BIT Numerical Mathematics (2022) 62:139–170
https://doi.org/10.1007/s10543-021-00875-y

BIT

Arbitrary high order A-stable and B-convergent numerical methods for ODEs via deferred correction

Saint-Cyr E. R. Koyaguerebo-Imé[1] · **Yves Bourgault**[1]

Received: 24 February 2020 / Accepted: 16 April 2021 / Published online: 29 April 2021

Abstract

This paper presents a sequence of deferred correction (DC) schemes built recursively from the implicit midpoint scheme for the numerical solution of general first order ordinary differential equations (ODEs). It is proven that each scheme is A-stable, satisfies a B-convergence property, and that the correction on a scheme DC2j of order 2j of accuracy leads to a scheme DC2j + 2 of order 2j + 2. The order of accuracy is guaranteed by a deferred correction condition. Numerical experiments with standard stiff and non-stiff ODEs are performed with the DC2, …, DC10 schemes. The results show a high accuracy of the method. The theoretical orders of accuracy are achieved together with a satisfactory stability.

Keywords Ordinary differential equations · High order time-stepping methods · Deferred correction · A-stability · B-convergence

Mathematics Subject Classification 65B05 · 65L04 · 65L05 · 65L12 · 65L20

Communicated by Antonella Zanna Munthe-Kaas.

The authors would like to acknowledge the financial support of the Discovery Grant Program of the Natural Sciences and Engineering Research Council of Canada (NSERC) and a scholarship to the first author from the NSERC CREATE program "Génie par la Simulation".

✉ Yves Bourgault
ybourg@uottawa.ca

Saint-Cyr E. R. Koyaguerebo-Imé
skoya005@uottawa.ca

[1] Department of Mathematics and Statistics, University of Ottawa, STEM Complex, 150 Louis-Pasteur Pvt, Ottawa, ON K1N 6N5, Canada

1 Introduction

In [10,20], Gustafsson and Kress introduced a new version of deferred correction (DC) strategy for the numerical solution of linear systems of ordinary differential equations (ODE) [10] and initial boundary value problems [20], under a monotonicity condition. Numerical experiments with one-dimensional linear parabolic and hyperbolic equations were performed and showed that the method is effective (orders 2, 4 and 6 of accuracy are achieved). We propose to extend the method from [10,20] to the time-discretization of more general time-evolution partial differential equations (PDEs). In this paper, we restrict to the case of the initial value problem (IVP)

$$\begin{cases} \dfrac{du}{dt} = F(t,u), \quad t \in [0,T], \\ u(0) = \quad u_0, \end{cases} \tag{1.1}$$

where the unknown u is from $[0,T]$ into a Banach space X, u_0 is a given data and F is a sufficiently differentiable function such that u exists and is sufficiently differentiable. The main objective is to show the properties of the numerical method (consistency, stability, convergence and order of accuracy). A complete analysis of the DC method applied to reaction-diffusion equations leads to an arbitrary high order and unconditionally stable method (see [18]).

The DC method is used to improve the order of accuracy of numerical methods of lower order. This method is explored by many authors, e.g. [1,2,6,7,10,12,21,23]. The method in [6] is an application of iterative deferred correction (IDC). The authors proved that an asymptotic improvement of order p can be accomplished, from a scheme of order p, at each step of the IDC procedure, provided suitable finite difference operators are employed. Numerical experiments are performed with the IDC applied to the trapezoidal rule, Taylor-2 and Adams-Bashforth of order 2. The results are promising even though they point out some difficulties of the proposed algorithms: inaccuracy for "large" time step and no asymptotic improvement for high levels of correction. The approaches in [1,2,7,10,12,21] are quite similar and consist in a linear perturbation of a low order scheme. However, solving stiff problems (problems extremely hard to solve by standard explicit methods [25]) is a challenge unfavorable for these methods. In particular, the method in [21], concerning a highly accurate solver for stiff ODEs, requires sufficiently small time steps for moderately stiff problems while convergence is reduced to order 2 for "very stiff" problems.

Our schemes are based on nonlinear perturbations (corrections) of the implicit midpoint rule and inherit the A-stable property of the trapezoidal rule [5] at any stage of the correction. Starting from an approximation $\left\{u^{2,n}\right\}_{n=0}^{N}$ of the exact solution u by the implicit midpoint rule on a uniform partition $0 = t_0 < t_1 < \cdots < t_N = T$ of $[0,T]$, at the stage $j = 1, 2, \ldots$ of the correction we obtain an approximation $\left\{u^{2j+2,n}\right\}_{n=0}^{N}$ of u, expected to be of order $2j+2$ of accuracy, on the same partition. Each approximate solution $\left\{u^{2j,n}\right\}_{n=0}^{N}$ to be corrected is subject to a deferred correction condition (DCC) which guarantees the improvement of the order of accuracy. We prove that if $\left\{u^{2j,n}\right\}_{n=0}^{N}$ satisfies the DCC and its correction $\left\{u^{2j+2,n}\right\}_{n=0}^{N}$ converges to u at the

discrete points $0 = t_0 < t_1 < \cdots < t_N = T$ (or is simply bounded, when X is finite dimensional) then $\left\{u^{2j+2,n}\right\}_{n=0}^{N}$ approximates u with order $2j+2$. Moreover, provided the function F is Lipschitz with respect to its second variable or satisfies a one-sided Lipschitz condition, each $\left\{u^{2j,n}\right\}_{n=0}^{N}$ satisfies the DCC and then converges with order $2j$ of accuracy, for arbitrary positive integer j. We also prove that each DC scheme involving $\left\{u^{2j,n}\right\}_{n=0}^{N}$ is B-stable. The theory is illustrated by numerical tests, for the schemes of order 2, 4, …, 10.

The paper is organized as follows: in Sect. 2 we recall some basic results from finite difference approximations and present the DC schemes; Sect. 3 deals with the consistency of the method; the analysis of convergence and order of accuracy together with a B-convergence result are given in Sect. 4; absolute stability is proved in Sect. 5, and Sect. 6 is devoted to numerical experiments.

2 Deferred correction schemes for the implicit midpoint rule

We suppose that $F \in C^{2p+2}([0,T]\times X, X)$, for a positive integer p, so that (1.1) has a unique solution $u \in C^{2p+3}([0,T],X)$. We simply denote by $\|\cdot\|$, the norm in the Banach space X. For a time step $k>0$, we denote $t_n = nk$ and $t_{n+1/2} = (n+1/2)k$, for each integer n. This implies that $t_0 = 0$. We consider the time steps k such that $0 = t_0 < t_1 < \cdots < t_N = T$ is a partition of $[0,T]$, for a non-negative integer N. The centered, forward and backward difference operators D, D_+ and D_-, respectively, related to k and applied to u, are defined as follows:

$$Du(t_{n+1/2}) = \frac{u(t_{n+1}) - u(t_n)}{k},$$
$$D_+u(t_n) = \frac{u(t_{n+1}) - u(t_n)}{k},$$

and

$$D_-u(t_n) = \frac{u(t_n) - u(t_{n-1})}{k}, n \geq 1.$$

The average operator is denoted by E:

$$Eu(t_{n+1/2}) = \widehat{u}(t_{n+1}) = \frac{u(t_{n+1}) + u(t_n)}{2}.$$

The composition of D_+ and D_- is defined recursively. They commute, that is $(D_+D_-)u(t_n) = (D_-D_+)u(t_n) = D_-D_+u(t_n)$, and satisfy the identities

$$(D_+D_-)^m u(t_n) = k^{-2m}\sum_{i=0}^{2m}(-1)^i\binom{2m}{i}u(t_{n+m-i}), \tag{2.1}$$

and

$$D_-(D_+D_-)^m u(t_n) = k^{-2m-1} \sum_{i=0}^{2m+1} (-1)^i \binom{2m+1}{i} u(t_{n+m-i}), \tag{2.2}$$

for each integer $m \geq 1$ such that $0 \leq t_{n-m-1} \leq t_{n+m} \leq T$. We have the estimate

$$\left\| D_+^{m_1} D_-^{m_2} u(t_n) \right\| \leq \max_{0 \leq t \leq T} \left\| \frac{d^{m_1+m_2} u}{dt^{m_1+m_2}}(t) \right\|, \tag{2.3}$$

provided $[t_{n-m_2}, t_{n+m_1}] \subset [0, T]$ and $m_1 + m_2 \leq 2p + 3$ (see [15, p. 249] or [17]).

If $\{u^n\}_n$ is a sequence of approximation of u at the discrete points t_n, the finite difference operators apply to $\{u^n\}_n$, and we define

$$Du^{n+1/2} = D_+u^n = D_-u^{n+1} = \frac{u^{n+1} - u^n}{k},$$

and

$$Eu^{n+1/2} = \widehat{u}^{n+1} = \frac{u^{n+1} + u^n}{2}.$$

From the centered finite difference approximation (see [17, Thm 5] or [3,4,13]) we have

$$\frac{du}{dt}(t_{n+1/2}) = \frac{u(t_{n+1}) - u(t_n)}{k} - \sum_{i=1}^{j} c_{2i+1} k^{2i} (D_+D_-)^i Du(t_{n+1/2})) + O(k^{2j+2}) \tag{2.4}$$

and

$$u(t_{n+1/2}) = \frac{u(t_{n+1}) + u(t_n)}{2} - \sum_{i=1}^{j} c_{2i} k^{2i} (D_+D_-)^i Eu(t_{n+1/2}) + O(k^{2j+2}), \tag{2.5}$$

for each integer $j = 1, 2, \ldots, p$. These approximations lead to the schemes

$$\begin{aligned} \frac{u^{n+1} - u^n}{k} - \sum_{i=1}^{j} c_{2i+1} k^{2i} (D_+D_-)^i Du^{n+1/2} \\ = F\left(t_{n+1/2}, \frac{u^{n+1} + u^n}{2} - \sum_{i=1}^{j} c_{2i} k^{2i} (D_+D_-)^i Eu^{n+1/2} \right). \end{aligned} \tag{2.6}$$

The schemes (2.6) are multi-steps and prone to stability restrictions. We resort to DC method to transform them into a sequence of one step schemes as follows: For $j = 0$,

we have the implicit midpoint rule

$$\frac{u^{2,n+1}-u^{2,n}}{k}=F\left(t_{n+1/2},\frac{u^{2,n+1}+u^{2,n}}{2}\right),\quad u^{2,0}=u_0. \tag{2.7}$$

For $j\geq 1$,

$$\begin{aligned}&\frac{u^{2j+2,n+1}-u^{2j+2,n}}{k}-\sum_{i=1}^{j}c_{2i+1}k^{2i}(D_+D_-)^iDu^{2j,n+1/2}\\&=F\left(t_{n+1/2},\frac{u^{2j+2,n+1}+u^{2j+2,n}}{2}-\sum_{i=1}^{j}c_{2i}k^{2i}(D_+D_-)^iEu^{2j,n+1/2}\right),\end{aligned} \tag{2.8}$$

$$u^{2j+2,0}=u_0. \tag{2.9}$$

The scheme (2.8)–(2.9) has unknowns $u^{2j+2,n}$, $n=1,2,\ldots,N$, and is deduced from (2.6) by substituting the unknown u^n under the summation symbols by $u^{2j,n}$. The index $2j$ indicates that $\left\{u^{2j,n}\right\}_n$ is expected to be an approximation of the exact solution u with order $2j$ of accuracy. We call the schemes (2.8)–(2.9) Deferred Correction of order $2j+2$ for the implicit midpoint rule, denoted DC2j+2.

Remark 2.1 The scheme (2.8)–(2.9), for $n=1,2,3,\ldots,j$, should involve unknowns $u^{2j,-1},\ldots,u^{2j,-j}$ which represent approximate solutions of (1.1) at the discrete points $t=-k,\ldots,-jk$, respectively. To avoid those approximations for $t<0$, we propose the following scheme which is efficient for the computation of $u^{2j+2,1},\ldots,u^{2j+2,j}$, using only points within the solution interval $[0,T]$.

$$\begin{aligned}&\frac{u^{2j+2,n+1}-u^{2j+2,n}}{k}-k^{-1}\sum_{i=1}^{j}c_{2i+1}^{j}k_j^{2i+1}(D_+D_-)^iD\bar{u}^{2j,(2j+1)n+j+1/2}\\&=F\left(t_{n+1/2},Eu^{2j+2,n+1/2}-\sum_{i=1}^{j}c_{2i}^{j}k_j^{2i}(D_+D_-)^iE\bar{u}^{2j,(2j+1)n+j+1/2}\right),\end{aligned} \tag{2.10}$$

$$u^{2j+2,0}=u_0. \tag{2.11}$$

The finite difference operator in (2.10) are related to the time step $k_j=k/(2j+1)$. The approximations $\left\{\bar{u}^{2j,m}\right\}_m$ and $\left\{u^{2j,n}\right\}_n$ are computed from the same scheme, (2.7) or (2.8)–(2.9), but for the time steps k_j and k, respectively. The scheme (2.10) results from the finite difference approximations

$$u'(t_{n+1/2})=\frac{u(t_{n+1})-u(t_n)}{k}-\frac{1}{k}\sum_{i=1}^{j}c_{2i+1}^{j}k_j^{2i+1}D(D_+D_-)^iu(\tau_{j+1/2})+O(k_j^{2j+2}) \tag{2.12}$$

Table 1 Coefficients of the approximations (2.12)–(2.13) for $j = 1, 2, 3, 4$

j	c_2^j	c_3^j	c_4^j	c_5^j	c_6^j	c_7^j	c_8^j	c_9^j
1	$\frac{9}{8}$	$\frac{9}{8}$						
2	$\frac{25}{8}$	$\frac{125}{24}$	$\frac{125}{128}$	$\frac{125}{128}$				
3	$\frac{49}{8}$	$\frac{343}{24}$	$\frac{637}{128}$	$\frac{13,377}{1920}$	$\frac{1029}{1024}$	$\frac{1029}{1024}$		
4	$\frac{81}{8}$	$\frac{243}{8}$	$\frac{1917}{128}$	$\frac{17,253}{640}$	$\frac{7173}{1024}$	$\frac{64,557}{7168}$	$\frac{32,733}{32,768}$	$\frac{32,733}{32,768}$

and

$$u(t_{n+1/2}) = \frac{u(t_{n+1}) + u(t_n)}{2} - \sum_{i=1}^{j} c_{2i}^j k_j^{2i} (D_+ D_-)^i E u(\tau_{j+1/2}) + O(k_j^{2j+2}), \quad (2.13)$$

where $t_n = \tau_0 < \tau_1 < \ldots < \tau_{2j+1} = t_{n+1}$, with $\tau_m = t_n + mk_j$, for $m = 1, 2, \ldots, 2j+1$. Table 1 gives the coefficients c_i^j for $j = 1, 2, 3, 4$.

Remark 2.2 Each $u^{2j+2,n+1}$, $n \geq j$, can be obtained by solving iteratively the system

$$x - a_n^j - kF(t_{n+1/2}, 0.5x + b_n^j) = 0, \quad (2.14)$$

where x is the unknown, and a_n^j and b_n^j are constants depending on $u^{2j+2,n}$ and $u^{2j,n+1+j}, u^{2j,n+j}, \ldots, u^{2j,n-j}$. The total number of vectors (in the solution space X) stored for the computation of $u^{2j+2,n+1}$ is $j^2 + 3j + 1$: $u^{2j+2,n}$ and the $u^{2i,q}$, for $i = 1, 2, \ldots, j$, and $n + (j-i+1)(j+i)/2 - 2i \leq q \leq n + 1 + (j-i+1)(j+i)/2$.

Remark 2.3 From Remark 2.2, only the implicit midpoint rule, DC2, and the DC schemes of the form (2.10)–(2.11) used at startup are implicit Runge-Kutta (RK) methods. Starting with DC4, all the DC2j methods of the form (2.8)–(2.9) are not RK methods. For instance, $u^{4,n+1}$ depends on $u^{4,n}$ and some of the $u^{2,i}$, which $u^{2,i}$ evolve independently and are not stages computed from $u^{4,n}$. As we will see in Sect. 5, the analysis of A-stability, in particular the proof of lemma 5.2, shows that it is impossible to write a recurrence $u^{2j+2,n+1} = R(z)\, u^{2j+2,n}$ from (2.8) when $j \geq 1$, as one would get by applying any RK method to Dahlquist equation. This is the main ingredient behind the A-stability of our DC2j methods independently of the order of accuracy.

3 Deferred correction condition (DCC)

In this section we give a sufficient condition for the scheme (2.8)–(2.9) to achieve order $2j+2$ of accuracy. Hereafter, the letter C will denote any constant independent from k, and that can be calculated explicitly in terms of known quantities. The exact value of C may change. We have the following definition:

Definition 3.1 (*Deferred Correction Condition*) Let u be the exact solution of the Cauchy problem (1.1). Given a positive integer j, a sequence $\left\{u^{2j,n}\right\}_{n=0}^{N}$ of approximations of u, at the discrete points $0 = t_0 < \cdots < t_N = T$, is said to satisfy the Deferred Correction Condition (DCC) for the implicit midpoint rule if $\left\{u^{2j,n}\right\}_{n=0}^{N}$ approximates u with order $2j$ of accuracy, and we have

$$\|(D_+D_-)D(u^{2j,n+1/2}-u(t_{n+1/2}))\|+\|D_+D_-(u^{2j,n+1}-u(t_{n+1}))\| \leq Ck^{2j}, \quad (3.1)$$

for $n = 1, 2, \ldots, N-2$ and $k \leq k_0$, where $k_0 > 0$ is fixed and C is a constant independent from k.

Remark 3.1 Condition (3.1) is equivalent to

$$\left\| \sum_{i=1}^{j} c_{2i} k^{2i} \left(D_+D_-\right)^i \left(u^{2j,n} - u(t_n)\right) \right\| \leq Ck^{2j+2}, \quad (3.2)$$

and

$$\left\| \sum_{i=1}^{j} (c_{2i+1} - c_{2i}) k^{2i} \left(D_+D_-\right)^i D \left(u^{2j,n+1/2} - u(t_{n+1/2})\right) \right\| \leq Ck^{2j+2}, \quad (3.3)$$

for $n = j, j+1, \ldots, N-j$. This is due to the transform

$$k^{2i} \left(D_+D_-\right)^i \left(u^{2j,n} - u(t_n)\right) = k^2 \sum_{l=0}^{i-1} (-1)^l \binom{2i-2}{l} D_+D_- \left(u^{2j,n} - u(t_n)\right)$$

and a similar transform for $k^i \left(D_+D_-\right)^i D \left(u^{2j,n+1/2} - u(t_{n+1/2})\right)$.

We have the following result:

Theorem 3.1 *Let u be the exact solution of (1.1) and $\left\{u^{2j,n}\right\}_{n=0}^{N}$, $j = 1, \ldots, p$, a sequence of approximations of u satisfying DCC for the implicit midpoint rule. Let $\left\{u^{2j+2,n}\right\}_{n=0}^{N}$ be the solution of (2.8)–(2.9) built from $\left\{u^{2j,n}\right\}_{n=0}^{N}$. We suppose that $u^{2j+2,1}, \ldots, u^{2j+2,j}$ are given and satisfy*

$$\|u^{2j+2,n} - u(t_n)\| \leq Ck^{2j+2}, \quad \text{for } n = 1, 2, \ldots, j, \quad (3.4)$$

where C is a constant independent from k. Furthermore, we suppose that one of the following four conditions holds:

(i) *F is Lipschitz with respect to the second variable x: there exists $\mu \geq 0$ such that*

$$\|F(t,x) - F(t,y)\| \leq \mu\|x - y\|, \quad \forall (t,x,y) \in [0,T] \times X \times X. \quad (3.5)$$

(ii) *X is finite dimensional, and $\left\{u^{2j+2,n}\right\}_{n=0}^{N}$ remains close to u in the sense that there exists $M > 0$ such that*

$$\|u^{2j+2,n} - u(t_n)\| \leq M, \quad \text{for each } n = 0, 1, \ldots, N. \tag{3.6}$$

(iii) *X is infinite dimensional, and $\left\{u^{2j+2,n}\right\}_n$ converges to the exact solution u.*

(iv) *X is a Hilbert space with inner product $(.,.)$, and F satisfies the following so-called one-sided Lipschitz condition, with a one-sided Lipschitz constant $\beta \in \mathbb{R}$:*

$$(F(t,x) - F(t,y), x - y) \leq \beta \|x - y\|^2, \quad \forall (t,x,y) \in [0,T] \times X \times X. \tag{3.7}$$

Then $\left\{u^{2j+2,n}\right\}_n$ approximates u with order $2j+2$ of accuracy, that is

$$\|u^{2j+2,n} - u(t_n)\| \leq C k^{2j+2}, \quad \text{for each } n = 0, 1, \ldots, N, \tag{3.8}$$

where C is a constant depending only on j, T, DCC, a Lipschitz constant on F and the derivatives of u up to order $2j+3$, for time steps k sufficiently small.

Proof 1. First we consider the case where the function $F = F(t,x)$ is Lipschitz with respect to the second variable x. Combining (1.1) and (2.8), we obtain the identity

$$\begin{aligned} D\Theta^{2j+2,n+1/2} &= \sigma^{2j+2,n+1/2} + (\Lambda^j - \Gamma^j) D \left(u^{2j,n+1/2} - u(t_{n+1/2})\right) \\ &+ F\left(t_{n+1/2}, \widehat{u}^{2j+2,n+1} - \Gamma^j \widehat{u}^{2j,n+1}\right) - F\left(t_{n+1/2}, \widehat{u}(t_{n+1}) - \Gamma^j \widehat{u}(t_{n+1})\right), \end{aligned} \tag{3.9}$$

where Λ^j and Γ^j are finite difference operators defined for arbitrary integer $j \geq 1$ by

$$\Lambda^j u(t_n) = \sum_{i=1}^{j} c_{2i+1} k^{2i} (D_+ D_-)^i u(t_n),$$

and

$$\Gamma^j u(t_n) = \sum_{i=1}^{j} c_{2i} k^{2i} (D_+ D_-)^i u(t_n),$$

provided $u(t_{n\pm i})$ exists for $i = 0, 1, 2, \ldots, j$. We have defined

$$\Theta^{2j+2,n} = \left(u^{2j+2,n} - u(t_n)\right) - \Gamma^j \left(u^{2j,n} - u(t_n)\right), \tag{3.10}$$

and

$$\sigma^{2j+2,n+1/2} = \left[u'(t_{n+1/2}) - Du(t_{n+1/2}) + \Lambda^j Du(t_{n+1/2})\right] - \left[F(t_{n+1/2}, u(t_{n+1/2})) - F(t_{n+1/2}, \widehat{u}(t_{n+1}) - \Gamma^j \widehat{u}(t_{n+1}))\right].$$

From (2.4) we have

$$\left\| u'(t_{n+1/2}) - Du(t_{n+1/2}) + \Lambda^j Du(t_{n+1/2}) \right\| \leq Ck^{2j+2},$$

and, since F is differentiable and u is sufficiently regular, we deduce from the mean value theorem and the approximation (2.5) that

$$\left\| F(t_{n+1/2}, u(t_{n+1/2})) - F(t_{n+1/2}, \widehat{u}(t_{n+1}) - \Gamma^j \widehat{u}(t_{n+1})) \right\| \leq Ck^{2j+2},$$

for each $n = 0, 1, \ldots, N$, where C is a constant depending only on j, T, a Lipschitz constant from F and the derivatives of u up to order $2j + 3$. The last two inequalities imply that

$$\left\| \sigma^{2j+2,n+1/2} \right\| \leq Ck^{2j+2}. \tag{3.11}$$

Since the sequence $\left\{u^{2j,n}\right\}_n$ satisfies DCC, from Remark 3.1 we have

$$\left\| \left(\Lambda^j - \Gamma^j\right) D\left(u^{2j,n+1/2} - u(t_{n+1/2})\right) \right\| \leq Ck^{2j+2}. \tag{3.12}$$

From the Lipschitz condition on F we have

$$\left\| F\left(t_{n+1/2}, \widehat{u}^{2j+2,n+1} - \Gamma^j \widehat{u}^{2j,n+1}\right) - F\left(t_{n+1/2}, \widehat{u}(t_{n+1}) - \Gamma^j \widehat{u}(t_{n+1})\right) \right\| \leq \mu \|\widehat{\Theta}^{2j+2,n+1}\|. \tag{3.13}$$

Substituting inequalities (3.11)–(3.13) in the identity (3.9), we deduce that

$$\|D\Theta^{2j+2,n+1/2}\| \leq Ck^{2j+2} + \mu \|\widehat{\Theta}^{2j+2,n+1}\|,$$

and it follows from the triangle inequality that

$$\|\Theta^{2j+2,n+1}\| \leq C\frac{k^{2j+3}}{2-\mu k} + \frac{2+\mu k}{2-\mu k}\|\Theta^{2j+2,n}\|,$$

for $0 \leq \mu k < 2$. We then deduce by induction on n that

$$\|\Theta^{2j+2,n}\| \leq C\frac{1}{2-\mu k}\left(\frac{2+\mu k}{2-\mu k}\right)^{n-j-1} k^{2j+2} + \left(\frac{2+\mu k}{2-\mu k}\right)^{n-j} \|\Theta^{2j+2,j}\|. \tag{3.14}$$

From hypothesis (3.4) and the DCC we have

$$\|\Theta^{2j+2,j}\| \leq \|u^{2j+2,j} - u(t_j)\| + \left\|\Gamma^j(u^{2j,j} - u(t_j))\right\| \leq Ck^{2j+2}, \quad (3.15)$$

where C is a constant independent from k. Moreover, the sequence $\left\{\left(\frac{2+\mu k}{2-\mu k}\right)^n\right\}_n$ is bounded above by $\exp(2\mu T/(2-\varepsilon))$, for $0 \leq \mu k \leq \varepsilon < 2$. Whence

$$\|\Theta^{2j+2,n}\| \leq Ck^{2j+2}.$$

Finally, by the triangle inequality, identity (3.10) and DCC, we have

$$\|u^{2j+2,n} - u(t_n)\| \leq \|\Theta^{2j+2,n}\| + \left\|\Gamma^j(u^{2j,n} - u(t_n))\right\| \leq Ck^{2j+2},$$

where C is a constant depending only on j, T, the DCC constant, μ and the derivatives of u up to order $2j+3$.

2. Suppose that $\left\{u^{2j+2,n}\right\}_{n=0}^N$ satisfies (3.6) and X is finite dimensional. We can write

$$F\left(t_{n+1/2}, \widehat{u}^{2j+2,n+1} - \Gamma^j\widehat{u}^{2j,n+1}\right) - F\left(t_{n+1/2}, \widehat{u}(t_{n+1}) - \Gamma^j\widehat{u}(t_{n+1})\right)$$
$$= \int_0^1 d_u F\left(t_{n+1/2}, \widehat{u}(t_{n+1}) - \Gamma^j\widehat{u}(t_{n+1}) + s\widehat{\Theta}^{2j+2,n+1}\right)\left(\widehat{\Theta}^{2j+2,n+1}\right) ds.$$

From (3.6) and the DCC there exists $k_1 > 0$ such that $0 < k \leq k_1 \leq k_0$ implies

$$\|\widehat{\Theta}^{2j+2,n+1}\| \leq M + Ck^{2j+2} \leq M + 1.$$

On the other hand, we have

$$\|\widehat{u}(t_{n+1}) - \Gamma^j\widehat{u}(t_{n+1})\| = \left\|\widehat{u}(t_{n+1}) - \sum_{i=1}^{j}\sum_{l=0}^{2i}(-1)^l c_{2i}\binom{2i}{l}u(t_{n+i-l})\right\| \leq R_{j+1}, \quad (3.16)$$

where

$$R_{j+1} := (j+1)\max_{0\leq t\leq T}\|u(t)\| \geq \left(1 + \sum_{i=1}^{j} 2^{2i}|c_{2i}|\right)\max_{0\leq t\leq T}\|u(t)\|. \quad (3.17)$$

It follows (3.13) for

$$\mu = \sup_{0\leq t\leq T, \|x\|\leq M+R_{j+1}+1} \|d_x F(t,x)\|.$$

Since F is differentiable and the set $\left\{x \in X : \|x\| \leq M + R_{j+1} + 1\right\}$ is compact in the finite dimensional linear space X, the supremum exists and is finite. The theorem is then deduced from the case (i).

3. If $\left\{u^{2j+2,n}\right\}_n$ converges to the exact solution u, taking the DDC and the finite difference formula (2.5) into account, we have

$$\left(\widehat{u}(t_{n+1}) - \Gamma^j \widehat{u}(t_{n+1}) + s\widehat{\Theta}^{2j+2,n+1}\right) - u(t_{n+1/2}) \to 0, \text{ as } k \to 0, \text{ for } 0 \leq s \leq 1.$$

It follows from the continuity of $u \mapsto d_u F(t,u)$ that there exists $0 < k_2 \leq k_0$ such that $0 < k \leq k_2$ implies

$$\|d_u F(t_{n+1/2}, \widehat{u}(t_{n+1}) - \Gamma\widehat{u}(t_{n+1}) + \tau\widehat{\Theta}^{2j+2,n+1})\| \leq 1 + \max_{0\leq t\leq T} \|d_u F\,(t, u(t))\,\|.$$

The theorem, in this case, follows by taking $\mu = 1 + \max_{0\leq t\leq T} \|d_u F\,(t, u(t))\,\|$ in (i).

4. Here we consider the case where X is a Hilbert space and F satisfies the monotonicity condition (3.7). Then, taking the inner product of the identity (3.9) with $\widehat{\Theta}^{2j+2,n+1}$, we deduce the inequality

$$\left(D\Theta^{2j+2,n+1/2}, \widehat{\Theta}^{2j+2,n+1}\right) \leq \left(\sigma^{2j+2,n+1/2}, \widehat{\Theta}^{2j+2,n+1}\right) + \beta\|\widehat{\Theta}^{2j+2,n+1}\|^2 \left((\Lambda^j - \Gamma^j)D(u^{2j,n+1/2} - u(t_{n+1/2})), \widehat{\Theta}^{2j+2,n+1}\right) \tag{3.18}$$

since, according to (3.7), we have

$$\left(F\left(t_{n+1/2}, \widehat{u}^{2j+2,n+1} - \Gamma\widehat{u}^{2j,n+1}\right) - F\left(t_{n+1/2}, \widehat{u}(t_{n+1}) - \Gamma\widehat{u}(t_{n+1})\right), \widehat{\Theta}^{2j+2,n+1}\right) \leq \beta\left\|\widehat{\Theta}^{2j+2,n+1}\right\|^2.$$

Inequalities (3.11)–(3.12) together with the Cauchy-Schwartz inequality yield

$$\left|\left(\sigma^{2j+2,n+1/2}, \widehat{\Theta}^{2j+2,n+1}\right)\right| \leq Ck^{2j+2}\|\widehat{\Theta}^{2j+2,n+1}\|,$$

and

$$\left|\left((\Lambda^j - \Gamma^j)D(u^{2j,n+1/2} - u(t_{n+1/2})), \widehat{\Theta}^{2j+2,n+1}\right)\right| \leq Ck^{2j+2}\|\widehat{\Theta}^{2j+2,n+1}\|,$$

where C is a constant depending only on j, T, a Lipschitz constant on F and the derivatives of u up to order $2j+3$. Substituting the last three inequalities into (3.18), we obtain

$$\left(D\Theta^{2j+2,n+1/2}, \widehat{\Theta}^{2j+2,n+1}\right) \leq Ck^{2j+2}\|\widehat{\Theta}^{2j+2,n+1}\| + \beta\|\widehat{\Theta}^{2j+2,n+1}\|^2,$$

and we deduce from the identity

$$\left(D\Theta^{2j+2,n+1/2}, \widehat{\Theta}^{2j+2,n+1}\right) = \frac{1}{2k}\left(\|\Theta^{2j+2,n+1}\|^2 - \|\Theta^{2j+2,n}\|^2\right)$$

and the inequality

$$\|\widehat{\Theta}^{2j+2,n+1}\| \leq \frac{1}{2}\left(\|\Theta^{2j+2,n+1}\| + \|\Theta^{2j+2,n}\|\right)$$

that

$$\|\Theta^{2j+2,n+1}\| \leq C\frac{k^{2j+3}}{2-\beta k} + \frac{2+\beta k}{2-\beta k}\|\Theta^{2j+2,n}\|.$$

The conclusion follows from the case (i), for $-2 \leq \beta k < 2$.

□

Remark 3.2 Theorem 3.1 shows that the correction may be applied for any other scheme satisfying DCC.

4 Convergence and order of accuracy

The aim of this section is to prove the following theorem:

Theorem 4.1 *Let $u \in C^{2p+3}([0,T],X)$ be the exact solution of the problem (1.1). Suppose that one of the four conditions (i)–(iv) of Theorem 3.1 holds, with condition (ii) or (iii) holding for all $j = 0,1,\ldots,p+1$. Then each sequence $\left\{u^{2j,n}\right\}_{n=0}^{N}$, $j = 1,2,\ldots,p+1$, solution of the scheme (2.7) or (2.8)–(2.9), approximates u with order $2j$ of accuracy. Furthermore, we have the estimate*

$$\|(D_+D_-)^m D(u^{2j,n+1/2} - u(t_{n+1/2}))\| + \|(D_+D_-)^m(u^{2j,n+1} - u(t_{n+1}))\| \leq Ck^{2j} \tag{4.1}$$

for $m = 0,1,\ldots,p-j$ and $n = m+j-1, m+j,\ldots,N-j-m$, where C is a constant depending only on p, T, and the derivatives of u and F up to order $2m+2j+1$ and $2m+2j-1$, respectively.

To prove this theorem we need Theorem 3.1 and the the following lemma:

Lemma 4.1 *Let $\left\{u^{2,n}\right\}_{n=0}^{N}$ be the solution of the scheme (2.7). Suppose that one of the conditions (i), (iii) or (iv) of Theorem 3.1 holds, or $\left\{u^{2,n}\right\}_{n=0}^{N}$ is bounded in the sense of the condition (ii) of this theorem. Then $\left\{u^{2,n}\right\}_{n=0}^{N}$ approximates u with order 2 of accuracy, and we have the inequality*

$$\|(D_+D_-)^m D(u^{2,n+1/2} - u(t_{n+1/2}))\| + \|(D_+D_-)^m(u^{2,n+1} - u(t_{n+1}))\| \leq Ck^2, \tag{4.2}$$

for $m = 0, 1, \ldots, p$ and $n = m, m+1, \ldots, N-m-1$, where C is a constant depending only on p, T, and the derivatives of u and F up to order $2m+3$ and $2m+1$, respectively.

Proof (Proof of Lemma 4.1) For the sake of simplification we suppose that $F = F(x)$. The general case can be handled by transforming (1.1) to an autonomous system. From the hypotheses of the Lemma, Theorem 3.1 implies that $\left\{u^{2,n}\right\}_{n=0}^{N}$ approximates u with order two of accuracy:

$$\|u(t_n) - u^{2,n}\| \leq Ck^2, \text{ for each } n = 0, 1, 2, \ldots, N, \tag{4.3}$$

where C is a constant depending only on T, F and the derivatives of u up to order 3. To establish (4.2) we proceed by induction on the integer $m = 0, 1, \ldots, p$.

1. Inequality (4.2) for $m = 0$.
   As in Theorem 3.1, we combine (1.1) and (2.7) and deduce the identity

$$D\Theta^{2,n+1/2} = \left[F\left(\widehat{u}^{2,n+1}\right) - F\left(\widehat{u}(t_{n+1})\right)\right] + \sigma^{2,n+1/2}, \tag{4.4}$$

   where

$$\Theta^{2,n} = u^{2,n} - u(t_n),$$

   and

$$\sigma^{2,n+1/2} = \left[u'(t_{n+1/2}) - Du(t_{n+1/2})\right] - \left[F\left(u(t_{n+1/2}\right) - F\left(\widehat{u}(t_{n+1})\right)\right].$$

   From Taylor's formula with integral remainder and the estimate (2.3), there exists a function g such that

$$\sigma^{2,n+1/2} = k^2 g(t_{n+1}),$$

   with

$$\|D_+^{m_1} D_-^{m_2} g(t_{n+1})\| \leq C, \text{ for } m_2 - 1 \leq n \leq N - m_1 - 1, \tag{4.5}$$

   for each nonnegative integers m_1 and m_2 such that $m_1 + m_2 \leq 2p$, where C is a constant depending only on T, F, and the derivatives of u up to order $m_1 + m_2 + 3$. We can write

$$F\left(\widehat{u}^{2,n+1}\right) - F\left(\widehat{u}(t_{n+1})\right) = \int_0^1 dF\left(K_1^{n+1}\right)\left(\widehat{\Theta}^{2,n+1}\right) d\tau_1,$$

   where

$$K_1^{n+1} = \widehat{u}(t_{n+1}) + \tau_1 \widehat{\Theta}^{2,n+1}.$$

The last identities substituted into (4.4) yield

$$D\Theta^{2,n+1/2} = \int_0^1 dF\left(K_1^{n+1}\right)(\widehat{\Theta}^{2,n+1})d\tau_1 + k^2 g(t_{n+1}). \tag{4.6}$$

Proceeding as in Theorem 3.1, we deduce from (4.3) and the regularity of u that

$$\left\| \int_0^1 dF\left(K_1^{n+1}\right)(\widehat{\Theta}^{2,n+1})d\tau_1 \right\| \leq C\|\widehat{\Theta}^{2,n+1}\|.$$

Therefore, taking the norm on both sides of (4.6), we deduce by the triangle inequality and the inequalities (4.3) and (4.5), for $m_1 = m_2 = 0$, that

$$\|D\Theta^{2,n+1/2}\| \leq C\|\widehat{\Theta}^{2,n+1}\| + k^2\|g(t_{n+1})\| \leq Ck^2, \tag{4.7}$$

where C is a constant depending only on T and the derivatives of u and F up to order 3 and 1, respectively. The last inequality combined with (4.3) implies that (4.2) holds for $m = 0$.

2. Here we are going to prove that inequality (4.2) remains true for $m+1$, assuming that it holds for an arbitrary integer m such that $0 \leq m \leq p-1$.
We apply $(D_+D_-)^m D_+$ to (4.6) and obtain

$$(D_+D_-)^{m+1}\,\Theta^{2,n+1} = (D_+D_-)^m\,D_+h(t_{n+1}) + k^2\,(D_+D_-)^m\,D_+g(t_{n+1}), \tag{4.8}$$

where we set

$$h(t_{n+1}) = \int_0^1 dF\left(K_1^{n+1}\right)(\widehat{\Theta}^{2,n+1})d\tau_1.$$

The main difficulty is to bound $(D_+D_-)^m\,D_+h(t_{n+1}) = D_+^{2m+1}h(t_{n+1-m})$. We have

$$\begin{aligned}
D_+h(t_n) &= \int_0^1 dF\left(K_1^{n+1}\right)(D_+\widehat{\Theta}^{2,n})d\tau_1 + \int_0^1\int_0^1 d^2F\left(K_2^n\right)\left(D_+K_1^n, \widehat{\Theta}^{2,n}\right)d\tau_1 d\tau_2,\\
D_+^2h(t_n) &= \int_0^1 dF(K_1^{n+2})(D_+^2\widehat{\Theta}^{2,n})d\tau_1 + \int_0^1\int_0^1 d^2F(K_2^{n+1})(D_+K_1^{n+1}, D_+\widehat{\Theta}^{2,n})d\tau^2\\
&\quad + \int_0^1\int_0^1 d^2F(K_2^{n+1})(D_+^2K_1^n, \widehat{\Theta}^{2,n+1})d\tau^2\\
&\quad + \int_0^1\int_0^1 d^2F(K_2^{n+1})(D_+K_1^n, D_+\widehat{\Theta}^{2,n})d\tau^2\\
&\quad + \int_0^1\int_0^1\int_0^1 d^3F\left(K_3^n\right)\left(D_+K_2^n, D_+K_1^n, \widehat{\Theta}^{2,n}\right)d\tau^3,
\end{aligned}$$

where $d\tau^i = d\tau_1 \dots d\tau_i$, and

$$K_{i+1}^n = K_i^n + \tau_{i+1}(K_i^{n+1} - K_i^n) = K_1^n + \sum_{l=1}^{i} \sum_{2 \le i_1 < \dots < i_l \le i+1} \tau_{i_1} \cdots \tau_{i_l} k^l D_+^l K_1^n. \tag{4.9}$$

It follows the general formula

$$D_+^q h(t_n) = \sum_{i=1}^{q+1} \sum_{|\alpha_i|=q} L_{i,\alpha_i}^{n,q}, \text{ for } q = 1, 2, \dots, 2p+1, \text{ and } n \le N - q, \tag{4.10}$$

where $\alpha_i = (\alpha_i^1, \dots, \alpha_i^{i-1}, \alpha_i^i) \in \{1, 2, \dots, q\}^{i-1} \times \{0, 1, \dots, q - i + 1\}$, and $L_{i,\alpha_i}^{n,q}$ is a linear combination, with properly chosen coefficients, of the quantities

$$L_{i,\alpha_i,\beta_i}^{n,q} = \int_{[0,1]^i} d^i F(K_i^{n+q+1-i}) \left(D_+^{\alpha_i^{i-1}} K_{i-1}^{n+\beta_i^{i-1}}, \dots, D_+^{\alpha_i^1} K_1^{n+\beta_i^1}, D_+^{\alpha_i^i} \widehat{\Theta}^{2,n+\beta_i^i} \right) d\tau^i,$$

where $\beta_i = (\beta_i^1, \dots, \beta_i^{i-1}, \beta_i^i) \in \{1, 2, \dots, q\}^{i-1} \times \{0, 1, \dots, q - i + 1\}$ with $\beta_i^l + \alpha_i^l \le q - l + 1$, for $l = 1, \dots, i$. From (4.9) and (4.3) we have

$$K_i^{n+1} = u(t_{n+1/2}) + O(k), \text{ for } i = 1, 2, \dots, 2p+2,$$

and we deduce that there exists $k_3 > 0$ such that $0 < k \le k_3$ implies

$$\left\| d^i F\left(K_i^n\right) \right\| \le C_i, \quad \text{for } i = 1, 2, \dots, 2p+2, \text{ and } n = 0, 1, \dots, N - i + 1, \tag{4.11}$$

where C_i is a constant depending only on k_3, T, and the derivatives of u and F up to order 3 and i, respectively. From the inductions hypothesis (4.2) and inequality (2.3) we have

$$\| D_+^r K_i^n \| \le C, \text{ for } 1 \le r \le i \le 2m+3, 1 \le n \le N - i - r + 1, \tag{4.12}$$

and

$$\| D_+^r \widehat{\Theta}^{2,n} \| \le Ck^2, \text{ for } 1 \le r \le 2m+1, 1 \le n \le N - r, \tag{4.13}$$

where C is a constant depending only on m, T, and the derivatives of u and F up to order $r+2$ and r, respectively. Each $L_{i,\alpha_i,\beta_i}^{n,q}$ being multilinear continuous, we deduce from (4.11)–(4.13) and the relation $\beta_i^l + \alpha_i^l \le q - l + 1$, for $l = 1, \dots, i$, that

$$\| L_{i,\alpha_i,\beta_i}^{n,q} \| \le Ck^2, \text{ for } 1 \le i \le q+1 \le 2m+2, \ n \le N - q.$$

It follows by the triangle inequality that (4.10) for $q = 2m+1$ yields

$$\| (D_+ D_-)^m D_+ h(t_{n+1}) \| = \left\| D_+^{2m+1} h(t_{n+1-m}) \right\| \le Ck^2,$$

for $n = m, m+1, \ldots, N-(m+1)-1$, where C is a constant depending only on p, T, and the derivatives of u and F up to order $2m+4$ and $2m+2$, respectively . Passing to the norm in identity (4.8), we deduce from (4.5) and the last inequality that

$$\| (D_+D_-)^{m+1}\, \Theta^{2,n+1} \| \leq Ck^2. \tag{4.14}$$

Otherwise, applying D_- to (4.8), inequalities (4.11)–(4.13) and (4.14) yield

$$\| (D_+D_-)^{m+1}\, h(t_{n+1}) \| = \left\| D_+^{2m+2} h(t_{n-m}) \right\| \leq Ck^2,$$

for $n = m, m+1, \ldots, N-(m+1)-1$, where C is a constant depending only on p, T, and the derivatives of u and F up to order $2m+5$ and $2m+3$, respectively. Therefore, passing to the norm in the identity obtained by applying D_- to (4.8), we deduce from (4.8) and the last inequality that

$$\| D_- (D_+D_-)^{m+1}\, \Theta^{2,n+1} \| \leq Ck^2, \tag{4.15}$$

for $n = m, m+1, \ldots, N-(m+1)-1$, with the constant C depending only on p, T, and the derivatives of u and F up to order $2m+5$ and $2m+3$, respectively. Inequalities (4.14) and (4.15) imply that the induction hypothesis is also true for $m+1$, and we deduce that (4.2) is true for each integer $m = 0, 1, \ldots, p$.

□

Proof (Proof of Theorem 4.1) We proceed by induction on $j = 1, 2, \ldots, p+1$. The case $j = 1$ is immediate from Lemma 4.1. Suppose that $\left\{u^{2j,n}\right\}_n^N$ approximates u with order $2j$ of accuracy and satisfies (4.1), for an arbitrary j such that $j \leq p$. We are going to prove that $\left\{u^{2j+2,n}\right\}_n^N$ approximates u with order $2j+2$ of accuracy and (4.1) holds substituting j by $j+1$.

From the induction hypothesis, $\left\{u^{2j,n}\right\}_n$ satisfies DCC. Because $\left\{u^{2j,n}\right\}_n$ and $\left\{\overline{u}^{2j,m}\right\}_m$ are computed from the same scheme DC2j, but for different time steps, $\left\{\overline{u}^{2j,m}\right\}_m$ also satisfies DCC. Therefore, as in 3.14, Theorem 3.1 applied to the approximation $\left\{u^{2j+2,n}\right\}_{n=0}^j$, built from $\left\{\overline{u}^{2j,m}\right\}_m$, yields

$$\|\overline{\Theta}^{2j+2,n}\| \leq C\frac{1}{2-\mu k}\left(\frac{2+\mu k}{2-\mu k}\right)^{n-1} k^{2j+2} + \left(\frac{2+\mu k}{2-\mu k}\right)^{n} \|\overline{\Theta}^{2j+2,0}\|,$$

where

$$\overline{\Theta}^{2j+2,n} = \left(u^{2j+2,n} - u(t_n)\right) - \Gamma^j \left(\overline{u}^{2j,(2j+1)n+j} - u(t_{(2j+1)n+j})\right), \text{ for } 1 \leq n \leq j.$$

According to the DCC and the condition $u^{2j+2,0} = u(t_0) = u_0$, we have

$$\left\|\overline{\Theta}^{2j+2,0}\right\| = \left\|\Gamma^j \left(\overline{u}^{2j,j} - u(t_j)\right)\right\| \leq Ck^{2j+2}.$$

By the triangle inequality and the DCC, the last two inequalities yield

$$\|u^{2j+2,n} - u(t_n)\| \leq Ck^{2j+2}, \text{ for } n = 0, 1, \ldots, j. \tag{4.16}$$

From the DCC on $\left\{u^{2j,n}\right\}_n$ and the inequality (4.16), Theorem 3.1 again implies that $\left\{u^{2j+2,n}\right\}_{n=0}^N$ approximates the exact solution u with order $2j+2$ of accuracy. Therefore, it is enough to establish (4.1) for $j+1$, $j \leq p$. To this end we rewrite identity (3.9) as follows

$$D\Theta^{2j+2,n+1/2} = H(t_{n+1}) + \sigma^{2j+2,n+1/2} + (\Lambda^j - \Gamma^j)D(u^{2j,n+1/2} - u(t_{n+1/2})), \tag{4.17}$$

with

$$H(t_{n+1}) = \int_0^1 d_u F\left(t_{n+1/2}, \widehat{u}(t_{n+1}) - \Gamma^j \widehat{u}(t_{n+1}) + \tau_1 \widehat{\Theta}^{2j+2,n+1}\right)\left(\widehat{\Theta}^{2j+2,n+1}\right) d\tau_1,$$

where $\Theta^{2j+2,n}$ and $\sigma^{2j+2,n+1/2}$ are as in Theorem 3.1. Proceeding as in Lemma 4.1 and taking the finite difference formulae (2.4) and (2.5) into account, we can write

$$\sigma^{2j+2,n+1/2} = k^{2j+2}\varepsilon_1(t_{n+1}),$$

where

$$\|D_+^{m_1} D_-^{m_2} \varepsilon_1(t_{n+1})\| \leq C, \text{ for } m_1 + m_2 \leq 2p - 2j \text{ and } m_2 - 1 \leq n \leq N - m_1 - 1,$$

C is a constant depending only on p, T, and the derivatives of u and F. According to the inequality (4.1) from the induction hypothesis, we may write

$$(\Lambda^j - \Gamma^j)D(u^{2j,n+1/2} - u(t_{n+1/2})) = k^{2j+2}\varepsilon_2(t_{n+1}),$$

where

$$\|D_+^{m_1} D_-^{m_2} \varepsilon_2(t_{n+1})\| \leq C, \text{ for } m_1 + m_2 \leq 2p - 2j + 2 \text{ and } m_2 - 1 \leq n \leq N - m_1 - 1.$$

Therefore, writing (4.17) as follows

$$D_-\Theta^{2j+2,n+1} = H(t_{n+1}) + k^{2j+2}G(t_{n+1}),$$

with

$$G(t_{n+1}) = \varepsilon_1(t_{n+1}) + \varepsilon_2(t_{n+1}),$$

the induction hypothesis and the reasoning from Lemma 4.1, substituting the functions h and g, respectively, by H and G, $\widehat{\Theta}^{2,n+1}$ by $\widehat{\Theta}^{2j+2,n+1}$, and k^2 by k^{2j+2}, yields

$$\|(D_+D_-)^m D\widehat{\Theta}^{2j+2,n+1/2}\| + \|(D_+D_-)^m \widehat{\Theta}^{2j+2,n+1}\| \leq Ck^{2j+2},$$

for $m = 0, 1, \ldots, p - j$ and $n = m + j - 1, m + j, \ldots, N - j - m$, where C is a constant depending only on p, T, and the derivatives of u and F up to order $2(m + j + 1) + 1$ and $2(m + j) + 1$, respectively. Inequality (4.1) holds for $\left\{u^{2j+2,n}\right\}_n$ by the triangle inequality from the last inequality. □

We end this section by the following corollary that gives an important convergence property of the DC method. This property is useful for a time-stepping method to solve stiff and large dimensional differential equations arising from the space discretization of time-dependent PDEs.

Corollary 4.1 *Suppose that the function F is from $\mathbb{R}^s \to \mathbb{R}^s$, for a positive integer s, and satisfies the one-sided Lipschitz condition (3.7). Then, each approximate solution $\left\{u^{2j,n}\right\}_{n=0}^{N}$ from $DC2j$ satisfies the inequality*

$$|u^{2j,n} - u(t_n)| \leq Ck^{2j}, \text{ for each } k \in (0, k_0), \tag{4.18}$$

where C is a constant independent from any global Lipschitz constant on F, and either $k_0 = 2/\beta$ for $\beta > 0$ or $k_0 = +\infty$ for $\beta \leq 0$.

Proof From the regularity assumption on F and u and the one sided-Lipschitz condition, we deduce from Theorem 4.1 that each $\left\{u^{2j,n}\right\}_{n=0}^{N}$, $j = 1, 2, \ldots$, satisfies DCC. Therefore, inequality (4.18) is immediate from the part 4 of Theorem 3.1. The constant C depends only on the derivatives of u up to order $2j + 1$ and, according to (3.16)–(3.17) and the mean value theorem, on the bound of the Jacobian F_y on the compact set $[0, T] \times \left\{y \in \mathbb{R}^s : |y| \leq R_j\right\}$. □

Remark 4.1 The convergence property satisfied by the schemes $DC2j$ in Corollary 4.1 is in fact B-convergence (see, e.g., [9,19]) since the constant C of the global error in (4.18) is independent from any global Lipschitz constant of the function F. Nevertheless, since in the definition of B-convergence the constant C depends on high order derivatives of the exact solution u, the identity

$$u''(t) = F_t(t, u(t)) + F_u(t, u(t)) \cdot u'(t)$$

can make any requirement on the independence of the constant C with respect to F_u somewhat artificial. The numerical test on Bernoulli ODE in Sect. 6 gives an application of Corollary 4.1.

Remark 4.2 From part 4 of the proof of Theorem 3.1, the global error for an approximate solution by a DC2j+2 method, $j = 0, 1, 2, \ldots$, of the IVP (1.1) under the one-sided Lipschitz condition (3.7) takes the form

$$\|u^{2j+2,n} - u(t_n)\| \leq C \left(\frac{2 + \beta k}{2 - \beta k}\right)^n k^{2j+2}, \tag{4.19}$$

whenever $-2 \leq \beta k < 2$. The constant C depends on the derivatives of the function F up to order $2j + 2$ and can be very large in magnitude. However, if $\beta < 0$ and k is not

necessarily small, the factor $\left(\frac{2+\beta k}{2-\beta k}\right)^n$ sharply decreases with n, so that $C\left(\frac{2+\beta k}{2-\beta k}\right)^n << 1$, leading to accurate approximate solutions. As the time step k gets larger, $\frac{2+\beta k}{2-\beta k}$ gets smaller and accuracy occurs from the first iterations. Nevertheless, when k is very small, $\left(\frac{2+\beta k}{2-\beta k}\right)^n$ is close to 1 for smaller values of n, so that the global error bound of our DC methods reduces to $C\,k^{2j+2}$ shortly after startup. This occurs when k is in the asymptotic region $k\mu < 2$, where μ is the global Lipschitz constant of F, μ large. For such small k, non B-convergent methods can also be used and may be competitive with our B-convergent DC methods, at least during this time interval following startup. This situation will be illustrated with the Bernoulli equation in Sect. 6.

5 Absolute stability

In this section we prove the absolute stability of the DC schemes. The notion of absolute stability is introduced by Dahlquist [5] to characterize methods able to solve stiff ODEs. Considering the following IVP,

$$\begin{cases} u' &= \lambda u \\ u(0) &= 1, \end{cases} \tag{5.1}$$

where λ is a complex number, we have the following definition (see [5,22]):

Definition 5.1 A numerical method is said to be absolutely stable if the corresponding solution for the problem (5.1) for fixed $k > 0$ and some $Re(\lambda) < 0$ is such that

$$\lim_{n\to+\infty} |u^n| = 0. \tag{5.2}$$

The region of absolute stability of a numerical method is defined as the subset of the complex plane

$$\mathcal{A} = \{z = \lambda k \in \mathbb{C} : (5.2) \text{ is satisfied } \}. \tag{5.3}$$

If $\mathcal{A} \cap \mathbb{C}_- = \mathbb{C}_-$, $\mathbb{C}_- = \{\lambda \in \mathbb{C} : Re(\lambda) < 0\}$, the numerical method is said to be A-stable.

Before establishing absolute stability results for the deferred correction schemes (2.7) and (2.8)–(2.9), we recall the following result.

Lemma 5.1 (See [27, formula (6)]) *Let P_m be a polynomial of degree m in one variable. Then the sum $\sum_{i=0}^{n} P_m(i)$ is a polynomial of degree $m+1$ in the variable n.*

Lemma 5.2 *Suppose that $F(t,u) = \lambda u$ and $u_0 = 1$ in the initial value problem (1.1), where λ is a complex number with negative real part ($\lambda \in \mathbb{C}_-$). Then the corresponding approximate solutions from the schemes (2.7) and (2.8)–(2.9) can be written as follows*

$$u^{2j+2,n} = \left(\frac{2+\lambda k}{2-\lambda k}\right)^{n-j} P_j(n), \text{ for } j = 0, 1, 2, \ldots, \text{ and } n \geq j, \tag{5.4}$$

where $P_j(n)$ is a polynomial of degree j in the variable n.

Proof We suppose that $\lambda k \neq -2$, otherwise we trivially have $u^{2j,n+1} = 0$, for $n \geq j$. Since $F(t,u) = \lambda u$, we can rewrite (2.8) as follows

$$u^{2j+2,n+1} = \frac{2+\lambda k}{2-\lambda k} u^{2j+2,n} + \frac{2}{2-\lambda k}\left(kD_-\Lambda^j u^{2j,n+1} - \lambda k \Gamma^j \widehat{u}^{2j,n+1}\right)$$

where, according to formulae (2.1) and (2.2), we have

$$\begin{aligned} kD_-\Lambda^j u^{2j,n} &= \sum_{i=1}^{j} c_{2i+1} k^{2i+1} D_-(D_+D_-)^i u^{2j,n} \\ &= \sum_{i=1}^{j}\sum_{m=0}^{2i+1} c_{2i+1}(-1)^m \binom{2i+1}{m} u^{2j,n+i-m}, \end{aligned}$$

and

$$\Gamma^j \widehat{u}^{2j,n} = \sum_{i=1}^{j} c_{2i} k^{2i} (D_+D_-)^i \widehat{u}^{2j,n} = \sum_{i=1}^{j}\sum_{m=0}^{2i} c_{2i}(-1)^m \binom{2i}{m} \widehat{u}^{2j,n+i-m}.$$

Combining the last three identities, we deduce that

$$u^{2j+2,n+1} = \frac{2+\lambda k}{2-\lambda k} u^{2j+2,n} + \frac{2}{2-\lambda k}\sum_{i=0}^{2j+1} \alpha_{j,i}(\lambda k) u^{2j,n+1+j-i}, \text{ for } n \geq j \geq 1, \tag{5.5}$$

where $\alpha_{j,i}$ is affine in λk. Under the hypothesis of the lemma, (2.7) matches the trapezoidal rule, and we have

$$u^{2,n} = \left(\frac{2+\lambda k}{2-\lambda k}\right)^n,$$

that is (5.4) is true for $j = 0$. Suppose that (5.4) holds for an arbitrary integer $j \geq 0$. From (5.5) we have

$$u^{2j+4,n} = \frac{2+\lambda k}{2-\lambda k} u^{2j+4,n-1} + \frac{2}{2-\lambda k}\sum_{i=0}^{2j+3} \alpha_{j+1,i}(\lambda k) u^{2j+2,n+1+j-i},$$

with $n \geq j+2$, and, substituting each $u^{2j+2,n+1+j-i}$ by the formula given by the induction hypothesis (5.4), we deduce that

$$u^{2j+4,n} = \frac{2+\lambda k}{2-\lambda k} u^{2j+4,n-1} + \left(\frac{2+\lambda k}{2-\lambda k}\right)^{n-j-1} Q_j(n),$$

where

$$Q_j(n) = \frac{2}{2-\lambda k} \sum_{i=0}^{2j+2} \alpha_{j+1,i}(\lambda k) \left(\frac{2+\lambda k}{2-\lambda k}\right)^{j+2-i} P_j(n+1+j-i).$$

It follows that

$$u^{2j+4,n} = \left(\frac{2+\lambda k}{2-\lambda k}\right)^{n-j-1} \left(u^{2j+4,j+1} + \sum_{i=j+2}^{n} Q_j(i)\right).$$

It is clear that $Q_j(n)$ is a polynomial of degree j in the variable n as $P_j(n)$. Therefore, according to the Lemma 5.1, $\sum_{i=j+2}^{n} Q_j(i)$ is a polynomial of degree $(j+1)$ in the variable n. Whence,

$$u^{2j+4,n} = \left(\frac{2+\lambda k}{2-\lambda k}\right)^{n-j-1} P_{j+1}(n),\ n \geq j+1,$$

where

$$P_{j+1}(n) = u^{2j+4,j+1} + \sum_{i=j+2}^{n} Q_j(i)$$

is a polynomial of degree $j+1$ in the variable n. We then deduce by induction that the lemma is true for arbitrary non-negative integer j. □

Theorem 5.1 *Each of the deferred correction schemes* (2.7) *and* (2.8)–(2.9) *is A-stable.*

Proof From Lemma 5.2 we have, for $Re(\lambda k) < 0$,

$$\lim_{n\to+\infty} |u^{2j+2,n}| = \lim_{n\to+\infty} \left|\left(\frac{2+\lambda k}{2-\lambda k}\right)^{n-j} P_j(n)\right| = \lim_{n\to+\infty} |P_j(n)|e^{(n-j)\ln\left|\frac{2+\lambda k}{2-\lambda k}\right|} = 0$$

since, under the condition $Re(\lambda k) < 0$, we have $\left|\frac{2+\lambda k}{2-\lambda k}\right| < 1$. □

6 Numerical experiments

In this section we evaluate the accuracy and order of convergence of the schemes $DC2, DC4, \ldots, DC10$, implemented using the Scilab programming language. The starting values are computed using the scheme (2.10)–(2.11).

We choose six standard problems for the evaluation. The first problem concerns B-convergence by considering a Bernoulli equation. The second problem is about long term integration with an oscillatory solution of large amplitude. The four other problems are about stiffness. The third and fourth problems (B5 modified and E5,

respectively) both involve complex eigenvalues of negative real parts, where the imaginary parts of the eigenvalues for the third problem have larger magnitudes while those from the fourth problem have smaller magnitudes. The fifth problem (Robertson) is nonlinear and stiff with real negative eigenvalues, and it also addresses B-convergence. The sixth problem is the van der Pol oscillator, which is stiff with arbitrary complex eigenvalues.

The first three problems have analytic solutions. For problems (6.4), (6.5) and (6.6) that do not have an analytic solution, we consider a small time step such that the approximate solutions with $DC6, \ldots, DC10$ are almost identical [to machine precision for problem (6.5)], and we choose one of the approximate solutions as reference solution.

For solutions $u = (u_1, \ldots, u_d) : [0, T] \to \mathbb{R}^d$, $1 \le d \le 6$, the absolute error on the approximate solutions $\left\{u^{2j,n}\right\}_{0 \le n \le N}$, $1 \le j \le 5$, is computed with the norm

$$\|u_i^{2j} - u_i\| = \max_{0 \le n \le N} |u_i^{2j,n} - u_i(t_n)|, \quad 1 \le i \le d.$$

For very large N we extract solutions at 2×10^6 or 3×10^6 discrete times evenly spread over the interval $[0, T]$.

For a comparison of accuracy, we implement in Scilab the backward differentiation formulae (BDF) of order 2, 4 and 6, and the explicit Runge-Kutta (RK) of order 4. The implemented BDF are run with exact starting values for the first three problems that have analytic solutions, while for problems four and five the starting values are provided by the function `stiff` (implementing BDF with adaptive steps) of the solver `ode` from Scilab. For the van der Pol oscillator, the comparison of our DC methods is done only with the solutions from `stiff` and `rkf` from the solver `ode`. For each of the problems, we give a table of absolute errors and orders of convergence for pairs of two consecutive time steps, for the approximate solutions with the DC methods. We denote by k_{max} the maximal time step allowed to compute an approximate solution with the solver `stiff` or `rkf` (see [8] for a discussion on maximal time steps).

6.1 Bernoulli differential equation

$$u'(t) = F(t, u) = -0.1u(t) - 1000u^{20}(t), \quad u(0) = 1, \quad t \in [0, 10]. \tag{6.1}$$

Table 2 gives the absolute error and the order of convergence for each pair of consecutive time steps, in the case of DC, BDF and RK4 methods. The dash for RK4 indicates that the method is unstable for the corresponding time steps.

This problem addresses B-convergence since the function F is one-sided Lipschitz with $\beta = -0.1$, when positive solutions are considered. The problem is strongly nonlinear with the magnitude of derivatives of the right hand side function F exponentially increasing with the order of the derivatives. Such derivatives of large magnitude generally limit the accuracy of high order methods that are not B-convergent. In fact, B-convergence provides at least two main advantages to time-stepping methods. First, when $\beta < 0$, the error estimate for our B-convergent methods remains valid for large

Table 2 Absolute error (order of convergence) for the Bernoulli problem

k	DC2	DC4	DC6	DC8	DC10
1	0.18	1.7e−2	1.8e−4	2.3e−4	1.3e−4
2.03e−3	3.71e−2	6.16e−4	7.14e−5	1.47e−6	9.42e−7
1.00e−3	2.48e−2 (0.62)	2.69e−4 (1.16)	4.86e−5 (0.54)	1.71e−6 (-0.18)	6.34e−7 (0.48)
1.00e−4	1.92e−3 (1.09)	2.93e−5 (0.96)	4.31e−6 (1.05)	3.72e−7 (0.65)	5.78e−8 (1.05)
1.00e−5	2.22e−5 (1.94)	1.30e−7 (2.35)	3.92e−9 (3.04)	1.9e−10 (3.27)	1.1e−11 (3.73)
5.00e−6	5.55e−6 (2.0)	1.04e−8 (3.70)	1.4e−10 (3.70)	4.4e−12 (5.50)	4.4e−13 (4.64)
3.33e−6	2.46e−6 (1.99)	2.59e−9 (3.33)	1.6e−11 (5.31)	4.5e−13 (5.63)	2.0e−13 (2.02)
2.25e−6	1.39e−6 (1.99)	8.7e−10 (3.79)	3.3e−12 (5.54)	4.2e−13 (0.16)	4.2e−13 (-2.66)

k	BDF2	BDF4	BDF6	RK4
1	0.14	0.83	6.1e−2	–
2.03e−3	4.29e−2	2.54e−2	1.91e−3	–
1.00e−3	3.22e−2	1.88e−2	1.39e−2	0.354
1.00e−4	6.61e−3 (0.69)	2.98e−3 (0.80)	1.79e−3 (0.89)	1.27e−3 (2.45)
1.00e−5	2.59e−4 (1.41)	1.92e−5 (2.19)	3.15e−6 (2.76)	4.91e−8 (4.41)
5.00e−6	7.29e−5 (1.91)	1.92e−6 (3.58)	1.35e−7 (5.11)	2.53e−9 (4.28)

$k > 0$, as stated in our Corollary 4.1, and the error can be small even for relatively large time step k, as discussed in Remark 4.2. As seen in Table 2, DC methods provide accurate approximate solutions for large time steps, and their accuracy increases with the order of the method. However, the order of convergence of the DC methods is suboptimal in this range of time steps. BDF methods are stable for large time steps, but they are less accurate than their corresponding DC methods (except for BDF2). As expected, RK4 is unstable for $k \geq k_0 \approx 2.03 \times 10^{-3}$.

A second main advantage of B-convergence is the relatively small size of the error constant that depends on β (among others) instead of the potentially much larger two-side Lipschitz constant μ, see Remarks 4.1 and 4.2. This should result in lower error when comparing a B-convergent method with a non B-convergent method of the same order for the same time step, in the range of small time steps. Table 2 suggests that the error constants are, respectively, about 10, 100, 1000 smaller for DC2, DC4, DC6 compared to BDF2, BDF4, BDF6 methods. Of course, care is needed when doing such crude comparisons of errors. For instance, some non B-convergent method may happen to be competitive on a specific problem for very small time steps, as discussed in Remark 4.2. For the smallest time steps, RK4 is more accurate than DC4 and any of the BDF methods, but as expected DC6-10 achieve better accuracy.

Finally, we note that $DC4$ and $DC6$ almost achieve their proper order for $k \leq 5 \times 10^{-6}$, and the order of convergence of $DC8$ and $DC10$ are not observed since these methods quickly achieve machine accuracy.

6.2 Oscillatory problem [14]

$$u' = \lambda u \cos(t), \quad u(0) = 1, \quad T = 10^6, \lambda = 10. \tag{6.2}$$

The exact solution is $u(t) = e^{\lambda \sin(t)}$. The original problem is set with $\lambda = 1$ in [14]. The author in [16] solved this problem with Runge-Kutta methods of orders 4 and 8, for $\lambda = 2$ and $T = 2580\pi$, to "illustrate the need of higher order methods when a long-term integration problem is considered". Table 3 gives the absolute error and the order of convergence for each pair of consecutive time steps. The BDF methods are run only for the smallest time step. The solvers `rkf` and `stiff` use adaptive time stepping with a maximal time step $k_{max} = 0.1$ and tolerances $rtol = 100 \times atol = 10^{-10}$.

The magnitude of the exact solution $u(t) = e^{10 \sin(t)}$ of the modified oscillatory problem is large, resulting in a relatively large absolute error obtained by the DC schemes (absolute errors of about 10^{-7} is possible for a good choice of stepsize). Moreover, the long term integration influences the accuracy of these schemes since they achieve absolute errors of about 10^{-9} when the solution interval is reduced to [0, 1000]. Nevertheless, each DC scheme converges with its proper order. The DC methods are considerably more accurate than standard methods (both with fixed and variable stepsizes) which are inaccurate for this problem. For instance, for BDF2 and `rkf`, the solutions remain bounded with bounds close to the maximal amplitude of the exact solution but the phase of the oscillation is completely wrong.

Table 3 Absolute error (order of convergence) for the oscillatory problem

k	DC2	DC4	DC6	DC8	DC10
5.00e−2	3418	456.26	42.665	3.2350	0.2132
2.50e−2	790.2 (2.1)	25.351 (4.2)	0.5959 (6.2)	1.17e−2 (8.1)	1.9e−4 (10.1)
1.25e−2	193.8 (2.0)	1.5493 (4.0)	9.17e−3 (6.0)	5.28e−5 (7.8)	2.79e−6 (6.1)
6.25e−3	48.23 (2.0)	9.67e−2 (4.0)	1.4e−4 (5.99)	2.78e−6 (0.0)	2.78e−6 (0.0)
1.56e−3	3.010 (2.0)	3.8e−4 (3.99)	4.72e−6 (2.5)	4.67e−6 (-0.3)	4.7e−6 (-0.3)

k	BDF2	BDF4	BDF6	`rkf`	`stiff`
1.56e−3	22,026.46	14,836.76	5578.40	22,026.46	2636.00

6.3 Problem B5 modified [8], stiff with complex eigenvalues of negative real parts and larger (in magnitude) imaginary parts

$$y' = \begin{bmatrix} -10 & \alpha & 0 & 0 & 0 & 0 \\ -\alpha & -10 & 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.1 \end{bmatrix} y,\ y(0) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},\ \alpha = 5000,\ T = 20. \quad (6.3)$$

This problem, originally set with $\alpha = 100$, is an illustration of ODEs resulting from a semi-discretization by finite element methods of parabolic PDEs [26]. We choose $\alpha = 5000$ to make the problem a little more difficult. Table 4 gives the absolute errors for the first component of the approximate solutions which is similar for the second component. The absolute errors for the others components quickly achieve machine precision. The solvers `stiff` and `rkf` are run with $k_{max} = 2 \times 10^{-5}$ and $atol = 10 \times rtol = 10^{-15}$.

Table 4 Absolute error (order of convergence) for the first component of the solution for $B5$ modified

k	DC2	DC4	DC6	DC8	DC10
2.000e−5	0.2152	6.51e−2	2.22e−2	8.00e−3	2.98e−3
5.000e−6	1.35e−2 (2)	2.59e−4 (4)	5.59e−6 (6)	1.27e−7 (8)	2.97e−9 (10)
2.500e−6	3.38e−3 (2)	1.62e−5 (4)	8.74e−8 (6)	4.9e−10 (8)	2.9e−12 (10)
1.250e−6	8.47e−4 (2)	1.01e−6 (4)	1.36e−9 (6)	1.9e−12 (8)	7.4e−14 (5.3)
3.125e−7	5.29e−5 (2)	4.00e−9 (4)	3.6e−13 (6)	7e−14 (2.4)	6.3e−14
6.250e−8	2.11e−6 (2)	6.3e−12 (4)	6.02e−13	2.33e−13	1.19e−13

k	BDF2	BDF4	BDF6	`rkf`	`stiff`
1.25e−6	3.38e−3	7.94e−8	2.3e−12	2.36e−6	6.6e−10

The imaginary parts of the Jacobian eigenvalues of the modified B5 problem are large. Even though the real parts of the eigenvalues are negative, we observe that smaller time steps are required by DC schemes to obtain accurate approximations. DC schemes achieve their proper order of convergence, but BDF methods perform better for this problem than DC schemes.

6.4 Problem E5 [8], stiff with complex eigenvalues of negative real parts and smaller (in magnitude) imaginary parts

$$\begin{aligned} y_1' &= -7.89 \times 10^{-10} y_1 - 1.1 \times 10^7 y_1 y_2 \\ y_2' &= 7.89 \times 10^{-10} y_1 - 1.13 \times 10^9 y_2 y_3 \\ y_3' &= 7.89 \times 10^{-10} y_1 - 1.1 \times 10^7 y_1 y_2 + 1.13 \times 10^3 y_4 - 1.13 \times 10^9 y_2 y_3 \\ y_4' &= 1.1 \times 10^7 y_1 y_2 + 1.13 \times 10^3 y_4 \\ y(0) &= (1.76 \times 10^{-3}, 0; 0; 0)^t, \; T = 1000. \end{aligned} \tag{6.4}$$

A reference solution is computed with $DC10$ for $k = 10^{-3}$. The solution of this problem has small magnitude in $[1.618 \times 10^{-3}, 1.76 \times 10^{-3}] \times [0, 1.46 \times 10^{-10}] \times [0, 8.27 \times 10^{-12}] \times [0, 1.38 \times 10^{-10}]$ and the eigenvalues of the Jacobian matrix $dF(y)$ along the solution curve belong to the region $[-20490, 3.68 \times 10^{-12}] \times [-9.17 \times 10^{-5}, 9.17 \times 10^{-5}]$ of the complex plane. Table 5 gives the absolute errors and order of convergence for the four components of the approximate solutions. For BDF, RK4 and `stiff`, the absolute errors are provided only for the first component. The absolute error on the other components is smaller by 2 (RK4) to 5 (`stiff`) orders of magnitude, as we should expect from the magnitude of the solution components. The implemented BDF methods are run with starting values deduced from the solver `stiff`. The implemented RK4 is unstable for time steps $k \geq 2 \times 10^{-4}$, and the absolute error is reported for $k = 10^{-4}$ in Table 5. The solver `stiff` is run with $k_{max} = 10^{-3}$ and $rtol = 10^8 \times atol = 10^{-15}$.

Imaginary parts of eigenvalues for the problem E5 are smaller, and larger time steps allow DC schemes to produce very accurate approximations, compared to the modified B5 problem. DC schemes perform better for this problem than BDF methods. They achieve their proper order of convergence but on a relatively small range of time steps, for higher order DC methods, since the solution is already very accurate for large time steps.

6.5 Robertson (1966) [11], stiff with real negative eigenvalues

$$\begin{aligned} y_1' &= -0.04 y_1 + 10^4 y_2 y_3 \\ y_2' &= 0.04 y_1 - 10^4 y_2 y_3 - 3.10^7 y_2^2 \\ y_3' &= 3.10^7 y_2^2 \\ y(0) &= (1, 0, 0)^t, \quad T = 10^5. \end{aligned} \tag{6.5}$$

Table 5 Absolute error (order of convergence) for the problem E5

k	DC2	DC4	DC6	DC8	DC10
100	2.79e−07	5.34e−08	8.31e−09	4.26e−09	1.04e−09
	8.30e−12	9.68e−13	6.86e−14	6.14e−14	1.66e−14
	4.47e−13	5.31e−14	3.28e−15	3.40e−15	8.42e−16
	7.85e−12	9.14e−13	6.54e−14	5.81e−14	1.57e−14
50	7.52e−08 (1.89)	1.02e−08 (2.38)	1.56e−09 (2.41)	8.53e−11 (5.64)	4.92e−11 (4.41)
	1.96e−12 (2.08)	6.46e−14 (3.90)	3.16e−14 (1.12)	2.94e−15 (4.38)	5.07e−16 (5.03)
	1.07e−13 (2.06)	3.73e−15 (3.83)	1.61e−15 (1.02)	2.21e−16 (3.94)	9.78e−17 (3.11)
	1.86e−12 (2.08)	6.14e−14 (3.89)	3.00e−14 (1.12)	2.85D-15 (4.35)	4.09D-16 (5.26)
10	3.16e−09 (1.99)	2.37e−11 (4.03)	5.26e−13 (5.23)	1.28e−14 (6.72)	4.51e−16 (8.89)
	7.77e−14 (1.99)	2.79e−16 (3.68)	3.02e−18 (5.74)	1.15e−19 (7.94)	7.28e−21 (8.09)
	4.31e−15 (1.97)	7.08e−17 (1.79)	5.91e−17 (0.24)	6.27e−17 (0.12)	6.84e−17 (0.09)
	7.34e−14 (1.99)	3.20e−16 (3.37)	6.18e−17 (1.79)	6.28e−17 (0.57)	6.84e−17 (0.11)

k	BDF2	BDF4	BDF6	RK4 ($k = 1.0e-4$)	`stiff`
10	5.7e−8	6.6e−10	3.5e−11	2.03e−16	1.29e−16

This is one of the three problems considered as stiffest in [11]. We compute a reference solution with DC10 for the time step $k = 1/6000$. The solution belongs to the region $[1.78 \times 10^{-2}, 1.00] \times [0, 3.58 \times 10^{-5}] \times [0, 0.983]$ and the eigenvalues of the Jacobian dF(y) along the solution curve belong to $[-9825.744, 0]$. Table 6 gives absolute errors and orders of convergence of DC methods for each component of the solution. For other methods, we give only the maximal errors on the three components of the approximate solutions. The solver `stiff` is run with $k_{max} = 1/600$ and $rtol = 100 \times atol = 10^{-15}$. The solver `rkf` fails in solving this problem for various tolerances and k_{max}, and Scilab reported: "it is likely that rkf45 is inefficient for solving this problem". The implemented BDF methods are run with starting values deduced from the solver `stiff` using the preceding tolerances.

The Robertson problem is stiff and addresses B-convergence since its Jacobian matrix has real negative eigenvalues with some having large magnitude. For this problem, DC schemes produce accurate approximate solutions even for large time steps, and high order DC methods can be avoided (DC6 is enough). The convergence is slow for $k > 1/300$, but faster convergence happens for k in the asymptotic region ($k < 1/300$). The DC schemes perform better than BDF methods at equal order and time step. A comparison of the errors for $k = 1/600$ suggests that the error constants might be 3 to 5 orders of magnitude smaller for DC than BDF methods.

6.6 van der Pol oscillator [8,24], stiff with arbitrary complex eigenvalues

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= \mu(1 - y_1^2)y_2 - y_1 \\ y_1(0) = 2, \quad y_2(0) &= 0, \quad T = 3000, \quad \mu = 1000. \end{aligned} \tag{6.6}$$

This problem was initially proposed for $T = 1$ and $\mu = 5$ in [8]. The actual version results from a suggestion by Shampine [24]. We compute a reference solution with DC8 for $k = 1.875 \times \times 10^{-6}$. The solution belong to the region $[-2, 2.000073] \times [-1323.04, 1231.35]$ of the real plane and the eigenvalues along the solution curve belong to the region $[-3000.29, 1123.17] \times [-1158.48, 1158.48]$ of the complex plane. Table 7 gives the absolute errors and orders of convergence. For `rkf` and `stiff`, we use $k_{max} = 7.5 \times 10^{-5}$ and $rtol = 10$, $atol = 10^{-16}$.

The van der Pol oscillator is stiff and the solution has a large magnitude. DC6 and DC8 reached their order of convergence. This shows that the DC strategy works well in spite of the fact that DC2 and DC4 would require much smaller time steps to produce reasonably accurate solutions. The order of convergence for DC10 is not observed, though the solutions obtained are accurate.

6.7 Discussion of the numerical results

In general, a careful assessment of the proof of Theorem 3.1 points out to the fact that, for a system with complex eigenvalues $\lambda = \lambda_1 + i\lambda_2$, we only need a time step k

Table 6 Absolute error (order of convergence) for Robertson problem

k	DC2	DC4	DC6	DC8	DC10
0.5	3.63e−5	4.46e−6	2.08e−6	2.91e−6	3.09e−6
	3.63e−5	4.46e−6	2.08e−6	2.91e−6	3.09e−6
	7.12e−5	4.37e−7	1.02e−7	4.12e−7	4.26e−7
1/300	4.7e−9 (1.8)	1.09e−9 (1.7)	4.0e−10 (1.7)	3.0e−10 (1.9)	2.0e−10 (1.9)
	7.4e−9 (1.7)	2.23e−8 (1.1)	4.16e−8 (0.8)	2.9e−8 (0.9)	2.5e−8 (0.9)
	4.7e−9 (1.9)	2.12e−8 (0.6)	4.12e−8 (0.6)	2.8e−8 (0.5)	2.5e−8 (0.6)
1/600	1.0e−9 (2.2)	1.5e−10 (2.8)	1.0e−12 (8.6)	9.9e−13 (8.)	7.5e−13 (8.2)
	5e−13 (14.)	3.0e−14 (19.6)	2.0e−16 (27.7)	2.0e−16 (27.1)	3.0e−16 (26.1)
	1.0e−9 (2.2)	1.5e−10 (7.1)	1.0e−12 (15.3)	9.9e−13 (15)	4.0e−13 (15.8)
1/6000	9.24e−12	7.31e−14	1.48e−14	4.57e−14	–
	5.38e−15	0.	0.	0.	–
	9.25e−12	2.07e−13	1.36e−13	8.27e−14	–

k	BDF2	BDF4	BDF6	`RK4`	`stiff`
0.5	5.3e−4	3.6e−5	4.1e−6	–	–
1/600	2.8e−6	1.2e−6	6.9e−7	–	7.28e−13

Table 7 Absolute error (order of convergence) for the van der Pol's equation

k	DC2	DC4	DC6	DC8	DC10
3.75e−5	3.0089	2.9999	2.9440	0.1838	3.12e−3
	1322.9	1327.5	1320.6	197.79	3.26792
1.50e−5	2.9769 (0)	2.9999 (0)	0.1080 (3.6)	1.90e−4 (7.5)	5.1e−5 (4.5)
	1333.3 (0)	1330.3 (0)	113.69 (2.7)	0.18281 (7.6)	5.1e−2 (4.5)
7.50e−6	2.8706 (0)	2.6947 (0)	1.60e−3 (6.0)	1.74e−6 (6.7)	1.27e−5 (1.9)
	1327.4 (0)	1286.5 (0)	1.6349 (6.1)	1.80e−3 (6.7)	1.29e−2 (1.9)
1.875e−6	0.74(0.9)	0.339 (1.5)	2.50e−7 (6.3)	–	2.88e−7 (2.7)
	659. (0.5)	373.2 (0.9)	2.91e−4 (6.2)	–	2.92e−4 (2.7)
–	`stiff`		`rkf`		
	2.16e−6		3.54e−2		
	3.48e−3		64.76		

such that $k\ \max\{\lambda_1, |\lambda_2|\} < 2$ for a good accuracy (faster convergence happens when $-\lambda_1 >> |\lambda_2|$). These situations are well illustrated by the test cases of Sects. 6.3 and 6.4, where the required time step for accuracy is much smaller for modified B5 than E5. However, time steps k such that $k\mu \simeq k|\lambda| < 2$, $\mu \simeq \max_{0\le t\le T} \|d_u F\,(t, u(t))\|$, is necessary for an asymptotic convergence with proper order. For example, in the case of the Bernoulli equation we have $\lambda \simeq -20{,}000.1 < 0$ and $\mu = 20{,}000.1$. Large time steps provide accurate approximations (as expected from B-convergent methods), but asymptotic convergences are observed only for $k\mu < 2$.

For the computational effort of the DC methods, we recall that to compute an approximate solution on discrete points $0 = t_0 < t_1 < \cdots < t_N = T$, $DC2$ solves N nonlinear systems while $DC2j$, $j \ge 2$, solves $j \times N$ systems. In the case of the Bernoulli equation, for example, $DC10$ achieves the maximal error of about 1.1×10^{-11} by solving approximately 5×10^6 nonlinear systems while the maximal absolute error for $DC2$ is about 8.9×10^{-7} for $N = 5 \times 10^6$. We did not report any CPU time since our code is written in Scilab, an interpreted language. All methods that we implemented are consequently interpreted, while `rkf` and `stiff` provided with Scilab are compiled. Nevertheless, the main burden in implicit time-stepping solvers is the resolution of nonlinear systems, and we have shown that higher order DC methods give the most accurate approximations by solving fewer systems of equations. This gives a clue on the CPU time required and the efficiency of these methods. High order DC methods should be competitive in situations where using fully implicit methods is unavoidable.

7 Conclusions

We have presented a new approach of deferred correction methods for the numerical solution of general first order ordinary differential equations. Proofs for consistency,

order of convergence and stability of the method are given, which rely on a recursive argument using a new deferred correction condition. The numerical experiments comply with the theory and show a high accuracy of the method and its satisfactory A-stable property and B-convergence. Globally, each DC scheme reaches its proper order of convergence and applies to any category of problem, providing accurate approximations for time steps not necessarily small. The accuracy of the DC schemes increases with the level of correction.

References

1. Auzinger, W.: Defect correction methods. In: Engquist, B. (ed.) Encyclopedia of Applied and Computational Mathematics, pp. 323–332. Springer, Berlin (2015)
2. Christlieb, A., Ong, B., Qiu, J.M.: Integral deferred correction methods constructed with high order Runge–Kutta integrators. Math. Comput. **79**, 761–783 (2010)
3. Chung, T.: Computational Fluid Dynamics, 2nd edn. Cambridge University Press, Cambridge (2010)
4. Dahlquist, G., Björck, A.K.: Numerical Methods in Scientific Computing, vol. I. SIAM, Philadelphia (2008)
5. Dahlquist, G.G.: A special stability problem for linear multistep methods. Nordisk Tidskr. Informationsbehandling (BIT) **3**, 27–43 (1963)
6. Daniel, J.W., Pereyra, V., Schumaker, L.L.: Iterated deferred corrections for initial value problems. Acta Cient. Venezolana **19**, 128–135 (1968)
7. Dutt, A., Greengard, L., Rokhlin, V.: Spectral deferred correction methods for ordinary differential equations. BIT **40**, 241–266 (2000)
8. Enright, W.H., Hull, T., Lindberg, B.: Comparing numerical methods for stiff systems of O.D.E:s. BIT **15**, 10–48 (1975)
9. Frank, R., Schneid, J., Ueberhuber, C.W.: The concept of B-convergence. SIAM J. Numer. Anal. **18**(5), 753–780 (1981)
10. Gustafsson, B., Kress, W.: Deferred correction methods for initial value problems. BIT **41**, 986–995 (2001)
11. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations. II. Stiff and Differential-Algebraic Problems, vol. 14. Springer, Berlin (1991)
12. Hansen, A.C., Strain, J.: On the order of deferred correction. Appl. Numer. Math. **61**, 961–973 (2011)
13. Hildebrand, F.B.: Introduction to Numerical Analysis. McGraw-Hill, New York (1974)
14. Hull, T.E., Enright, W.H., Fellen, B.M., Sedgwick, A.E.: Comparing numerical methods for ordinary differential equations. SIAM J. Numer. Anal. **9**, 603–637 (1972)
15. Isaacson, E., Keller, H.B.: Analysis of Numerical Methods. Wiley, New York (1966)
16. Karouma, A.: A class of contractivity preserving Hermite–Birkhoff–Taylor high order time discretization methods. Ph.D. thesis, Université d'Ottawa/University of Ottawa (2015)
17. Koyaguerebo-Imé, S.C.E., Bourgault, Y.: Finite difference and numerical differentiation: General formulae from deferred corrections. arXiv preprint arXiv:2005.11754 (2020)
18. Koyaguerebo-Imé, S.C.R., Bourgault, Y.: Arbitrary high-order unconditionally stable methods for reaction-diffusion equations via deferred correction: Case of the implicit midpoint rule. arXiv:2006.02962v2 (2020)
19. Kraaijevanger, J.: B-convergence of the implicit midpoint rule and the trapezoidal rule. BIT **25**(4), 652–666 (1985)
20. Kress, W., Gustafsson, B.: Deferred correction methods for initial boundary value problems. J. Sci. Comput. **17**(1–4), 241–251 (2002)
21. Kushnir, D., Rokhlin, V.: A highly accurate solver for stiff ordinary differential equations. SIAM J. Sci. Comput. **34**, A1296–A1315 (2012)
22. Quarteroni, A., Sacco, R., Saleri, F.: Numerical Mathematics, vol. 37, 2nd edn. Springer, Berlin (2007)
23. Schild, K.H.: Gaussian collocation via defect correction. Numer. Math. **58**, 369–386 (1990)
24. Shampine, L.F.: Evaluation of a test set for stiff ODE solvers. ACM Trans. Math. Softw. **7**, 409–420 (1981)

25. Spijker, M.N.: Stiffness in numerical initial-value problems. J. Comput. Appl. Math. **72**, 393–406 (1996)
26. Stewart, K.: Avoiding stability-induced inefficiencies in BDF methods. J. Comput. Appl. Math. **29**, 357–367 (1990)
27. Tuenter, H.: The Frobenius problem, sums of powers of integers, and recurrences for the Bernoulli numbers. J. Number Theory **117**, 376–386 (2006)