbf{L}\, \theta^k \qquad k = 1, 2, \tag{57a}$$
$$1 - \gamma k = \mathbf{e}_q \mathbf{L}\, \theta^k \qquad k = 1, 2, 3. \tag{57b}$$

Here, $\theta = (\theta_1, \theta_2, \cdots, \theta_q)^T$. For the three-stage ($q = 3$) IM-BDFs scheme (19), the order conditions are

$$\theta^k - \gamma\, k\, \theta^{k-1} = \mathbf{L}\, \theta^k \qquad k = 1, \tag{58a}$$
$$1 - \gamma k = \mathbf{e}_q \mathbf{L}\, \theta^k \qquad k = 1, 2, 3. \tag{58b}$$

6 Discussion

This paper presents a class of one-step multi-stage backward differentiation formulas, called composite BDFs or C-BDFs, as a generalization of the TR-BDF2 composite scheme. These schemes are equivalent to special type SDIRK methods, and regarded as the one-step analogs of the multi-step BDFs. Unlike the standard BDFs, however, the C-BDFs do not need external startup calculation while maintaining the full accuracy of the scheme. The C-BDF is easily implementable, and uses the Forward Euler (FE) and the Backward Euler (BE) methods as fundamental building blocks.

Note that, for each $i > 1$, the right hand side of the TR-BDFs (17) or the IM-BDFs (19), is an interpolation or extrapolation of the past intermediate solutions $\{w_j\}_{j=0}^{i-1}$. If the interpolated or extrapolated solution is denoted by

$$v_i \equiv \sum_{j=0}^{i-1} \alpha_{i,j}\, w_j, \tag{59}$$

each stage in (17) or (19) can be regarded as a time integration of the evolution equation (1) by a timestep $\gamma\Delta t$ with the simplest Backward Euler (BE) method and the initial data given by v_i. Naturally, a good initial guess for iteratively solving the nonlinear equation,

$$w_i - \gamma \Delta t f(w_i) = v_i, \tag{60}$$

can be computed with the Forward Euler (FE) method, i.e.,

$$w_i^{(0)} = v_i + \gamma \Delta t f(v_i). \tag{61}$$

In this sense, each of the C-BDFs, (17) and (19), is made up of the simplest FE and BE processes, which are interleaved with interpolation and extrapolation operations on the intermediate solutions. Once the BE method successfully works, the multi-stage C-BDF can be readily implemented.

In this paper, we intend to emphasize the composite formula as a more algorithmic way of doing the special type SDIRK scheme rather than their equivalence. Further investigation on the possible new insight of the C-BDFs schemes from the perspective of backward differentiation is in progress. In the future, we will also study on reliable and efficient error estimation techniques for the C-BDFs.

References

[1] A. H. AL-Rabeh, Optimal order diagonally implicit Runge-Kutta methods, BIT 33 (1993) 620–633.

[2] R. K. Alexander, Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s, SIAM J. Numer. Anal. 14 (6) (1977) 1006–1021.

[3] R. K. Alexander, Stability of Runge-Kutta methods for stiff ordinary differential equations, SIAM J. Numer. Anal. 31 (4) (1994) 1147–1168.

[4] R. E. Bank, W. M. Coughran, J. W. Fichtner, E. H. Grosse, D. J. Rose, R. K. Smith, Transient simulation of silicon devices and circuits, IEEE Transactions on Electron Devices 32 (10) (1985) 1992–2007.

[5] K. Burrage, J. C. Butcher, Stability criteria for implicit Runge-Kutta methods, SIAM J. Numer. Anal. 16 (1979) 46–57.

[6] J. C. Butcher, The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods, Wiley, 1987.

[7] J. C. Butcher, Numerical Methods for Ordinary Differential Equations, Wiley, 2003.

[8] J. R. Cash, The integration of stiff initial value problems in O.D.E.S using modified extended backward differentiation formulae, Comp. and Maths. With Applics. 9 (1983) 645–657.

[9] K. Dekker, J. Verwer, Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations, North-Holland, 1984.

[10] P. Deuflhard, F. Bornemann, Scientific Computing with Ordinary Differential Equations, Springer, 2002.

[11] A. Dutt, L. Greengard, V. Rokhlin, Spectral deferred correction methods for ordinary differential equations, BIT 40 (2) (2000) 241–266.

[12] C. W. Gear, Numerical Initial Value Problems in Ordinary Differential Equations, Prentice Hall, 1971.

[13] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, Springer-Verlag, 1991.

[14] A. C. Hindmarsh, ODEPACK, a systematized collection of ODE solvers, in: R. S. Stepleman et al. (ed.), Scientific Computing, North-Holland, Amsterdam, 1983, pp. 55–64.

[15] M. E. Hosea, L. F. Shampine, Analysis and implementation of TR-BDF2, Appl. Numer. Math. 20 (1-2) (1996) 21–37.

[16] J. P. Keener, J. Sneyd, Mathematical Physiology, Springer-Verlag, New York, 1998.

[17] J. Lambert, Numerical Methods for Ordinary Differential Equations, Wiley, 1991.

[18] A. T. Layton, M. L. Minion, Conservative multi-implicit spectral deferred correction methods for reacting gas dynamics, J. Comput. Phys. 194 (2) (2004) 697–715.

[19] N. N. Lebedev, Special Functions and Their Applications, Prentice-Hall, Inc., 1965.

[20] J. D. Murray, Mathematical Biology: I. An Introduction, Third Edition, Springer, 2002.

[21] J. D. Murray, Mathematical Biology: II. Spatial Models and Biomedical Applications, Third Edition, Springer, 2004.

[22] S. P. Nørsett, P. Thomsen, Local error control in SDIRK-methods, BIT 26 (1986) 100–113.

[23] L. R. Petzold, Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations, SIAM J. Sci. Stat. Comput. 4 (1983) 136–148.

[24] L. F. Shampine, Numerical Solution of Ordinary Differential Equations, Chapman & Hall, 1994.

[25] L. F. Shampine, C. W. Gear, A users view of solving stiff ordinary differential equations, SIAM Review 21 (1979) 1–17.

[26] L. F. Shampine, M. K. Gordon, Computer Solution of Ordinary Differential Equations: the Initial Value Problem, Freeman, 1975.

[27] W.-J. Ying, A multilevel adaptive approach for computational cardiology, Ph.D. thesis, Department of Mathematics, Duke University (2005).