Error Estimates for Deferred Correction Methods in Time

Wendy Kress*

Abstract

In this paper, we consider the deferred correction principle for high order accurate time discretization of partial differential equations (PDEs) and ordinary differential equations (ODEs). Deferred correction is based on a lower order method, here we use second order accurate A-stable methods. Solutions of higher order accuracy are computed successively. The computational complexity for calculating higher order solutions is comparable to the complexity of the lower order method. There is no stability restraint on the size of the time-step. Error estimates are derived and the application of the schemes to initial boundary value problems is discussed in detail. The theoretical results are supported by a series of numerical experiments.

1 Introduction

When choosing a method for the time discretization of a PDE or ODE, two aspects must be considered, stability and accuracy. Often, explicit time marching methods are used. However, when dealing with stiff problems and parabolic and higher order PDEs, one often encounters severe stability restrictions on the time-step for explicit methods. If, in addition, a method of high order accuracy is required, one faces the problem that standard high order explicit methods usually have very small stability regions. Thus, one needs to consider implicit methods which usually have better stability properties. Implicit methods commonly used include implicit Runge-Kutta methods which can be constructed to be A-stable, i.e., rendering no stability restriction on the time-step, for arbitrarily high order. Multistep methods like the backward differentiation formulae (BDF) are also frequently used. Implicit BDF methods are, however, not A-stable for orders higher than two and unstable for orders higher than six.

In order to obtain an A-stable p-th order accurate Runge-Kutta scheme, one needs a $\frac{p}{2}$-stage fully implicit Runge-Kutta method which involves the solution

*Department of Information Technology, Scientific Computing, University of Uppsala, Sweden. Email: wendy@tdb.uu.se

of a $\frac{p}{2}N \times \frac{p}{2}N$ system in each time-step, where N is the size of the system of ODEs. There are ways to reduce the computational complexity to the solution of p systems of size $N \times N$ in each time-step, when considering singular diagonally implicit Runge-Kutta methods (SDIRK).

In this paper, we consider a p-th order accurate method that requires the solution of only $\frac{p}{2}$ $N \times N$ systems per time-step and has very good stability properties. The method is not A-stable in the usual sense, but we consider a different linear stability concept. The stability and the performance of the method will depend on the smoothness of the initial condition and the forcing term and, for PDEs, on the treatment of the boundary conditions. However, there is no stability restriction on the size of the time-step.

The principle of the deferred correction method based on the implicit midpoint rule (IMR) has been presented in [12] and [18]. There, stability estimates have been derived for PDE with time-independent coefficients and special cases of time-dependent coefficients.

In this paper, we present the deferred correction method based on the second order implicit BDF2 scheme. We derive error estimates for the BDF2 and the IMR based schemes, which are valid for quite general time-dependent coefficient problems. We discuss in detail the smoothness requirements to obtain the desired order of accuracy in the deferred correction scheme both for the BDF2 scheme and the IMR. Special attention needs to be given to the choice of the additional initial condition needed for the BDF2 based scheme and for initial boundary value problems, a special formulation of the boundary conditions needs to be used. A number of numerical test problems is presented.

The paper is organized as follows. In Section 2, we present the general principle of the deferred correction scheme. In Section 3, we describe the method in detail for the implicit BDF2 scheme as an underlying scheme. We give error estimates and investigate the smoothness requirements for optimal convergence order for the BDF2 based deferred correction scheme for problems with time-independent coefficients. In Section 4, we derive similar error estimates for the IMR based deferred correction scheme. The treatment of problems with time-dependent coefficients is discussed in Section 5. In Section 6, a modified implementation of boundary conditions for PDEs is presented. It ensures the optimal order of accuracy of the deferred correction scheme. A series of numerical experiments is performed in Section 7, supporting the theoretical analysis in the previous sections. In Section 8, we investigate the performance of the deferred correction methods for problems which require the use of implicit methods. The performance of the deferred correction scheme is investigated both analytically and experimentally for two classes of stiff problems. In the last part of the paper, we briefly investigate a different approach to obtaining high order accurate methods for the time integration, using the so called *modified equation approach*.

2 The deferred correction principle

We use the deferred correction method to obtain a high order accurate time discretization method which can be applied to systems of PDEs and ODEs. When considering PDEs, we assume that the equations have already been discretized in space, yielding a system of ODEs, i.e., the *method of lines* is used. We also assume that the boundary conditions are incorporated in the ODE. We therefore consider a general system of ODEs,

$$\begin{aligned} u'(t) &= \mathcal{L}(u(t), t), \\ u(0) &= u_0. \end{aligned} \tag{1}$$

We now describe the general concept of the deferred correction method to obtain arbitrarily high order methods. We start by solving the system of ODEs (1) with a k-th order accurate scheme, preferably A-stable, to obtain a solution $u^{k,n} \approx u(t_n) = u(n\Delta t)$ for $n = 1, \ldots$. We now consider the local truncation error of the method. It is obtained by inserting the exact solution to (1) into the scheme. Using Taylor series expansion, we can write it in the form

$$e(t, \Delta t) = \sum_{i=k}^{2k-1} \Delta t^k \mathcal{D}_i(u(t), t) + \mathcal{O}\left(\Delta t^{2k}\right). \tag{2}$$

Here, $\mathcal{D}_i$ are higher order differential operators. One can now discretize the differential operators in (2) and apply them to the unknown solution. This yields an approximation of the error

$$e_h(u, t, \Delta t) = \sum_{i=k}^{2k-1} \Delta t^k D_{ih}(u(t), t),$$

where D_{ih} are k-th order accurate discretizations of $\mathcal{D}_i$. One way of constructing a $2k$-th order method is to eliminate the lower order error terms by adding $e_h(u, t, \Delta t)$ to the equation to solve, i.e., one solves a discretized version of

$$\begin{aligned} u'(t) &= \mathcal{L}(u(t), t) + e_h(u, t, \Delta t), \\ u(0) &= u_0. \end{aligned}$$

This straightforward method usually will not work. The stability of the method will in general deteriorate very fast. In fact, most methods obtained this way are unstable.

The deferred correction method uses a different way of eliminating the lower order terms in the error. Instead of applying e_h to the unknown approximation, it is applied to the lower order accurate solution $u^{k,n}$ of the original scheme. In the next step, one solves

$$\begin{aligned} u'(t) &= \mathcal{L}(u(t), t) + e_h\left(u^{k,n}, t, \Delta t\right), \\ u(0) &= u_0, \end{aligned}$$

with the same k-th order scheme as before. This yields an approximation of order $2k$. This process can now be repeated, calculating the new truncation error, discretizing the derivatives in it, and evaluating it at the previously calculated approximation. With this procedure, one obtains arbitrarily high order schemes which have good stability properties.

The concept of deferred correction for boundary value problems has been introduced by Fox [8]. It has been taken up and developed by Pereyra in a series of papers [15], [21], [22], [23], [24]. In [5], initial value problems for systems of ODEs are considered. Recently, the deferred correction ideas for initial value problems for ODEs have been used in [4], [6]. As mentioned before, the scheme discussed here has been introduced in [12] and [18]. In [11], the deferred correction principle is used in both space and time. There are several related methods, referred to by *defect correction*, *iterative improvement*, *difference correction*. In [25], a survey over the different methods is given.

The above papers focus mainly on boundary value problems and for initial value problems, the focus is on systems of ODEs. In our work, we consider PDE applications, i.e., the system of ODEs arises from an initial boundary value problem, which has been discretized in space. Although the analysis in this paper is valid for general ODE, we investigate in detail the consequences when the original problem is a PDE.

2.1 Notation

We now summarize some notation used in the following text. We will only consider constant step sizes $\Delta t = t_{n+1} - t_n$. For any function $f(t)$, we use the notation $f^n = f(t_n)$ and $f^{n+1/2} = f\left(t_{n+1/2}\right) = f(t_n + \Delta t/2)$. Difference and averaging operators are defined as

$$\begin{aligned} D_+ u^n &= D_- u^{n+1} = \tfrac{u^{n+1}-u^n}{\Delta t}\,, \\ D_0 u^n &= \tfrac{u^{n+1}-u^{n-1}}{2\Delta t}\,, \\ E_+ u^n &= \tfrac{u^{n+1}+u^n}{2}\,. \end{aligned}$$

The l^2-scalar product of u, $v \in \mathbf{R}^N$ is defined by

$$(u, v) = \frac{1}{N} \sum_{i=1}^{N} u_i v_i\,,$$

and the induced norm is denoted by $\|u\| = \sqrt{(u, u)}$.

3 The BDF2 based deferred correction scheme

We now turn to the case where the underlying method for the deferred correction scheme is the fully implicit second order BDF2 scheme. We describe in detail

how a fourth order time discretization is obtained. For simplicity, consider the linear system

$$\begin{aligned} u'(t) &= A(t)\,u(t) + F(t)\,, \\ u(0) &= u_0\,, \end{aligned} \tag{3}$$

where $A(t) \in \mathbf{R}^{N\times N}$ is a matrix function and $F(t)$, $u(t) \in \mathbf{R}^N$ are vector functions. Discretizing (3) with the BDF2 scheme yields

$$\frac{3}{2}D_+u^{2,n} - \frac{1}{2}D_-u^{2,n} = A(t_{n+1})\,u^{2,n+1} + F^{n+1}\,. \tag{4}$$

The first index on u denotes the order of the method, the second index denotes the time level. In order to solve (4), we need two initial conditions. The first condition will naturally be $u^{2,0} = u_0$. We explain how to choose the second condition later on in this section. Following the procedure described in Section 2, we calculate the local truncation error of the BDF2 scheme by Taylor expansion,

$$\begin{aligned} e(u,t_{n+1},\Delta t) &= \tfrac{3}{2}D_+u(t_n) - \tfrac{1}{2}D_-u(t_n) - A(t_{n+1})\,u(t_{n+1}) - F^{n+1} \\ &= -\tfrac{1}{3}\Delta t^2 u^{(3)}(t_{n+1}) + \tfrac{1}{4}\Delta t^3 u^{(4)}(t_{n+1}) + \mathcal{O}\left(\Delta t^4\right)\,. \end{aligned}$$

Here, $u^{(j)}(t)$ denotes the j-th derivative of u. We now replace the first two error terms by centered difference quotients of the second order accurate solution.

$$\begin{aligned} &\frac{1}{3}\Delta t^2 u^{(3)}(t_{n+1}) - \frac{1}{4}\Delta t^3 u^{(4)}(t_{n+1}) = \\ &\quad \tfrac{1}{3}\Delta t^2 D_+D_-D_0u^{2,n+1} - \tfrac{1}{4}\Delta t^3 \left(D_+D_-\right)^2 u^{2,n+1} + \mathcal{O}\left(\Delta t^4\right)\,. \end{aligned} \tag{5}$$

In doing so, we assume that not only the solution is approximated to order two with the BDF2 scheme, but also that the third difference quotient of the approximate solution is a second order approximation to the third derivative of the exact solution. The exact requirements are discussed in Theorem 3.2. We assume for now that (5) holds.

The fourth order approximation is now obtained by solving

$$\begin{aligned} &\frac{3}{2}D_+u^{4,n} - \frac{1}{2}D_-u^{4,n} = A(t_{n+1})\,u^{4,n+1} + F^{n+1} \\ &\quad -\tfrac{1}{3}\Delta t^2 D_+D_-D_0u^{2,n+1} + \tfrac{1}{4}\Delta t^3 \left(D_+D_-\right)^2 u^{2,n+1}\,. \end{aligned} \tag{6}$$

One can continue the process to obtain arbitrarily high order solutions.

$$\begin{aligned} &\frac{3}{2}D_+u^{2j,n} - \frac{1}{2}D_-u^{2j,n} = A(t_{n+1})\,u^{2j,n+1} + F^{n+1} \\ &\quad + \textstyle\sum_{k=2}^{j}\Big(c_k\Delta t^{2k-2}\left(D_+D_-\right)^{k-1} D_0u^{2j-2,n+1} \\ &\quad + d_k\Delta t^{2k-1}\left(D_+D_-\right)^{k} u^{2j-2,n+1}\Big)\,, \qquad j = 2,3,\ldots\,. \end{aligned} \tag{7}$$

The constants c_k and d_k are determined by the local truncation error of the $(2k-2)$ -nd order deferred correction step. Note that negative time-levels of

$u^{2j-2,n}$ are needed in (7). These can be obtained by extrapolation or solving the PDE backwards in time for the number of time-levels required. A more detailed discussion of this can be found in [12].

In the remainder of this section and the following section, we restrict our observations to the case of time-independent operators A. A generalization to time-dependent A follows in Section 5.

We now present error estimates for the BDF2 based scheme. As mentioned before, we need to make sure that not only the solution itself is approximated to the correct order of accuracy, but that this is also the case for the difference quotients approximating the derivatives which occur in the local error terms. It turns out that this depends highly on the action of A.

In order to obtain a stability estimate for the deferred correction scheme, we need stability for the underlying scheme. For the remainder of this article, we assume that A is *semibounded*, i.e.,

$$(u, Au) \leq 0 \quad \forall u \in \mathbf{R}^N .$$

Similar estimates to the ones obtained in this paper can be obtained with the relaxed condition

$$(u, Au) \leq \alpha \|u\|^2 \quad \forall u \in \mathbf{R}^N .$$

To simplify the proofs, we use the first assumption. As shown in [12], the original problem (3) is wellposed, i.e., for the exact solution, we have

$$\|u(t)\| \leq const \left(\|u_0\| + \max_{s \leq t} \|F(s)\| \right) .$$

The constant may depend on t.

A similar stability estimate holds for the BDF2 scheme.

Theorem 3.1. *If $u^{2,n}$ satisfies (4) and A is semibounded, the stability result*

$$\left\|u^{2,n}\right\| \leq const \left(\max\left(\left\|u^{2,0}\right\|, \left\|u^{2,1}\right\| \right) + \max_{\nu \leq n} \|F^{\nu}\| \right)$$

holds.

Proof. This result can be found in [13, Ch. V.6]. There it is shown that

$$\begin{aligned} E &:= \left(\tfrac{3}{2} D_+ u^{2,n} - \tfrac{1}{2} D_- u^{2,n}, u^{2,n+1} \right) \Delta t \\ &= \left\| \begin{pmatrix} u^{2,n+1} \\ u^{2,n} \end{pmatrix} \right\|_G^2 - \left\| \begin{pmatrix} u^{2,n} \\ u^{2,n-1} \end{pmatrix} \right\|_G^2 \\ &\quad + \left\| \tfrac{1}{2} u^{2,n+1} - u^{2,n} + \tfrac{1}{2} u^{2,n-1} \right\|^2 , \end{aligned}$$

where we consider the norm

$$\left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|_G^2 = g_{11}(u,u) + 2 g_{12}(u,v) + g_{22}(v,v) ,$$

with

$$G = \frac{1}{4}\begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}.$$

This norm is equivalent to the usual l^2-norm. On the other hand, we have

$$\begin{aligned} E &= \left(Au^{2,n+1}, u^{2,n+1}\right)\Delta t + \left(F^{n+1}, u^{2,n+1}\right)\Delta t \leq \left\|F^{n+1}\right\|\left\|u^{2,n+1}\right\|\Delta t \\ &\leq c\left\|F^{n+1}\right\|\left(\left\|\begin{pmatrix} u^{2,n+1} \\ u^{2,n} \end{pmatrix}\right\|_G + \left\|\begin{pmatrix} u^{2,n} \\ u^{2,n-1} \end{pmatrix}\right\|_G\right)\Delta t\,, \end{aligned}$$

where we have used the equivalence of the norms. Thus,

$$\left\|\begin{pmatrix} u^{2,n+1} \\ u^{2,n} \end{pmatrix}\right\|_G - \left\|\begin{pmatrix} u^{2,n} \\ u^{2,n-1} \end{pmatrix}\right\|_G \leq c\left\|F^{n+1}\right\|\Delta t\,.$$

Because of the equivalence of norms, we obtain

$$\begin{aligned} \left\|u^{2,n+1}\right\| &\leq c\left\|\begin{pmatrix} u^{2,n+1} \\ u^{2,n} \end{pmatrix}\right\|_G \\ &\leq c\left(\left\|\begin{pmatrix} u^{2,1} \\ u^{2,0} \end{pmatrix}\right\|_G + const\,\max_{\nu\leq n+1}\left\|F^{\nu}\right\|\right) \\ &\leq const\left(\max\left(\left\|u^{2,0}\right\|, \left\|u^{2,1}\right\|\right) + \max_{\nu\leq n+1}\left\|F^{\nu}\right\|\right). \end{aligned}$$

□

Using the result for the BDF2 scheme, we can now give estimates for the difference quotient of the errors for the deferred correction scheme based on the BDF2 method.

Theorem 3.2. *Let $u^{2j,n}$ be the solution to the $2j$-th order deferred correction method (7) based on the BDF2 scheme and let A be independent of t and semi-bounded. The error $e^{2j,n} = u(t_n) - u^{2j,n}$ satisfies*

$$\begin{aligned} &\left\|D_+^p e^{2j,n}\right\| \leq \\ &\quad const\left(\sum_{k=0}^{j-1}\Delta t^{2k}\max\left(\left\|D_+^{p+3k}e^{2(j-k),0}\right\|, \left\|D_+^{p+3k}e^{2(j-k),1}\right\|\right)\right. \\ &\quad \left.+\Delta t^{2j}\max_t\left\|u^{(p+3j)}(t)\right\|\right) + \mathcal{O}\left(\Delta t^{2j+1}\right), \quad p = 0, 1, \ldots\,. \end{aligned} \tag{8}$$

Here, $u^{(p+3j)}(t)$ denotes the $(p+3j)$-th derivative of the exact solution to (3). An estimate for $e^{2j,n}$ can be found by setting $p = 0$.

Proof. The proof is done via induction over j. For $j = 1$, the difference quotient $D_+^p e^{2,n}$ satisfies the following equation.

$$\begin{aligned} &\frac{3}{2}D_+D_+^p e^{2,n} - \frac{1}{2}D_-D_+^p e^{2,n} = AD_+^p e^{2,n+1} \\ &\qquad -\Delta t^2 c_2 D_+^p u'''(t_{n+1}) - \Delta t^3 d_2 D_+^p u^{(4)}(t_{n+1}) + \mathcal{O}\left(\Delta t^4\right). \end{aligned}$$

From Theorem 3.1, we obtain

$$\begin{aligned}\left\|D_+^p e^{2,n}\right\| &\leq const\Bigg(\max\left(\left\|D_+^p e^{2,0}\right\|,\left\|D_+^p e^{2,1}\right\|\right)\\ &\quad +\Delta t^2\left(\max_{t\leq t_n}\left\|D_+^p u'''(t)\right\|+\Delta t\max_{t\leq t_n}\left\|D_+^p u^{(4)}(t)\right\|\right)\Bigg)+\mathcal{O}\left(\Delta t^4\right)\\ &\leq const\Bigg(\max\left(\left\|D_+^p e^{2,0}\right\|,\left\|D_+^p e^{2,1}\right\|\right)\\ &\quad +\Delta t^2\max_t\left\|u^{(p+3)}(t)\right\|\Bigg)+\mathcal{O}\left(\Delta t^3\right),\end{aligned}$$

where we assume smoothness of the exact solution u. Assume now that (8) is true for j substituted by $j-1$. The difference quotient $D_+^p e^{2j,n}$ fulfills

$$\frac{3}{2}D_+D_+^p e^{2j,n}-\frac{1}{2}D_-D_+^p e^{2j,n}=AD_+^p e^{2j,n+1}+F_1^{n+1}+F_2^{n+1},$$

where

$$\begin{aligned}F_1^n &= \sum_{k=2}^{j}c_k\Delta t^{2k-2}\left(D_+D_-\right)^{k-1}D_0\left(D_+^p e^{2j-2,n}\right)\\ &\quad +d_k\Delta t^{2k-1}\left(D_+D_-\right)^k\left(D_+^p e^{2j-2,n}\right),\\ F_2^n &= -c_{j+1}\Delta t^{2j}D_+^p u^{(2j+1)}(t_n)-d_{j+1}\Delta t^{2j+1}D_+^p u^{(2j+2)}(t_n)\\ &\quad +\mathcal{O}\left(\Delta t^{2j+2}\right).\end{aligned}$$

With Theorem 3.1, we obtain the estimate

$$\begin{aligned}\left\|D_+^p e^{2j,n}\right\| &\leq const\Bigg(\max\left(\left\|D_+^p e^{2j,0}\right\|,\left\|D_+^p e^{2j,1}\right\|\right)\\ &\quad +\max_{\nu\leq n}\left(\|F_1^\nu\|+\|F_2^\nu\|\right)\Bigg).\end{aligned}\tag{9}$$

For F_2^n, we easily obtain the following bound, for the case where the exact solution is smooth.

$$\max_n\|F_2^n\| \leq const\Delta t^{2j}\left(\max_t\left\|u^{(p+2j+1)}(t)\right\|\right)+\mathcal{O}\left(\Delta t^{2j+1}\right).$$

For F_1^n, we have

$$\begin{aligned}\max_n\|F_1^n\| &\leq const\left(\max_n\sum_{k=2}^{j}\Delta t^{2k-2}\left\|D_+^{p+2k-1}e^{2j-2,n}\right\|\right)\\ &\leq \Delta t^2 const\left(\max_n\left\|D_+^{p+3}e^{2j-2,n}\right\|\right),\end{aligned}$$

and we can use the induction assumption to arrive at

$$\begin{aligned}
&\max_n \|F_1^n\| \leq \\
&\quad const \quad \Delta t^2 \left(\sum_{k=0}^{j-2} \Delta t^{2k} \max \left(\left\| D_+^{p+3+3k} e^{2(j-1-k),0} \right\|, \left\| D_+^{p+3+3k} e^{2(j-1-k),1} \right\| \right) \right. \\
&\qquad \left. + \Delta t^{2j-2} \max_t \left\| u^{(p+3+3(j-1))}(t) \right\| \right) + \mathcal{O}\left(\Delta t^{2j+1}\right) \\
&\quad = \quad const \left(\sum_{k=1}^{j-1} \Delta t^{2k} \max \left(\left\| D_+^{p+3k} e^{2(j-k),0} \right\|, \left\| D_+^{p+3k} e^{2(j-k),1} \right\| \right) \right. \\
&\qquad \left. + \Delta t^{2j} \max_t \left\| u^{(p+3j)}(t) \right\| \right) + \mathcal{O}\left(\Delta t^{2j+1}\right) .
\end{aligned}$$

Inserting this into (9), we see that (8) is satisfied for general j and the proof is complete. □

Theorem 3.2 gives an estimate of the error in terms of the exact solution u and the initial error. No restriction on the time-step is needed. In order to achieve the desired order of accuracy for $u^{2j,n}$, we need sufficient smoothness of the exact solution. In addition, we need to establish that

$$\max \left(\left\| D_+^{3k} e^{2(j-k),0} \right\|, \left\| D_+^{3k} e^{2(j-k),1} \right\| \right) = \mathcal{O}\left(\Delta t^{2(j-k)}\right) .$$

This requirement will lead to conditions for choosing $u^{2j,1}$. Note that the choice of $u^{2j,1}$ is crucial for the deferred correction principle to work, and the usual technique of using a one step method as a start up scheme will in general not give good results. This will also be seen in the experiments presented in Section 7. It turns out that when choosing

$$u^{2,1} = u(0) + u'(0)\Delta t + u^{(2)}(0)\frac{\Delta t^2}{2} + u^{(3)}(0)\frac{\Delta t^3}{2} + \frac{\Delta t^4}{24}\left(-u^{(4)}(0) + 4Au^{(3)}(0)\right), \tag{10}$$

then $D_+^3 e^{2,0}$ and $D_+^3 e^{2,1}$ are of order Δt^2. The derivation of this can be found in Appendix A. The time derivatives of the initial data are not naturally given in the problem, but can be obtained by using the original differential equation (3) at $t = 0$. For the sixth order deferred correction scheme, a similar argumentation leads to the choice

$$\begin{aligned}
u^{2,1} \quad = \quad & u(0) + u'(0)\Delta t + u^{(2)}(0)\tfrac{\Delta t^2}{2} + u^{(3)}(0)\tfrac{\Delta t^3}{2} \\
& + \left(-\tfrac{1}{24}u^{(4)}(0) + \tfrac{1}{6}Au^{(3)}(0)\right)\Delta t^4 \\
& + \left(\tfrac{1}{6}u^{(5)}(0) + \tfrac{1}{24}Au^{(4)}(0) + \tfrac{1}{6}A^2u^{(3)}(0)\right)\Delta t^5 \\
& + \left(-\tfrac{17}{240}u^{(6)}(0) - \tfrac{1}{40}Au^{(5)}(0) - \tfrac{1}{12}A^2u^{(4)}(0) + \tfrac{1}{24}A^3u^{(3)}(0)\right)\Delta t^6 \\
& + \left(\tfrac{167}{1440}u^{(7)}(0) + \tfrac{109}{1440}Au^{(6)}(0) \right. \\
& \left. + \tfrac{73}{480}A^2u^{(5)}(0) + \tfrac{3}{32}A^3u^{(4)}(0) + \tfrac{1}{8}A^4u^{(3)}(0) \right)\Delta t^7 ,
\end{aligned} \tag{11}$$

and

$$\begin{aligned} u^{4,1} &= u(0) + u'(0)\Delta t + u^{(2)}(0)\tfrac{\Delta t^2}{2} + u^{(3)}(0)\tfrac{\Delta t^3}{6} \\ &\quad + u^{(4)}(0)\tfrac{\Delta t^4}{24} \\ &\quad + \left(-\tfrac{1}{9}Au^{(4)}(0) - \tfrac{5}{72}u^{(5)}(0) - \tfrac{1}{9}A^2u^{(3)}(0)\right)\Delta t^5 \\ &\quad + \left(\tfrac{31}{240}u^{(6)}(0) + \tfrac{13}{180}Au^{(5)}(0) + \tfrac{1}{18}A^2u^{(4)}(0) - \tfrac{1}{36}A^3u^{(3)}(0)\right)\Delta t^6 . \end{aligned} \tag{12}$$

See Appendix A for the detailed derivation. When choosing $u^{2,1}$ and $u^{4,1}$ as in (11) and (12), one can show by Taylor expansion around $\Delta t = 0$ that

$$\max\left(\left\|D_+^6 e^{2,0}\right\|, \left\|D_+^6 e^{2,1}\right\|\right) \approx \left\|A^5 u^{(3)}(0)\right\| \Delta t^2 ,$$

and

$$\max\left(\left\|D_+^3 e^{4,0}\right\|, \left\|D_+^3 e^{4,1}\right\|\right) \approx \left\|A^4 u^{(3)}(0)\right\| \Delta t^4 .$$

We have left out terms including lower powers of A and higher powers of Δt. Inserting this into (8) for $j = 3$ and $p = 0$, we obtain the estimate

$$\left\|e^{6,n}\right\| \leq const \Delta t^6 \left(\left\|A^5 u^{(3)}(0)\right\| + \max_t \left\|u^{(9)}(t)\right\|\right) + \mathcal{O}\left(\Delta t^6\right) , \tag{13}$$

again having omitted terms including lower powers of A. We see here that the sixth order deferred correction error is indeed of order six if we have a uniform bound on $A^5 u^{(3)}(0)$. Thus, the performance of the method depends on the smoothness of the exact solution and on the boundedness of the initial data under application of the discrete differential operator A. For operators A arising from the spatial discretization of a PDE, the bound should be independent of the spatial grid-size. This will be discussed in more detail in Section 6.

4 Deferred correction based on the implicit midpoint rule

We now derive error estimates for the IMR based scheme. The implicit midpoint rule for discretizing a system of linear ODEs (3) is

$$D_+ u^{2,n} = AE_+ u^{2,n} + F^{n+1/2} . \tag{14}$$

Again, we assume constant A. Time-dependent A are discussed in Section 5. The scheme is A-stable and we have the following stability estimate for semibounded matrices A.

$$\left\|u^{2,n}\right\| \leq \left\|u^{2,0}\right\| + \sum_{\nu=0}^{n-1} \left\|F^{\nu+1/2}\right\| \Delta t . \tag{15}$$

The deferred correction method works in a similar way as for the BDF2 case. The local truncation error for the second order IMR is

$$D_+ u(t_n) - AE_+ u(t_n) - F^{n+1/2} = \frac{\Delta t^2}{24} u^{(3)}\left(t_{n+1/2}\right) - \frac{\Delta t^2}{8} A u^{(2)}\left(t_{n+1/2}\right) + \mathcal{O}\left(\Delta t^4\right) .$$

The first deferred correction step is solving

$$\begin{aligned} D_+u^{4,n} &= AE_+u^{4,n} + F^{n+1/2} \\ &\quad + \tfrac{\Delta t^2}{24} D_+D_+D_-u^{2,n} - \tfrac{\Delta t^2}{8} AE_+D_+D_-u^{2,n}\,. \end{aligned}$$

Higher order solutions can be obtained by repeating the process of calculating the local truncation error and inserting the previously calculated solution,

$$\begin{aligned} D_+u^{2j,n} &= AE_+u^{2j,n} + F^{n+1/2} \\ &\quad + \textstyle\sum_{k=2}^{j}\Big(c_k\Delta t^{2k-2}D_+\left(D_+D_-\right)^{k-1}u^{2j-2,n} \\ &\quad + d_k\Delta t^{2k-2}AE_+\left(D_+D_-\right)^{k-1}u^{2j-2,n}\Big)\,. \end{aligned} \tag{16}$$

In [12] and [18], a stability analysis was performed. The main stability result is the following theorem.

Theorem 4.1. *Assume that A is constant and semibounded. Then the solution of order p of the deferred correction algorithm satisfies the estimate*

$$\begin{aligned} \|u^{p,n}\| \leq const\Bigg(&\max_{j\leq p/2-1}\left\|A^ju_0\right\| \\ &+ \max\nolimits_{\substack{0\leq\nu\leq n-1+p(p-2)/8\\ j\leq p/2-1}}\left\|A^jF^{\nu+1/2}\right\|\Bigg)\,. \end{aligned} \tag{17}$$

Here, the constant depends on p and on t_n but not on A.

Error estimates were not proven in [12] and [18]. As for the BDF2 based scheme, one has to show smoothness of the intermediate lower order solutions. In the following, we give estimates for the error $e^{2j,n} = u\left(t_n\right) - u^{2j,n}$. The error fulfills the equation

$$\begin{aligned} D_+e^{2j,n} &= AE_+e^{2j,n} + \sum_{k=2}^{j}\Big(c_k\Delta t^{2k-2}D_+\left(D_+D_-\right)^{k-1}e^{2j-2,n} \\ &\quad + d_k\Delta t^{2k-2}AE_+\left(D_+D_-\right)^{k-1}e^{2j-2,n}\Big) \\ &\quad - c_{j+1}\Delta t^{2j}u^{(2j+1)}\left(t_{n+1/2}\right) - d_{j+1}\Delta t^{2j}Au^{(2j)}\left(t_{n+1/2}\right) + \mathcal{O}\left(\Delta t^{2j+2}\right)\,. \end{aligned} \tag{18}$$

To prove the main estimate, we need the following lemma which holds for the solution $u^{2,n}$ to the IMR scheme. The lemma has already been proven in [12].

Lemma 4.1. *Assume that A is a semibounded matrix. The divided differences of the solution of (14) satisfy the estimate*

$$\begin{aligned} \|D_+^pu^{2,n}\| &\leq \|A^pu_0\| + \sum_{l=0}^{p-1}(\|A^pF^{l+1/2}\|\Delta t + \|(AE_+)^lD_+^{p-l-1}F^{1/2}\|) \\ &\quad + \sum_{k=0}^{n-1}\|D_+^pF^{k+1/2}\|\Delta t, \quad j=0,1,2,\ldots. \end{aligned}$$

Proof. We have by induction

$$\begin{aligned} D_+(D_+^p u^{2,n}) &= (AE_+)(D_+^p u^{2,n}) + D_+^p F^{n+1/2}, \\ D_+^p u^{2,0} &= (AE_+)^p u^{2,0} + \sum_{l=0}^{p-1} (AE_+)^l D_+^{p-l-1} F^{1/2}. \end{aligned} \tag{19}$$

After multiplication of (14) by A^p and applying estimate (15), we immediately obtain

$$\|A^p u^{2,n}\| \le \|A^p u_0\| + \sum_{l=0}^{n-1} \|A^p F^{l+1/2}\| \Delta t,$$

and therefore

$$\|(AE_+)^p u^{2,0}\| \le \|A^p E_+^p u^{2,0}\| \le \max_{0\le k\le p} \|A^p u^{2,k}\| \le \|A^p u_0\| + \sum_{l=0}^{p-1} \|A^p F^{l+1/2}\| \Delta t.$$

The final estimate follows by applying estimate (15) to (19). □

We can now prove the following error estimate for the IMR based deferred correction method.

Theorem 4.2. *Let $u^{2j,n}$ be the $2j$-th order solution to the deferred correction method (16) based on the implicit midpoint rule. Then the error $e^{2j,n} = u(t_n) - u^{2j,n}$ satisfies*

$$\begin{aligned} \left\|D_+^p e^{2j,n}\right\| \le const \Bigg(& \max_{\substack{0\le i\le j-1 \\ i\le k\le j-1}} \left\|A^{3k+p-i} e^{2(j-k),0}\right\| \Delta t^{2k} \\ & + \Delta t^{2j} \Big(\max_{\substack{0\le l\le k \\ 1\le k\le j}} \left\|A^l u^{(2j+p+k-l)}(t)\right\| \\ & + \max_{\substack{0\le s\le t_\nu \\ 0\le i\le j-1 \\ i\le k\le j-1 \\ 0\le l\le 3k+p-i}} \left\|A^l u^{(k+p-l+2j-i)}(s)\right\| \Big) \Bigg) + \mathcal{O}(\Delta t^{2j+1}). \end{aligned} \tag{20}$$

Here, t_ν denotes a small number that depends on the deferred correction step j but not on t_n and the constant depends on t_n but not on A.

Proof. We do the proof by induction over j. For $j = 1$, we use Lemma 4.1 to obtain

$$\begin{aligned} \left\|D_+^p e^{2,n}\right\| \le const \Bigg(& \left\|A^p e^{2,0}\right\| + \max_{0\le l\le p-1} \left\|A^p (c_2 u''' + d_2 A u'')(t_{l+1/2})\right\| \Delta t^3 \\ & + \max\nolimits_{0\le l\le p-1} \left\|(AE_+)^l D_+^{p-l-1} (c_2 u''' + d_2 A u'')(t_{1/2})\right\| \Delta t^2 \\ & + \max\nolimits_t \left\|D_+^p (c_2 u''' + d_2 A u'')(t)\right\| \Delta t^2 \Bigg) + \mathcal{O}(\Delta t^4). \end{aligned}$$

The second term is $\mathcal{O}(\Delta t^3)$ and the third term can be estimated by

$$\begin{aligned} \max_{0\le l\le p-1} \left\|(AE_+)^l D_+^{p-l-1} (c_2 u''' + d_2 A u'')(t_{1/2})\right\| \Delta t^2 \le \\ const \max\nolimits_{\substack{0\le s\le t_\nu \\ 0\le l\le p}} \left\|A^l u^{(p-l+2)})(s)\right\| \Delta t^2, \end{aligned}$$

where t_ν is a small number independent of n. We obtain

$$\begin{aligned}\left\|D_+^p e^{2,n}\right\| &\leq const\Bigg(\left\|A^p e^{2,0}\right\| + \max_{\substack{0\leq s\leq t_\nu\\ 0\leq l\leq p}} \left\|A^l u^{(p-l+2)}(s)\right\| \Delta t^2 \\ &\quad + \max_{\substack{t\\ 0\leq l\leq 1}} \left\|A^l u^{(p+3-l)}(t)\right\| \Delta t^2\Bigg) + \mathcal{O}(\Delta t^3)\,.\end{aligned}$$

For the induction step from $j-1$ to j, we again use Lemma 4.1 for $e^{2j,n}$ for $j \geq 2$ to obtain

$$\begin{aligned}\|D_+^p e^{2j,n}\| \leq const\Bigg(&\|A^p e^{2j,0}\| \\ &+ \max_{0\leq l\leq p-1} \|A^p(c_{j+1}u^{(2j+1)} + d_{j+1}Au^{(2j)})(t_{l+1/2})\|\Delta t^{2j+1} \\ &+ \max_{0\leq l\leq p-1} \|A^l E_+^l D_+^{p-l-1}(c_{j+1}u^{(2j+1)} + d_{j+1}Au^{(2j)})(t_{1/2})\|\Delta t^{2j} \\ &+ \max_t(\|D_+^p(c_{j+1}u^{(2j+1)} + d_{j+1}Au^{(2j)})(t)\|)\Delta t^{2j} \\ &+ I + II + III\Bigg) + \mathcal{O}(\Delta t^{2j+2})\,,\end{aligned} \tag{21}$$

where

$$\begin{aligned} I &= const \max_{0\leq l\leq p-1}\Bigg(\sum_{k=2}^{j} \Delta t^{2k+1}(\|A^p(D_+D_-)^{k-1}D_0 e^{2j-2,l}\| \\ &\quad + \|A^{p+1}E_+(D_+D_-)^{k-1}e^{2j-2,l}\|)\Bigg) \\ &\leq const \max_{0\leq l\leq \nu} \Delta t^2(\|A^p D_+^2 e^{2j-2,l}\| + \|A^{p+1}D_+ e^{2j-2,l}\|)\,,\end{aligned}$$

and

$$\begin{aligned} II &= const \max_{0\leq l\leq p-1}\Bigg(\sum_{k=2}^{j} \Delta t^{2k+1}(\|(AE_+)^l(D_+D_-)^{k-1}D_0 D_+^{p-1-l} e^{2j-2,l}\| \\ &\quad + \|(AE_+)^{l+1}(D_+D_-)^{k-1}D_+^{p-1-l}e^{2j-2,l}\|)\Bigg) \\ &\leq const \max_{\substack{0\leq l\leq p-1\\ 0\leq m\leq \nu}} (\|A^l E_+^l D_+^{p-l+2} e^{2j-2,m}\| \\ &\quad + \|A^{l+1}E_+^l D_+^{p-l+1} e^{2j-2,m}\|)\Delta t^2\,,\end{aligned}$$

where we have introduced ν as a small number. We have

$$\begin{aligned} I+II &\leq const\left(\max_{\substack{m\leq\nu\\0\leq l\leq p+1}}\|A^l D_+^{p-l+2}e^{2j-2,m}\|\Delta t^2\right)\\ &\leq const\Bigg(\max_{\substack{0\leq i\leq j-2\\ i\leq k\leq j-2}}\left\|A^{3k+p+2-i}e^{2(j-1-k),0}\right\|\Delta t^{2k+2}\\ &\quad+\Delta t^{2j}\Bigg(\max_{\substack{0\leq s\leq t_\nu\\0\leq l\leq p+1\\0\leq l'\leq k\\1\leq k\leq j-1}}\left\|A^{l+l'}u^{(2j-2+p-l+2+k-l')}(s)\right\|\\ &\quad+\max_{\substack{0\leq s\leq t_\nu\\0\leq l\leq p+1\\0\leq i\leq j-2\\ i\leq k\leq j-2\\0\leq l'\leq 3k+p-l+2-i}}\left\|A^{l'+l}u^{(k+p-l+2-l'+2j-2-i)}(s)\right\|\Bigg)\Bigg)\\ &\quad+\mathcal{O}(\Delta t^{2j+1})\,. \end{aligned}$$

Here, we have used the induction assumption for j replaced by $j-1$. We now perform variable transformations $i+1\to i$ and $l+l'\to l$ and, for the first and last term, we use a variable transformation $k+1\to k$ to obtain

$$\begin{aligned} I+II &\leq const\Bigg(\max_{\substack{1\leq i\leq j-1\\ i\leq k\leq j-1}}\left\|A^{3k+p-i}e^{2(j-k),0}\right\|\Delta t^{2k}\\ &\quad+\Delta t^{2j}\Bigg(\max_{\substack{0\leq s\leq t_\nu\\1\leq k\leq j-1\\0\leq l\leq p+k+1}}\left\|A^l u^{(2j+p-l+k)}(s)\right\|\\ &\quad+\max_{\substack{0\leq s\leq t_\nu\\1\leq i\leq j-1\\ i\leq k\leq j-1\\0\leq l\leq 3k+p-i}}\left\|A^l u^{(k+p-l+2j-i)}(s)\right\|\Bigg)\Bigg)+\mathcal{O}(\Delta t^{2j+1})\\ &\leq const\Bigg(\max_{\substack{1\leq i\leq j-1\\ i\leq k\leq j-1}}\left\|A^{3k+p-i}e^{2(j-k),0}\right\|\Delta t^{2k}\\ &\quad+\max_{\substack{0\leq s\leq t_\nu\\0\leq i\leq j-1\\ i\leq k\leq j-1\\0\leq l\leq 3k+p-i}}\left\|A^l u^{(k+p-l+2j-i)}(s)\right\|\Delta t^{2j}\Bigg)+\mathcal{O}(\Delta t^{2j+1})\,. \end{aligned}$$

For III, we have

$$\begin{aligned}
III &= const \max_n \Bigg(\sum_{k=2}^{j} \Delta t^{2k+1} (\|D_+^p (D_+D_-)^{k-1} D_0 e^{2j-2,n}\| \\
&\quad + \|AD_+^p E_+ (D_+D_-)^{k-1} e^{2j-2,l}\|) \Bigg) \\
&\leq const \max_n (\|D_+^{p+3} e^{2j-2,n}\| + \|AD_+^{p+2} e^{2j-2,n}\|) \Delta t^2 \\
&\leq const \max_{\substack{0\leq i\leq j-2 \\ i\leq k\leq j-2}} \Bigg(\left\|A^{3k+3+p-i} e^{2(j-1-k),0}\right\| \Delta t^{2k+2} \\
&\quad + \Delta t^{2j} \Big(\max_{\substack{t \\ 0\leq l\leq k \\ 1\leq k\leq j-1}} (\left\|A^l u^{(2j-2+p+3+k-l)}(t)\right\| \\
&\quad + \left\|A^{l+1} u^{(2j-2+p+2+k-l)}(t)\right\|) \\
&\quad + \max_{\substack{0\leq s\leq t_\nu \\ 0\leq i\leq j-2 \\ i\leq k\leq j-2 \\ 0\leq l\leq 3k+p+2-i}} (\left\|A^l u^{(k+p+3-l+2j-2-i)}(s)\right\| \\
&\quad + \left\|A^{l+1} u^{(k+p+2-l+2j-2-i)}(s)\right\|) \Big) \Bigg) + \mathcal{O}(\Delta t^{2j+1}) .
\end{aligned}$$

Again, performing variable transforms, one obtains,

$$\begin{aligned}
III &\leq const \Bigg(\max_{\substack{0\leq i\leq j-2 \\ i+1\leq k\leq j-1}} \left\|A^{3k+p-i} e^{2(j-k),0}\right\| \Delta t^{2k} \\
&\quad + \Delta t^{2j} \Big(\max_{\substack{t \\ 2\leq k\leq j \\ 0\leq l\leq k}} \left\|A^l u^{(2j+p+k-l)}(t)\right\| \\
&\quad + \max_{\substack{0\leq s\leq t_\nu \\ 0\leq i\leq j-2 \\ i+1\leq k\leq j-1 \\ 0\leq l\leq 3k+p-i}} \left\|A^l u^{(k+p-l+2j-i)}(s)\right\| \Big) \Bigg) + \mathcal{O}(\Delta t^{2j+1}) .
\end{aligned}$$

For the remaining terms in estimate (21), we have

$$\begin{aligned}
\max_{0\leq l\leq p-1} \|A^l E_+^l D_+^{p-l-1} (c_{j+1} u^{(2j+1)} + d_{j+1} A u^{(2j)})(t_{1/2})\| \Delta t^{2j} \leq \\
const \max_{\substack{0\leq s\leq t_\nu \\ 0\leq l\leq p}} \|A^l u^{(2j+p-l)}(s)\| \Delta t^{2j} ,
\end{aligned}$$

and

$$\begin{aligned}
\max_t (\|D_+^p (c_{j+1} u^{(2j+1)}(t) + d_{j+1} A u^{(2j)})(t)\|) \Delta t^{2j} \leq \\
const \max_{\substack{t \\ 0\leq l\leq 1 \\ 1\leq k\leq 1}} \|A^l u^{(2j+p+k-l)}(t)\| \Delta t^{2j} .
\end{aligned}$$

Combining these results, one obtains that estimate (20) also holds for j. □

As in the BDF case, the estimate depends on the smoothness of the initial steps and the exact solution. In addition to terms, where A is applied to the initial data, there are also terms present, where A is applied to the exact solution of the semidiscrete equation at later time-levels.

5 Time-dependent coefficients

The estimates in Sections 3 and 4 have been proven for time-independent A for simplicity. In this section, we show how to generalize the proofs to time-dependent $A(t)$. We restrict ourselves to the IMR based scheme.

We have in mind a PDE and need to obtain estimates that are independent of the space-step Δx. In Theorem 5.1, we give an estimate for time-dependent $A(t)$ of the form $A(t) = A_1(t) Q$, where $A_1(t)$ is a smooth matrix function that is uniformly bounded for all Δx and Q is a time-independent, not necessarily bounded operator. The operator Q usually corresponds to a discretization of a spatial differential operator and the matrix function $A(t)$ contains the coefficients. For the simple example $u_t = a(x,t)\,\partial_x u(x,t)$, we could have

$$A(t) = \begin{pmatrix} a(x_1,t) & 0 & 0 & 0 \\ 0 & a(x_2,t) & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & a(x_n,t) \end{pmatrix} Q\,,$$

where Q is a discretization of ∂_x.

We introduce the following commutators, which are used in the next theorem.

$$\begin{aligned} [A,B]_0 &= B\,, \\ [A,B]_1 &= [A,B] = AB - BA\,, \\ [A,B]_j &= [A,[A,B]_{j-1}] \quad j = 2,3,\ldots\,. \end{aligned}$$

We have the following result, which is used in the proof of the next theorem.

Lemma 5.1. *For matrices Q and A, we have*

$$Q^l A = \sum_{j=0}^{l} \binom{l}{j} [Q,A]_j Q^{l-j} \quad l = 0,1,\ldots\,.$$

Proof. We do the proof by induction over l. For $l = 0$, the statement is trivial. Assume now that the statement of the lemma holds for l replaced with $l-1$. We now show that it also holds for l. We have

$$\begin{aligned} Q^l A &= Q\left(Q^{l-1}A\right) = \textstyle\sum_{j=0}^{l-1} \binom{l-1}{j} Q[Q,A]_j Q^{l-1-j} \\ &= \textstyle\sum_{j=0}^{l-1} \binom{l-1}{j} [Q,A]_{j+1} Q^{l-(j+1)} + \sum_{j=0}^{l-1} \binom{l-1}{j} [Q,A]_j Q^{l-j} \\ &= \textstyle\sum_{j=1}^{l} \binom{l-1}{j-1} [Q,A]_j Q^{l-j} + \sum_{j=0}^{l-1} \binom{l-1}{j} [Q,A]_j Q^{l-j} \\ &= \textstyle\sum_{j=1}^{l-1} \left(\binom{l-1}{j-1} + \binom{l-1}{j}\right) [Q,A]_j Q^{l-j} \\ &\quad + \binom{l-1}{l-1} [Q,A]_l + \binom{l-1}{0} AQ^l\,. \end{aligned}$$

We now use the fact that

$$\binom{l-1}{l-1} + \binom{l-1}{j} = \binom{l}{j},$$

and

$$\binom{l-1}{l-1} = \binom{l}{l} = \binom{l-1}{0} = \binom{l}{0} = 1,$$

to write

$$\begin{aligned} Q^l A &= \sum_{j=1}^{l-1} \binom{l}{j} [Q,A]_j Q^{l-j} + [Q,A]_l + AQ^l \\ &= \sum_{j=0}^{l} \binom{l}{j} [Q,A]_j Q^{l-j}. \end{aligned}$$

□

We can now prove the following estimate, which is the basic ingredient to the proofs in Sections 3 and 4. The estimate can be generalized to operators of the form $A(t) = \sum_{i=1}^{M} A_i(t) Q_i$, the proof becoming quite technical.

Theorem 5.1. *Let u^n be the solution to*

$$D_+ u^n = A\left(t_{n+1/2}\right) Q E_+ u^n + F^{n+1/2}, \tag{22}$$

with $A(t)Q$ semibounded for all $t \geq 0$, where $A(t)$ is a smooth matrix function. The derivatives of $A(t)$ are assumed to be uniformly bounded for all Δx. Under the assumption that the commutators $[Q,A]_j$ are uniformly bounded for sufficiently high j, we have

$$\left\| Q^l D_+^m u^n \right\| \leq K(t_n) \left(\max_{\substack{l'+m' \leq l+m \\ m' \leq m}} \left(\left\| Q^{l'} D_+^{m'} u^0 \right\| + \max_{n'} \left\| Q^{l'} D_+^{m'} F^{n'+1/2} \right\| \right) \right). \tag{23}$$

Proof. We do the proof by induction over l and m. For $l = m = 0$, the statement follows by the stability of the implicit midpoint rule. First, let $m = 0$ and assume that (23) holds for l replaced with $l-1$. To show that the statement holds for l, we use the difference equation (22). We write $A(t) = A$ as a simplification. We have

$$D_+ Q^l u^n = Q^l D_+ u^n = Q^l A Q E_+ u^n + Q^l F^{n+1/2}.$$

As proven in Lemma 5.1,

$$\begin{aligned} D_+ Q^l u^n &= \sum_{j=0}^{l} \binom{l}{j} [Q,A]_j E_+ Q^{l-j+1} u^n + Q^l F^{n+1/2} \\ &= AQE_+ Q^l u^n + l[Q,A] E_+ Q^l u^n \\ &\quad + \sum_{j=2}^{l} \binom{l}{j} [Q,A]_j E_+ Q^{l-j+1} u^n + Q^l F^{n+1/2}. \end{aligned} \tag{24}$$

Using the induction assumption and the fact that $[Q,A]_j$ are bounded, we obtain

$$\left\|\sum_{j=2}^{l}\binom{l}{j}[Q,A]_j E_+ Q^{l-j+1}u^n\right\| \leq K(t_n)\left(\max_{l'\leq l-1}\left(\left\|Q^{l'}u^0\right\|+\max_{n'}\left\|Q^{l'}F^{n'+1/2}\right\|\right)\right).$$

We now take the scalar product of (24) with $E_+Q^l u^n 2\Delta t$ to obtain

$$\begin{aligned}\left\|Q^l u^{n+1}\right\|^2 - \left\|Q^l u^n\right\|^2 &\leq l\left\|[Q,A]\right\|_\infty \frac{\Delta t}{2}\left(\left\|Q^l u^{n+1}\right\|+\left\|Q^l u^n\right\|\right)^2\\ &+\left\|\textstyle\sum_{j=2}^{l}\binom{l}{j}[Q,A]_j E_+ Q^{l-j+1}u^n\right\|\left(\left\|Q^l u^{n+1}\right\|+\left\|Q^l u^n\right\|\right)\Delta t\\ &+\left\|Q^l F^{n+1/2}\right\|\left(\left\|Q^l u^{n+1}\right\|+\left\|Q^l u^n\right\|\right)\Delta t,\end{aligned}$$

where $\left\|[Q,A]\right\|_\infty$ is the uniform bound on $[Q,A]$. From this we obtain

$$\begin{aligned}\left\|Q^l u^{n+1}\right\| &\leq \frac{1+\frac{\Delta t}{2}l\left\|[Q,A]\right\|_\infty}{1-\frac{\Delta t}{2}l\left\|[Q,A]\right\|_\infty}\left\|Q^l u^n\right\|\\ &\quad+\frac{K(t_n)\Delta t}{1-\frac{\Delta t}{2}l\left\|[Q,A]\right\|_\infty}\left(\max_{l'\leq l-1}\left(\left\|Q^{l'}u^0\right\|+\max_{l'\leq l}\max_{n'}\left\|Q^{l'}F^{n'+1/2}\right\|\right)\right)\\ &\leq K(t_n)\max_{l'\leq l}\left(\left\|Q^{l'}u^0\right\|+\max_{n'}\left\|Q^{l'}F^{n'+1/2}\right\|\right),\end{aligned}$$

where $K(t)$ is used to denote a generic constant. The above statement is of course only true if $1-\frac{\Delta t}{2}l\left\|[Q,A]\right\|_\infty > 0$. This is, however, true for reasonable choices of Δt. Thus, the estimate (23) is true for $m=0$ and $l\geq 0$. We now assume that the statement is valid for m replaced by $m'<m$ and all $l\geq 0$. We need to show that it is also valid for m. We begin by showing that $D_+^m u^n$ is bounded. By induction we then show the statement for all l. We have

$$D_+\left(D_+^m u^n\right) = D_+^m\left(D_+ u^n\right) = D_+^m\left(AQE_+u^n\right)+D_+^m F^{n+1/2}.$$

For the difference quotient of a matrix vector product $A(t_n)u^n$, we have

$$D_+^m\left(A(t_n)u^n\right) = \sum_{j=0}^{m}\binom{m}{j}\left(E_+^{m-j}D_+^j A(t_n)\right)\left(D_+^{m-j}E_+^j u^n\right). \tag{25}$$

This can easily be shown to be true. Thus,

$$\begin{aligned}D_+\left(D_+^m u^n\right) &= D_+^m\left(AQE_+u^n\right)+D_+^m F^{n+1/2}\\ &= \left(E_+^m AQ\right)E_+D_+^m u^n+\textstyle\sum_{j=1}^{m}\binom{m}{j}\left(D_+^j E_+^{j-m}A\right)E_+^{j+1}QD_+^{m-j}u^n\\ &\quad+D_+^m F^{n+1/2}.\end{aligned} \tag{26}$$

By the induction assumption, we know that

$$\max_{1\leq j\leq m}\left\|QD_+^{m-j}u^n\right\| \leq K(t_n)\left(\max_{\substack{m'+l'\leq m\\ m'\leq m-j}}\left(\left\|Q^{l'}D_+^{m'}u^0\right\|+\max_{n'}\left\|Q^{l'}D_+^{m'}F^{n'+1/2}\right\|\right)\right).$$

We multiply (26) by $2\Delta t E_+\left(D_+^m u^n\right)$ and use the semiboundedness of $E_+^m AQ$, which follows directly from the semiboundedness of AQ for all t. We obtain,

$$\left\|D_+^m u^n\right\| \leq K(t_n)\left(\max_{m'+l'\leq m}\left(\left\|Q^{l'}D_+^{m'}u^0\right\|+\max_{n'}\left\|Q^{l'}D_+^{m'}F^{n'+1/2}\right\|\right)\right).$$

Here we have used the boundedness of the derivatives of A. Thus, statement (23) is true for $l = 0$ and general m. For the induction step to prove the statement for general l, we have

$$\begin{aligned} D_+Q^lD_+^mu^n &= Q^lD_+^m\left(AQE_+u^n\right) + Q^lD_+^mF^{n+1/2} \\ &= Q^lE_+^mAQE_+D_+^mu^n + \textstyle\sum_{j=1}^m \binom{m}{j} Q^l\left(D_+^jE_+^{j-m}A\right)E_+^{j+1}QD_+^{m-j}u^n \\ &\quad + Q^lD_+^mF^{n+1/2}, \end{aligned}$$

where we have used (25). We now use Lemma 5.1 to obtain

$$\begin{aligned} D_+Q^lD_+^mu^n &= E_+^mAQE_+Q^lD_+^mu^n + l[Q,E_+^mA]E_+Q^lD_+^mu^n \\ &\quad + \textstyle\sum_{j=2}^l \binom{l}{j}[Q,E_+^mA]_jE_+Q^{l+1-j}D_+^mu^n \\ &\quad + \textstyle\sum_{j=1}^m \binom{m}{j}\left(D_+^jE_+^{j-m}A\right)E_{j+1}QD_+^{m-j}Q^lu^n \\ &\quad + l[Q,E_+^{m-j}D_+^jA]E_+^{j+1}Q^lD_+^{m-j}u^n \\ &\quad + \textstyle\sum_{i=2}^l \binom{l}{i}[Q,E_+^{m-j}D_+^jA]_jQ^{l+1-j}E_+^{j+1}D_+^{m-j}u^n + Q^lD_+^mF^{n+1/2}. \end{aligned}$$

The third term is bounded by induction over l and the fourth, fifth and sixth terms are bounded by induction over m. Thus, when multiplying by $E_+Q^lD_+^mu^n$, one arrives at the theorem. □

Note: The assumptions for the commutators $[Q,A]_j$ are usually valid for operators $A(t)Q$ arising from the spatial discretizations of PDEs, aside from possible complications due to boundary conditions. The matrix elements of $[Q,A]$ often are approximations to derivatives of the original coefficients of the PDE, and $[Q,A]_j$ will correspond to high order derivatives. If these derivatives are bounded, the commutators are bounded independent of Δx. This is quite easy to see for first order differential operators. For higher order differential operators, an additional step needs to be taken in the above proof.

When investigating the proofs for the estimates in Theorems 3.2 and 4.2, we see that we can now prove similar estimates for $A = A(t)$, under the assumptions stated in Theorem 5.1.

6 Special treatment of the boundary conditions

In the previous sections, we have found that the estimates for $\left\|e^{2j,n}\right\|$ involve terms of the form $A^ku^{(l)}(0)$ and for the IMR case also $A^ku^{(l)}(t)$. In order for the estimates to be useful for PDE discretizations, we need them to be independent of the space-step Δx. Usually, A is a discretization of a spatial differential operator involving boundary conditions for the underlying PDE. If we assume smooth solutions in space and no boundary conditions are present, we can expect the above terms to be bounded. One has to be careful in the presence of boundary conditions, however. When formulating the boundary

conditions in the usual way, one finds that $A^l u$ in general has a bound that deteriorates for decreasing Δx. As an example, consider the PDE

$$\partial_t u = -\partial_x u + f$$

with the boundary condition $u(0,t) = g(t)$. A spatial discretization of the PDE is

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_N \end{pmatrix}_t = \begin{pmatrix} 0 & -\frac{1}{2\Delta x} & 0 & \cdots & \cdots \\ \frac{1}{2\Delta x} & 0 & -\frac{1}{2\Delta x} & 0 & \cdots \\ 0 & \frac{1}{2\Delta x} & 0 & -\frac{1}{2\Delta x} & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ \cdots & \cdots & 0 & \frac{1}{\Delta x} & -\frac{1}{\Delta x} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_N \end{pmatrix} + \begin{pmatrix} f_1 + \frac{g}{2\Delta x} \\ f_2 \\ f_3 \\ \vdots \\ f_N \end{pmatrix} .$$

By calculating $A^k u$, one sees that the first element is proportional to $\frac{1}{\Delta x^k}$. This leads to unusable bounds in the estimates for $\Delta x \to 0$. We now present two techniques for modifying the boundary conditions in order to obtain bounded $A^k u$.

6.1 Modified boundary conditions: MBC1

The first technique, which we refer to by MBC1, is based on taking a time derivative of the original boundary condition. Instead of using the original boundary condition, we prescribe the condition

$$\partial_t^b u(0,t) = g^{(b)}(t) ,$$

and introduce new variables

$$v_i(t) = \partial_t^i u(0,t) \quad i = 0, \ldots, b-1 .$$

We add the following rows to the system of ODEs

$$\begin{aligned} v_i'(t) &= v_{i+1}(t) \quad i = 0, \ldots, b-2 , \\ v_{b-1}'(t) &= g^{(b)}(t) . \end{aligned} \tag{27}$$

Note that a similar modification of the boundary conditions has been done in [1] for high order Runge-Kutta methods. In the above example, we obtain the following system for $b = 2$.

$$\begin{pmatrix} v_1 \\ v_0 \\ u_1 \\ \vdots \\ u_N \end{pmatrix}_t = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & \frac{1}{2\Delta x} & 0 & -\frac{1}{2\Delta x} & 0 & \cdots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \cdots & \cdots & \cdots & 0 & \frac{1}{\Delta x} & -\frac{1}{\Delta x} \end{pmatrix} \begin{pmatrix} v_1 \\ v_0 \\ u_1 \\ \vdots \\ u_N \end{pmatrix} + \begin{pmatrix} g'' \\ 0 \\ f_1 \\ \vdots \\ f_N \end{pmatrix} .$$

Here

$$A\begin{pmatrix} v_1 \\ v_0 \\ u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} = \begin{pmatrix} 0 \\ v_1 = u_t(0,t) = u_{0x} + f_0 \\ u_{1x} + \mathcal{O}\left(\Delta x^2\right) \\ u_{2x} + \mathcal{O}\left(\Delta x^2\right) \\ \vdots \\ u_{Nx} + \mathcal{O}(\Delta x) \end{pmatrix},$$

and

$$A^2\begin{pmatrix} v_1 \\ v_0 \\ u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ u_{1xx} + \frac{f_0}{2\Delta x} + \mathcal{O}\left(\Delta x^2\right) \\ u_{2xx} + \mathcal{O}\left(\Delta x^2\right) \\ \vdots \\ u_{Nxx} + \mathcal{O}(\Delta x) \end{pmatrix}.$$

As indicated by this example we can see that for problems with time-independent A and where the spatial derivatives of the forcing function are zero at the boundary, we obtain $\left\|A^k u\right\| \sim \frac{1}{\Delta x^{k-b}}$. Thus, for sufficiently high values of b, the modification will render bounded $A^k u$.

6.2 Modified boundary conditions: MBC2

The second technique, which we refer to as MBC2, works for general forcing functions, but again for time-independent A. Assume that the ODE

$$u_t(t) = Au(t) + F(t)$$

arises from the PDE

$$\begin{aligned} u_t(x,t) &= Lu(x,t) + f(x,t)\,, \\ u(0,t) &= g(t)\,, \end{aligned}$$

where L is a linear differential operator that may depend on x. We assume that it does not depend on t. We now introduce new variables

$$v_i(t) = L^i u(0,t) \quad i = 0, \ldots, b-1\,,$$

and add the following equations to the system

$$\begin{aligned} v_i'(t) &= v_{i+1}(t) + L^i f(0,t) \quad i = 0, \ldots, b-2\,, \\ v_{b-1}'(t) &= g^{(b)}(t) - \textstyle\sum_{j=0}^{b-1} L^j \partial^{b-1-j} f(0,t)\,. \end{aligned} \tag{28}$$

With this, $A^k u$ is bounded if b is chosen sufficiently large. Thus, we have found a way to modify the boundary conditions for problems with time-independent coefficients, in order to obtain estimates that are independent of Δx as $\Delta x \to 0$.

7 Numerical experiments

In the previous sections, we have developed estimates for the errors of the deferred correction scheme. Smoothness requirements for correct order of accuracy have been given. We have also given ways of choosing the initial conditions in the BDF2 based scheme and the boundary conditions in order for these requirements to hold. In this section, we conduct a number of numerical experiments to support the results of the previous sections. In Section 7.1, we investigate the effect of the choice of the second initial condition for the BDF2 based deferred correction. In Section 7.2, we demonstrate the performance of the scheme for a problem with time-dependent coefficients. In Section 7.3, we consider the suggested methods MBC1 and MBC2 for modifying the boundary conditions. This is done for a hyperbolic equation. A similar experiment, not shown here, has been performed for a parabolic problem, with similar results. Actually, the diffusiveness of the parabolic equation reduces the error coming from the boundary conditions since oscillations are damped out in time. Thus, in practice, the complications from the boundary are not as serious as in the hyperbolic case. We also show some test cases with nonzero forcing functions, both for time-independent and time-dependent coefficients. In addition to the problems shown here, we have considered problems with non-smooth solutions and a short discussion follows at the end of this section.

7.1 Effect of the choice of the initial conditions for the BDF2 based scheme

We consider a problem with periodic boundary conditions to illustrate the effect of the choice of the second initial condition for the BDF2 based deferred correction scheme. Consider the following system,

$$\begin{aligned} u_t &= v_x + u_{xx} + F_1\,, \qquad & v_t &= u_x + v_{xx} + F_2\,, \\ u\,(x,0) &= 2\cos(x)\,, & v\,(x,0) &= 0\,, \end{aligned} \tag{29}$$

for $0 \le t \le 2\pi$ and with periodic solutions

$$\begin{pmatrix} u \\ v \end{pmatrix}(x+2\pi,t) = \begin{pmatrix} u \\ v \end{pmatrix}(x,t)\ .$$

We choose F_1 and F_2 such that the exact solution is given by

$$\begin{aligned} u\,(x,t) &= \cos(x+t) + \cos(x-t)\,, \\ v\,(x,t) &= \cos(x+t) - \cos(x-t)\,. \end{aligned} \tag{30}$$

For the spatial discretization, we use the following sixth order Padé approximation on a staggered grid, i.e., we store the u and v solution at alternating points, $u_j^n = u\,(j\Delta x, n\Delta t)$, $v_j^n = u\left(\left(j-\frac{1}{2}\right)\Delta x, n\Delta t\right)$ for $j = 1,\ldots,N$, $\Delta x = 2\pi/N$,

$n = 1, \ldots, M$, $\Delta t = 2\pi/M$ and use the approximation

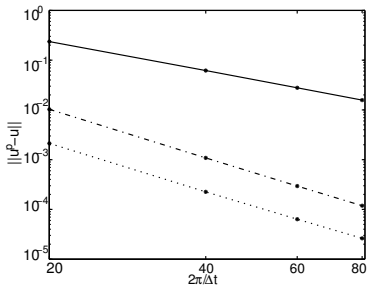
$$\frac{9(u_x)_{j-1/2}+62(u_x)_{j+1/2}+9(u_x)_{j+3/2}}{80} = \frac{17u_{j+1}+189u_j-189u_{j-1}-17u_{j-2}}{240\Delta x},$$
$$\frac{2(u_{xx})_{j-1}+11(u_{xx})_j+2(u_{xx})_{j+1}}{11} = \frac{3(u_{j+2}+16u_{j+1}-34u_j+16u_{j-1}+u_{j-2})}{44\Delta x^2}.$$

For more details on compact discretizations on regular and staggered grids, see Fornberg [7] or Lele [19]. The periodicity is taken into account by replacing any $j \leq 0$ by $j \to j + N$ and $j > N$ by $j \to j - N$. We use small spatial steps to isolate the time error in order to illustrate the order of accuracy for the time discretization. The time discretization is done with the deferred correction scheme using the BDF2 scheme as the base scheme. We consider the second, fourth and sixth order deferred correction schemes for decreasing Δt and Δx, keeping $\Delta x/\Delta t$ constant. We investigate the order of convergence and the error. In all of the numerical experiments, we show the error in the l^2-norm,

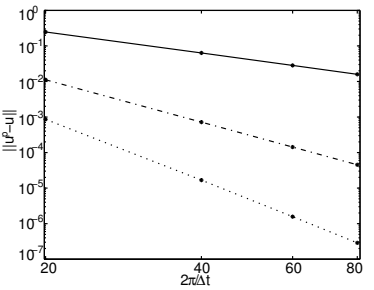
$$\|u^p - u\|^2 = \sum_{n=1}^{M}\sum_{j=1}^{N} \frac{1}{\Delta t \Delta x} |u_j^{p,n} - u\,(x_j, t_n)\,|^2 .$$

and give the order of accuracy o_p, determined by the error on the finest and coarsest grids.

In Figure 1(a), $u^{2,1}$, $u^{4,1}$ and $u^{6,1}$ are chosen to be the exact solution at $t = \Delta t$. In Figure 1(b), the modified starting values according to (11) and (12) are used. One can clearly see that the use of the modified starting values results in a



(a) exact starting values, order of accuracy: $o_2 = 2.0$, $o_4 = 3.2$, $o_6 = 3.2$.



(b) modified starting values, order of accuracy: $o_2 = 2.0$, $o_4 = 4.0$, $o_6 = 5.8$.

Figure 1: l^2-error for problem (29) using BDF2 based deferred correction, $\Delta x = \Delta t/2$, (— 2^{nd} order, $- \cdot -$ 4^{th} order, $\cdots$ 6^{th} order).

smaller error and better convergence order, although for large time-steps both methods result in similar errors. However, for a time-step $\Delta t = 2\pi/80$, the error for the sixth order deferred correction using the modified starting values is around 1% of that using the exact solution as a second initial data.

7.2 Time-dependent coefficients

We now illustrate the validity of the error estimate even for time-dependent coefficients. We use the IMR based deferred correction for the following problem.

$$\begin{aligned} u_t = a(x,t)\, v_x + b(x,t)\, u_{xx} + F_1\,, \qquad & v_t = a(x,t)\, u_x + b(x,t)\, v_{xx} + F_2\,, \\ u(x,0) = 2\cos(x)\,, \qquad & v(x,0) = 0\,, \end{aligned} \tag{31}$$

for $0 \le t \le 2\pi$ with periodic solutions $u(x+2\pi,t) = u(x,t)$ and $v(x+2\pi,t) = v(x,t)$ and

$$\begin{aligned} a(x,t) &= \sin(x+t)\,, \\ b(x,t) &= 2 + \sin(x+t)\,. \end{aligned}$$

Again, we choose F such that the exact solution is given by (30). For the spatial discretization, we use the same scheme as in the first experiment. In Figure 2, we can see that the deferred correction scheme gives good results for this problem. The biggest improvement is obtained from second to fourth order accuracy.

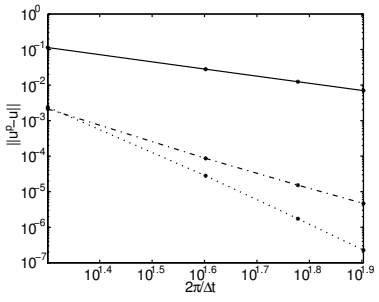


Figure 2: l^2-error for problem (31), time-dependent coefficients, IMR based scheme, $\Delta x = \Delta t/2$, (— 2^{nd} order, $-\cdot-$ 4^{th} order, $\cdots$ 6^{th} order), order of accuracy: $o_2 = 2.0$, $o_4 = 4.4$, $o_6 = 6.7$.

7.3 Modified boundary conditions

In the next example, we focus on the need for modified boundary conditions for both the IMR approach and the BDF2 approach. We first consider a problem without forcing terms.

$$\begin{aligned} u_t + u_x &= 0\,, \quad 0 \le x \le 2\pi\,, \quad 0 \le t \le 2\pi\,, \\ u(0,t) &= \cos(t)\,, \quad u(x,0) = \cos(x)\,. \end{aligned} \tag{32}$$

The exact solution is given by $u(x,t) = \cos(x-t)$.

For this and the remaining examples, we use the following sixth order Padé approximation for the spatial discretization,

$$\frac{(u_x)_{j-1} + 3\,(u_x)_j + (u_x)_{j+1}}{3} = \frac{-u_{j-2} - 28u_{j-1} + 28u_{j+1} + u_{j+2}}{36\Delta x}\,. \tag{33}$$

We treat the *numerical* boundary conditions by introducing ghost points and defining the values at the ghost points by high order extrapolation. We modify the *physical* boundary conditions according to Section 6. For the time discretization, we use both deferred correction versions discussed above, based on the implicit midpoint rule and the BDF2 scheme. For the modification of the boundary conditions, both modification techniques, MBC1 and MBC2, yield the same results in this case, since no forcing term is present. We compare the results for different values of b. The results are shown in Figures 3 and 4 and Tables 1 and 2. We have chosen values of b between 0 and 6. In the three

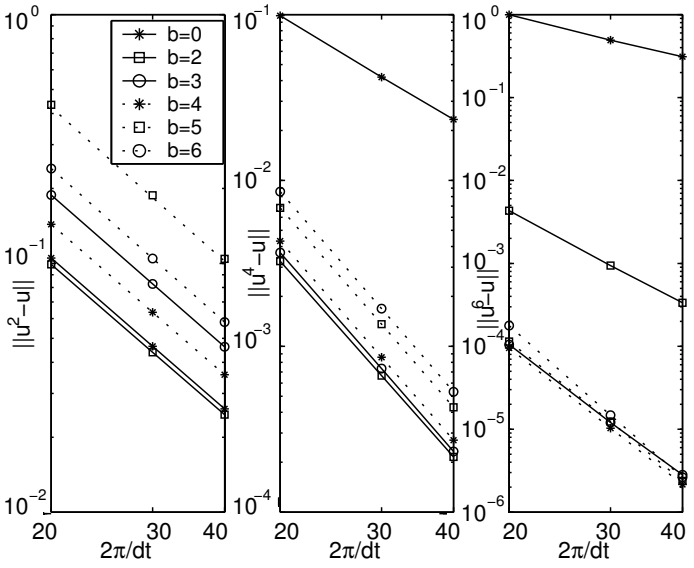


Figure 3: l^2-error for problem (32), IMR based deferred correction.

b	o_2	o_4	o_6
0	2.0	2.1	1.7
2	2.0	3.9	3.7
3	2.0	4.0	5.2
4	2.0	4.0	5.5
5	2.1	4.0	5.6
6	2.1	4.0	6.1

Table 1: Convergence order for problem (32), IMR based deferred correction.

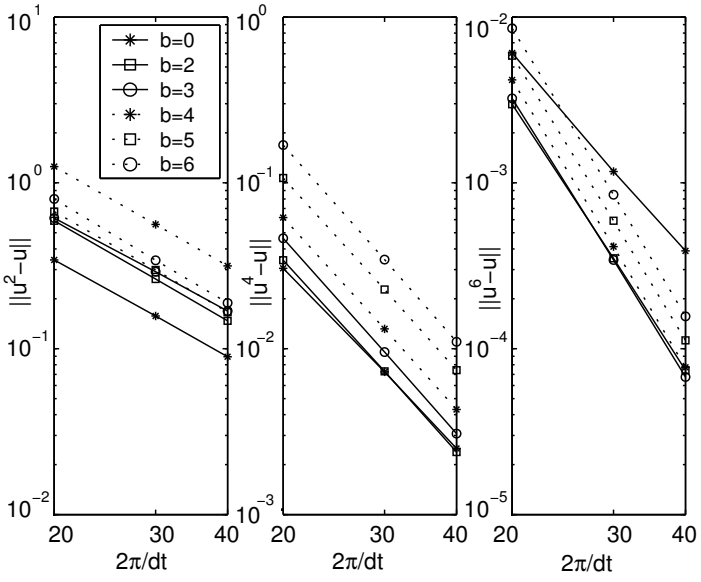


Figure 4: l^2-error for problem (32), BDF2 based deferred correction.

b	o_2	o_4	o_6
0	1.9	3.6	4.0
2	2.0	3.8	5.3
3	1.9	3.9	5.6
4	2.0	3.8	5.8
5	2.0	3.9	5.7
6	2.1	3.9	5.8

Table 2: Convergence order for problem (32), BDF2 based deferred correction.

plots in each figure, the l^2-error is depicted for the second, fourth and sixth order deferred correction scheme. Without the modified boundary conditions, the error for the sixth order IMR deferred correction scheme actually increases compared to the original IMR. One can see an improvement in both the order of accuracy and the accuracy itself for suitable values of b. For the BDF scheme, the improvement is noticeable in the sixth order deferred correction, where $b = 2$ gives the best results. When comparing the experimental results to the error estimates in the previous sections, one obtains better results than expected. When assuming that the initial errors are zero, Theorem 4.2 yields the following estimate,

$$\|e^{4,n}\| \leq \Delta t^4 (\max_t \|A^2 u^{(6)}(t)\| + \max_{0\leq s\leq t_\nu} \|A^3 u^{(2)}(s)\|) + \mathcal{O}(\Delta t^4)\,, \tag{34}$$

and

$$\|e^{6,n}\| \leq \Delta t^6 (\max_t \|A^3 u^{(9)}(t)\| + \max_{0\leq s\leq t_\nu} \|A^6 u^{(2)}(s)\|) + \mathcal{O}(\Delta t^6)\,. \tag{35}$$

We have included the highest powers of A only. Here, t_ν is small and independent of n. For the BDF2 based scheme, the corresponding estimates are

$$\left\|e^{4,n}\right\| \leq const\Delta t^4 \left(\left\|A^2 u^{(3)}\left(0\right)\right\| + \max_t \left\|u^{(6)}\left(t\right)\right\|\right) + \mathcal{O}\left(\Delta t^4\right)\,,$$

and

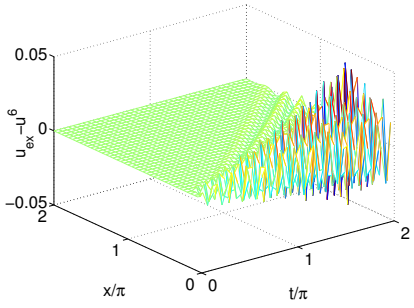
$$\left\|e^{6,n}\right\| \leq const\Delta t^6 \left(\left\|A^5 u^{(3)}\left(0\right)\right\| + \max_t \left\|u^{(9)}\left(t\right)\right\|\right) + \mathcal{O}\left(\Delta t^6\right)\,.$$

The estimates (34) and (35) contain two terms. The first term is bounded when taking $b = 2$ in the fourth order case and $b = 3$ in the sixth order case. In the experiments for the IMR based scheme, significant improvements in both the size of the error and the order of accuracy are observed for exactly these values. The last term in (34) and (35) suggests the need for a higher value of b. However, in our experiments, the improvement for higher values of b is not very significant. For the BDF based scheme, there are no terms with A applied to the solution $u(t)$, but only terms with A applied to the initial data. Thus, the problem at the boundary is not as severe. However, we still see an improvement for the sixth order scheme when modifying the boundary condition. In Figures 5(a) – 5(d), we show the error for both sixth order deferred correction schemes for $b = 0$ and $b = 3$ respectively. One can see the different behavior of the error for the BDF based scheme and the IMR based scheme. Consider first, the IMR based scheme. In Figures 5(a) and 5(b), one can see oscillations at the boundary due to the first term in (35), i.e., A acting on the exact solution at time t. By taking $b = 3$, the oscillations due to this term have been eliminated, but small oscillations due to the last term, A acting on the initial data, are still present. For the BDF based scheme, the error of the deferred correction scheme only depends on A acting on the initial data. In accordance to this, we see in Figures 5(c) and 5(d) that all oscillations in the error originate at $t = 0$, $x = 0$ and are quickly damped out for positive t. For the case $b = 3$, all oscillations are eliminated. This is actually a lower value than expected.

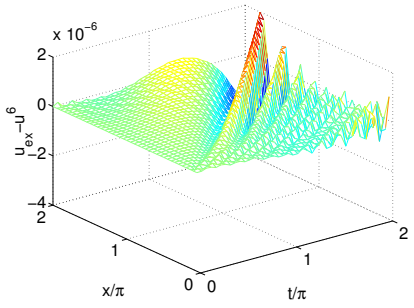
We now consider the following scalar equation with a forcing term. Consider

$$\begin{aligned} u_t + \tfrac{1}{2} u_x &= \tfrac{1}{2}\sin(x-t)\,, \quad 0 \leq x \leq 2\pi\,, \quad 0 \leq t \leq 2\pi\,, \\ u(0,t) &= \cos(t)\,, \quad u(x,0) = \cos(x)\,. \end{aligned} \tag{36}$$

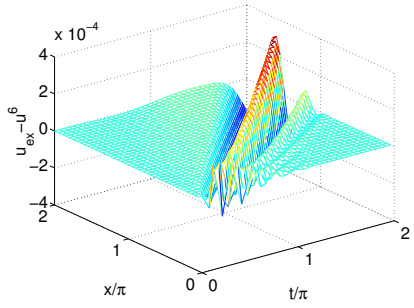
The exact solution is again given by $u(x,t) = \cos(x-t)$. We use the same spatial discretization as above. We now use $b = 4$. This time, we compare the results for the deferred correction method using the implicit midpoint rule and the BDF2 scheme and also the two different boundary conditions, MBC1 and MBC2. The results are shown in Figures 6(a) – 6(d) and Table 3. As expected, MBC2 works better in both cases. We note that the IMR based scheme gives a lower error than the BDF2 based scheme. However, the BDF scheme is not as



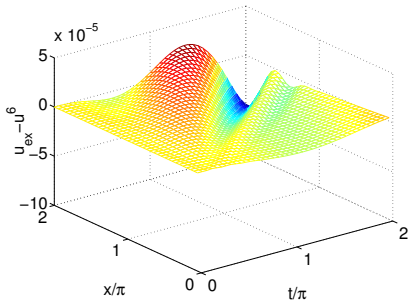
(a) 6^{th} order IMR based scheme, $b = 0$.



(b) 6^{th} order IMR based scheme, $b = 3$.



(c) 6^{th} order BDF based scheme, $b = 0$.



(d) 6^{th} order BDF based scheme, $b = 3$.

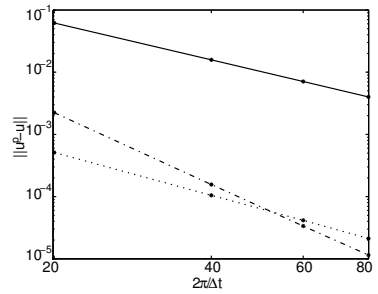
Figure 5: Error of the sixth order deferred correction scheme for problem (32), $\Delta t = 2\pi/40$.

susceptible as the IMR scheme to the presence of a forcing function when using MBC1.

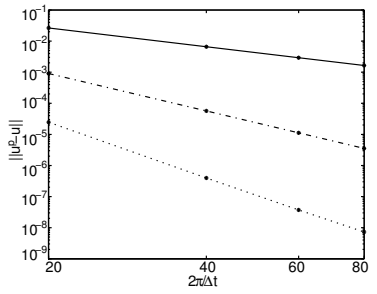
Finally, we present an example with time and space varying coefficients. Here, for convenience, we have only used MBC1 as a method to modify the boundary conditions.

$$\begin{aligned} u_t + \tfrac{x+1}{t+1} u_x &= \sin(x-t)\left(\tfrac{x+1}{t+1} - 1\right), \quad 0 \le x \le 2\pi, \quad 0 \le t \le 2\pi, \\ u(0,t) &= \cos(t), \quad u(x,0) = \cos(x). \end{aligned} \tag{37}$$

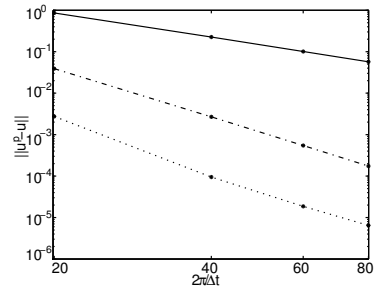
The exact solution is again given by $u(x,t) = \cos(x-t)$. The l^2-errors for $b = 4$ are shown in Figures 7(a) and 7(b). Again, the order of accuracy of the BDF2 based scheme is closer to the optimal value. We also see that the IMR based scheme, although more sensitive to the boundary conditions, tends to have smaller errors.



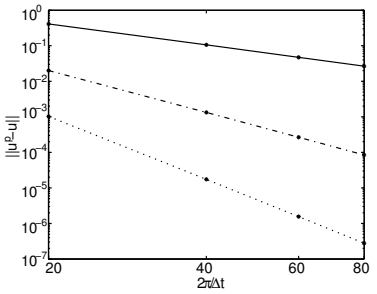
(a) IMR based scheme, MBC1.



(b) IMR based scheme, MBC2.



(c) BDF based scheme, MBC1.



(d) BDF based scheme, MBC2.

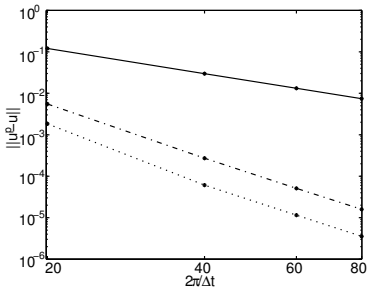
Figure 6: l^2-error for problem (36), $\Delta x = \Delta t/2$, (— 2^{nd} order, $-\cdot-$ 4^{th} order, $\cdots$ 6^{th} order).

7.4 Non-smooth problems

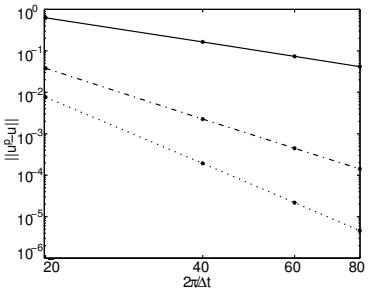
We have also investigated the performance of the IMR based deferred correction scheme for problems with non-smooth solutions. When considering problems that are discontinuous in time only, the deferred correction scheme does not improve the accuracy compared to the base scheme. The error for all deferred correction steps remains approximately the same. When considering a PDE with a non-smooth solution both in space and time, we do observe destruction of accuracy when attempting to obtain higher order deferred correction approximations. This is as we would expect. It becomes clear in our estimates that the performance of the scheme is highly sensitive to the action of A on the exact solution. In this case, $A^j u$ will grow with decreasing Δx. We expect the behavior to be somewhat better for the BDF2 based scheme, as the dependence on A is not as strong as in the IMR case and initial oscillations are damped out due to the diffusivity of the BDF2 scheme.

	o_2	o_4	o_6
IMR, MBC1	2.0	3.8	2.3
IMR, MBC2	2.0	4.0	5.9
BDF, MBC1	2.0	3.9	4.4
BDF, MBC2	2.0	3.9	5.9

Table 3: Convergence order for problem (36).



(a) IMR based scheme, order of accuracy: $o_2 = 2.01$, $o_4 = 4.22$, $o_6 = 4.51$.



(b) BDF2 based scheme, order of accuracy $o_2 = 1.96$, $o_4 = 4.03$, $o_6 = 5.35$.

Figure 7: l^2-error for problem (37), $\Delta x = \Delta t/4$, (— 2^{nd} order, $- \cdot -$ 4^{th} order, $\cdots$ 6^{th} order).

8 Stiff problems

The deferred correction methods discussed in this paper are based on A-stable second order implicit methods. It is often argued that since each time-step for implicit methods is much more expensive to compute, implicit methods are justified only in cases where explicit methods require very small time-steps due to stability restrictions. In this section, we discuss two typical problems, where explicit methods usually fail because of their limited stability domain. One large class of problems which we will not discuss further here, is the class of parabolic and higher order PDE. A restriction on the time-step for methods that are not A-stable is usually proportional to high powers of Δx, leading to quite severe restrictions on Δt. Another class of problems are stiff differential equations, where the time-step is severely restricted only due to stability requirements and not necessarily by the accuracy requirements. Here, implicit methods with large stability domains are necessary. Still another case, where explicit methods usually require excessively small time-steps, is found in problems with a nonuniform spatial grid due to the geometry of the problem, rather than accuracy needs (e.g., corners in domains, the origin of a ball in spherical coordinates). Typically, the restriction of the time-step in order to obtain a stable scheme depends on the smallest space-step used which in some cases is

extremely small.

In this section, we investigate the performance of the deferred correction methods based on the implicit midpoint rule for two of the types of problems described above and discuss the expected performance of the deferred correction method. In Section 8.3, we present numerical experiments, verifying the discussion in this section.

8.1 Problems with spatial grid refinement due to geometry

We begin by discussing the case of a nonuniform grid which is refined due to geometrical restraints rather than accuracy requirements. The permissible time-step for an explicit method is governed by the smallest space-step used. This can be a very severe restriction. The example we look at to illustrate this problem is

$$u_t + u_x = g(x,t) \ ,$$

with a uniformly smooth solution $u(x,t)$. The discretization in space is calculated on a stretched grid. Another way of describing this is by instead considering a modified problem on a uniform grid,

$$v_t + \frac{1}{f'(y)} v_y = g(f(y),t) \ .$$

Here $f(y)$ is a function transforming a uniform grid y_i into a stretched grid $x_i = f(y_i)$. The solution to this problem is $v(y,t) = u(f(y),t)$. To imitate the presence of very small step sizes, we consider a function with $f'(x) << 1$ for some x. We will show that the bounds for the error in time are independent of the function f. To obtain a bound for the error, one can look at the estimate in Theorem 4.2. The error is bounded by terms of the form $A^p\tilde{V}$, where the components $\tilde{V}_j = \tilde{v}(y_j,t)$ consist of time derivatives of the exact solution $V_j(t) = v(y_j,t)$. In the above example, we can assume that $\tilde{v}(y,t) = \tilde{u}(f(y),t)$ where $\tilde{u}$ is a smooth function. We have

$$\begin{aligned} \left(A\tilde{V}\right)_j &= \tfrac{1}{f'(y_j)} \left(Q\tilde{V}\right)_j \\ &\approx \tfrac{1}{f'(y_j)} \partial_y \tilde{v}(y_j,t) = \tfrac{1}{f'(y_j)} \partial_y \tilde{u}(f(y_j),t) \\ &= \partial_x \tilde{u}(f(y_j),t) \ , \end{aligned}$$

where Q is the discretization of ∂_y. In general,

$$\left(A^p\tilde{V}\right)_j \approx \partial_x^p \tilde{U}(f(y_j),t) \ .$$

Hence, the terms $A^p\tilde{V}$ are bounded, independent of $f(y)$.

8.2 Multi timescale problems

Another set of problems that requires implicit methods is the case of problems where two timescales are present in the differential equation, whereas in the solution, only the slow scale is present. Explicit methods require a fine grid in time, corresponding to the fastest timescale, although the solution already is resolved on a coarser grid. For this, we consider the equation

$$u_t + au_x = f(x-t)\,, \quad a >> 1\,, \tag{38}$$

with periodic solutions in space and initial data $u_0(x)$. The forcing term f will be chosen to completely eliminate fast timescales from the problem. To investigate the timescales involved in (38), we perform a Fourier transformation in space, to obtain

$$\hat{u}_t(\omega,t) + ai\omega\hat{u}(\omega,t) = \hat{f}(\omega)\, e^{-i\omega t}\,.$$

One can easily calculate the solution to this problem as

$$\hat{u}(\omega,t) = e^{-i\omega at}\left(\hat{u}(\omega,0) + \frac{1}{i\omega(1-a)}\hat{f}(\omega)\right) - \frac{1}{i\omega(1-a)}e^{-i\omega t}\hat{f}(\omega)\,.$$

We can observe the presence of two timescales in the solution, $e^{-i\omega at}$ and $e^{-i\omega t}$. We consider the case where $\hat{u}(0) = -\frac{1}{i\omega(1-a)}\hat{f}$, i.e., only the slow scale is present in the solution. In order to resolve the solution, one can then use a coarse grid in time.

We now consider the approximate solution from the implicit midpoint rule. This is of interest, since it appears as a forcing function in the difference equation for the fourth order deferred correction. We have

$$D_+u^{2,n} = -aQE_+u^{2,n} + F^{n+\frac{1}{2}}\,, \tag{39}$$

where Q is a discrete approximation to ∂_x and $F_j^{n+\frac{1}{2}} = f\left(x_j - t_{n+\frac{1}{2}}\right)$.

In Appendix B, we perform a discrete Fourier transformation in space. One can see that the approximate solution has two different timescales. The fast scale is not completely eliminated. This means that for large a, part of the error is highly oscillatory in time. For the deferred correction process, this has the consequence that $D_+^3e^{2,n}$, which is present as a forcing term in the equation for $e^{4,n}$, is proportional to a^3. So, even though the error of the fourth order deferred correction method asymptotically still has the correct order of accuracy, the error constant can be very large for large a. For the higher order deferred correction steps of order $2j$, we expect the error to be proportional to $a^{3(j-1)}$. Another way to see this, is to look at the estimates in Theorem 4.2. The error $\left\|e^{2j,n}\right\|$ is bounded by $\left\|A^{3(j-1)}u^{(2)}(s)\right\|$. In our case, $\left\|A^{3(j-1)}u^{(2)}(s)\right\|$ is proportional to $a^{3(j-1)}$. A possible remedy to this might be postprocessing the intermediate solutions or using a special choice of initial conditions to eliminate the fast timescales. The latter has been discussed in [16].

We can conclude that we expect very good behavior for stiff problems arising from geometric constraints. However, for stiff problems with different timescales a suitable time-step for the deferred correction scheme will depend on the fastest timescale in the problem.

8.3 Numerical experiments

In order to validate the theoretical discussion in this section, we perform two numerical experiments. We first consider a problem with a highly irregular grid with partially small step sizes in space. We consider the equation

$$\begin{aligned} u_t + u_x &= 0\,, \\ u\,(x,0) &= \cos\,(x)\ , \end{aligned} \tag{40}$$

with periodic solutions in space. The exact solution is $u\,(x,t) = \cos\,(x-t)$. We want to investigate the performance of the deferred correction method with a nonuniform spatial grid $x_j = f\left(\frac{2\pi j}{N}\right)$. For $f\,(y)$, we choose $f\,(y) = y + \sin\,(y)$. We perform a transformation $y \to f\,(y)$ and solve

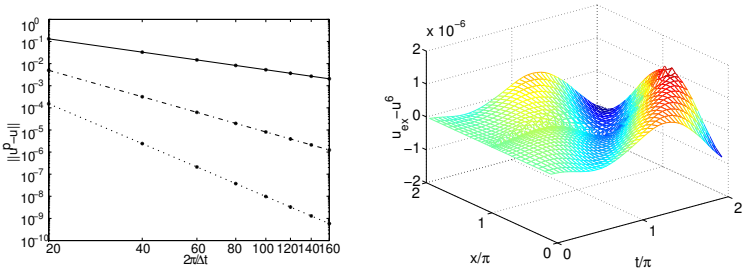
$$v_t + \frac{1}{f'\,(y)} v_y = 0\,, \tag{41}$$

which has the solutions $v\,(y,t) = \cos\,(f\,(y) - t)$. This is now discretized on a uniform grid $y_j = \frac{2\pi j}{N}$. Note that $f'\,(\pi) = 0$. To avoid problems in the discretization, one has to make sure that π is not a grid point. The results of the computations for the deferred correction based on the implicit midpoint rule are presented in Figure 8(a), where the l^2-error is shown for the second, fourth and sixth order deferred correction scheme. In Figure 8(b), the error for the sixth order deferred correction method is shown. Although the grid is very fine around $x = \pi$, the time-step can be chosen quite large, without any problems.

In order to illustrate the problem of stiffness, we consider

$$\begin{aligned} u_t + a u_x &= (1-a)\sin\,(x-t)\ , \\ u\,(x,0) &= \cos\,(x)\ , \end{aligned} \tag{42}$$

with a periodic solution $u\,(x+2\pi,t) = u\,(x,t)$. The solution to this problem is $u\,(x,t) = \cos\,(x-t)$, independent of a. The results of the deferred correction based on IMR for $a = 10$ are shown in Figure 9(a), where the l^2-error is shown for the second, fourth and sixth order methods. One can see that the performance of higher order deferred correction methods is not very good for large time-steps, although the IMR itself performs well. When decreasing the time-steps, eventually, the error for the higher order methods decreases below the error of the IMR method and is approximately of correct order of accuracy. In Figure 9(b), the error of the sixth order deferred correction scheme is shown for $\Delta t = \frac{2\pi}{40}$ and $\Delta x = \frac{2\pi}{80}$. One can see that the temporal behavior is like $\sin\,(10t)$.



(a) (b)

Figure 8: (a) l^2-error for problem (41) using deferred correction based on IMR, $\Delta x = \Delta t/2$, (— 2^{nd} order, $-\cdot-$ 4^{th} order, $\cdots$ 6^{th} order), order of accuracy: $o_2 = 2.00$, $o_4 = 3.99$, $o_6 = 6.00$. (b) Error for the sixth order approx., $\Delta t = 2\pi/40$ for problem (41).

9 Time compact schemes

To conclude, we briefly discuss an alternative approach to obtaining high order time discretization schemes. In the above scheme, wide discretization stencils in time are applied to previously calculated solutions. This may lead to storage problems when dealing with very large systems of equations. We now introduce the concept of time compact schemes which avoid the wide stencils in time.

The time compact approach can be used for both ODEs and PDEs. When solving PDEs it is favorable to return to the continuous problem

$$u_t = Lu + f\,, \tag{43}$$

where we assume that L is a linear spatial differential operator with constant coefficients. The method is applicable to a nonlinear problems with varying coefficients as well.

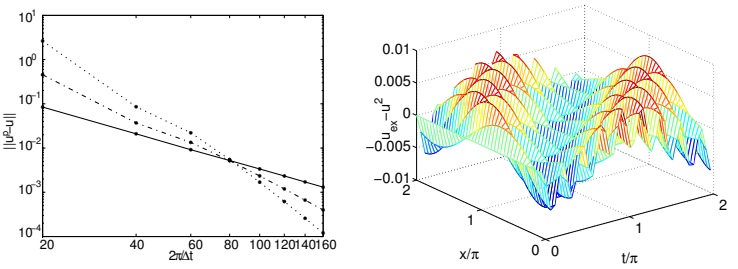
We first discretize (43) in time with the IMR scheme. We obtain the local truncation error

$$e\left(u,t,\Delta t\right) = \frac{\Delta t^2}{24}u_{ttt} - \frac{\Delta t^2}{8}Lu_{tt} + \mathcal{O}\left(\Delta t^4\right)\,.$$

Now, instead of directly discretizing the local truncation error, one can use the original differential equation to transfer time derivatives into spatial derivatives. The local truncation error in time becomes

$$\begin{aligned}\tilde{e}\left(u,\Delta t,t\right) &= \tfrac{\Delta t^2}{24}\left(L^3u + L^2f + Lf_t + f_{tt}\right)\\ &\quad -\tfrac{\Delta t^2}{8}\left(L^3u + L^2f + Lf_t\right) + \mathcal{O}\left(\Delta t^4\right)\,.\end{aligned}$$

There are several possibilities of developing a higher order time discretization from this. One way is to first calculate a second order accurate solution by



(a) (b)

Figure 9: (a) l^2-error for problem (42) using IMR based deferred correction, $a = 10$, $\Delta x = \Delta t/2$, (— 2^{nd} order, $- \cdot -$ 4^{th} order, $\cdots$ 6^{th} order), order of accuracy: $o_2 = 2.00$, $o_4 = 3.39$, $o_6 = 4.81$. (b) Error of the second order approx for problem (42), $a = 10$, $\Delta t = 2\pi/40$.

solving

$$D_+ u^{2,n} = L_h E_+ u^{2,n} + f^{n+1/2}\,,$$

where L_h is a discretization of the spatial operator L. The solution is then used in the same way as in the previously discussed deferred correction procedure, i.e., to obtain a fourth order accurate scheme in time, one solves

$$\begin{aligned} D_+ u^{4,n} &= L_h E_+ u^{4,n} + f^{n+1/2} + \tfrac{\Delta t^2}{24}\left(\left(L^3\right)_h E_+ u^{2,n} + L^2 f + L f_t + f_{tt}\right) \\ &\quad - \tfrac{\Delta t^2}{8}\left(\left(L^3\right)_h u^{2,n} + L^2 f + L f_t\right)\,. \end{aligned}$$

The operator $\left(L^3\right)_h$ is a second order accurate discretization of the spatial operator L^3. Another possibility to obtain a higher order scheme is to develop a direct method. For this, we apply the local truncation error to the unknown approximation, i.e., we solve

$$\begin{aligned} D_+ u^{4,n} &= L_h E_+ u^{4,n} + f^{n+1/2} + \tfrac{\Delta t^2}{24}\left(\left(L^3\right)_h E_+ u^{4,n} + L^2 f + L f_t + f_{tt}\right) \\ &\quad - \tfrac{\Delta t^2}{8}\left(\left(L^3\right)_h u^{4,n} + L^2 f + L f_t\right)\,. \end{aligned}$$

We investigate this class of schemes further in [17], where we consider the wave equation in one and two space dimensions.

Compared to the deferred correction scheme, the time compact scheme has the advantage of only needing to save two time-steps of the lower order solution. One disadvantage is the fact that it can only be applied if the time derivatives can be cancelled using the original equations. This is for example not the case for the incompressible Navier Stokes equation, where no equation for the time derivative of the pressure is present.

A similar class of methods can be developed with the *method of modified equations* which has been considered in several works, e.g., [2], [9], [10] and [14],

mainly as a tool for analyzing existing finite difference schemes. Higher order discretizations in time based on the modified equation have been mostly applied using explicit methods in time for the second order wave equation, [3], [20], [26]. In [27], an extensive study on modified equation methods based on the explicit forward Euler scheme and the implicit midpoint rule for ordinary differential equations have been studied. Different schemes, based on several versions of the modified equation are studied, including a version that is similar to the deferred correction scheme discussed in the previous sections. They have performed a number of numerical studies for linear and nonlinear ODE, showing the validity of the methods. However, especially for the deferred correction method, no rigorous proofs for stability have been made. We would like to point out that, as also seen in the results in this paper, the stability of the deferred correction method is not as straightforward as suggested in [27], especially when considering PDE discretizations.

10 Summary

In this paper, we have investigated several aspects of the deferred correction method in time applied to PDEs. The requirements for optimal order of accuracy have been investigated by deriving estimates for the error in terms of the problem data.

From the estimates and numerical experiments, it could be seen that sufficient smoothness of the data is necessary in order to achieve the desired order of accuracy. For initial boundary value problems, this involves careful investigation of the boundary conditions and for multistep methods as underlying schemes, the numerical initial conditions have to be chosen in a special way, to ensure smoothness of the intermediate solutions. We have presented a method of modifying the boundary conditions so that for time-independent A, the terms in the estimates stay bounded as we decrease the space-step. The theoretical results have been validated by a number of numerical experiments.

We have also investigated the performance of the deferred correction method for two problems that typically require the use of implicit methods with large stability domains. In the example of a stiff scalar PDE with two timescales in the problem, it is seen that the error will in general consist of a fast and a slow timescale, even though the fast scale may be hidden in the exact solution. The higher order schemes are still of the correct order. However, the error constant might be very large, rendering a larger error than for lower order methods. For the case of a nonuniform grid, where parts of the computational grid are extremely fine due to geometric reasons, the deferred correction performs well independent of the nonuniformity of the grid.

A Choice of initial data for the BDF2 based scheme

We present the details of calculations for obtaining the correct choice for $u^{2,1}$ and $u^{4,1}$, see (10) and (11). We have

Theorem A.1. *Let $u^{2,n}$ and $u^{4,n}$ be the solution to the second and fourth order deferred correction scheme based on the BDF2 scheme. If we choose $u^{2,1}$ and $u^{4,1}$ such that*

a)

$$e^{2,1} = -\frac{1}{3}u^{(3)}(0)\Delta t^3 + \frac{1}{12}u^{(4)}(0)\Delta t^4 - \frac{1}{6}Au^{(3)}(0)\Delta t^4$$

then $D_+^3 e^{2,0} = \mathcal{O}\left(\Delta t^2\right)$ and $D_+^3 e^{2,1} = \mathcal{O}\left(\Delta t^2\right)$,

b)

$$\begin{aligned} e^{2,1} &= -\tfrac{1}{3}u^{(3)}(0)\Delta t^3 \\ &\quad +\tfrac{1}{12}u^{(4)}(0)\Delta t^4 - \tfrac{1}{6}Au^{(3)}(0)\Delta t^4 \\ &\quad -\tfrac{19}{120}u^{(5)}(0)\Delta t^5 - \tfrac{1}{24}Au^{(4)}(0)\Delta t^5 - \tfrac{1}{6}A^2u^{(3)}(0)\Delta t^5 \\ &\quad +\tfrac{11}{90}u^{(6)}(0)\Delta t^6 + \tfrac{1}{40}Au^{(5)}(0)\Delta t^6 \\ &\quad +\tfrac{1}{12}A^2u^{(4)}(0)\Delta t^6 - \tfrac{1}{24}A^3u^{(3)}(0)\Delta t^6 \\ &\quad -\tfrac{3}{32}A^3u^{(4)}(0)\Delta t^7 - \tfrac{109}{1440}Au^{(6)}(0)\Delta t^7 - \tfrac{389}{3360}u^{(7)}(0)\Delta t^7 \\ &\quad -\tfrac{73}{480}A^2u^{(5)}(0)\Delta t^7 - \tfrac{1}{8}A^4u^{(3)}(0)\Delta t^7 . \end{aligned}$$

$$\begin{aligned} e^{4,1} &= \tfrac{1}{9}Au^{(4)}(0)\Delta t^5 + \tfrac{7}{90}u^{(5)}(0)\Delta t^5 + \tfrac{1}{9}A^2u^{(3)}(0)\Delta t^5 \\ &\quad -\tfrac{23}{180}u^{(6)}(0)\Delta t^6 - \tfrac{13}{180}Au^{(5)}(0)\Delta t^6 \\ &\quad -\tfrac{1}{18}A^2u^{(4)}(0)\Delta t^6 + \tfrac{1}{36}A^3u^{(3)}(0)\Delta t^6 . \end{aligned}$$

then $D_+^6 e^{2,0} = \mathcal{O}\left(\Delta t^2\right)$, $D_+^6 e^{2,1} = \mathcal{O}\left(\Delta t^2\right)$, $D_+^3 e^{4,0} = \mathcal{O}\left(\Delta t^4\right)$ and $D_+^3 e^{4,1} = \mathcal{O}\left(\Delta t^4\right)$.

Proof. For the error $e^{2,n} = u(t_n) - u^{2,n}$, we have

$$e^{2,n} = (3I - 2\Delta t A)^{-1}\left(4e^{2,n-1} - e^{2,n-2} + 2\Delta t f^n\right), \quad n = 2, \dots,$$

with

$$f^n = \frac{3}{2}D_+u(t_{n-1}) - \frac{1}{2}D_-u(t_{n-1}) - u'(t_n) .$$

Taking $e^{2,0} = 0$ and $e^{2,1}$ as an unknown, one can calculate the first couple of time-steps. Using Taylor expansion around $\Delta t = 0$, one arrives at

$$\begin{aligned} D_+^3 e^{2,0}\Delta t^3 &= \tfrac{4}{9}e^{2,1} - \tfrac{14}{27}Ae^{2,1}\Delta t + \tfrac{4}{9}A^2e^{2,1}\Delta t^2 \\ &\quad + \left(\tfrac{200}{243}A^3e^{2,1} + \tfrac{4}{27}u^{(3)}(0)\right)\Delta t^3 \\ &\quad + \left(\tfrac{656}{729}A^4e^{2,1} - \tfrac{8}{81}Au^{(3)}(0) - \tfrac{1}{27}u^{(4)}(0)\right)\Delta t^4 + \mathcal{O}\left(\Delta t^5\right) . \end{aligned}$$

If we choose

$$e^{2,1} = -\frac{1}{3}u^{(3)}(0)\,\Delta t^3 + \frac{1}{12}u^{(4)}(0)\,\Delta t^4 - \frac{1}{6}Au^{(3)}(0)\,\Delta t^4\,,$$

all terms up to order five vanish.

We now turn to the case of the requirements on the initial data for the sixth order solution to have optimal order of accuracy. Following Theorem 3.2, we need

$$D_+^6 e^{2,0} = \mathcal{O}\left(\Delta t^2\right)\,, \tag{44}$$

and

$$D_+^3 e^{4,0} = \mathcal{O}\left(\Delta t^4\right)\,. \tag{45}$$

Doing a similar calculation as above, one obtains that the following choice of $e^{2,1}$ guarantees that (44) holds.

$$\begin{aligned} e^{2,1} \;=\; & -\tfrac{1}{3}u^{(3)}(0)\,\Delta t^3 \\ & +\tfrac{1}{12}u^{(4)}(0)\,\Delta t^4 - \tfrac{1}{6}Au^{(3)}(0)\,\Delta t^4 \\ & -\tfrac{19}{120}u^{(5)}(0)\,\Delta t^5 - \tfrac{1}{24}Au^{(4)}(0)\,\Delta t^5 - \tfrac{1}{6}A^2u^{(3)}(0)\,\Delta t^5 \\ & +\tfrac{11}{90}u^{(6)}(0)\,\Delta t^6 + \tfrac{1}{40}Au^{(5)}(0)\,\Delta t^6 \\ & +\tfrac{1}{12}A^2u^{(4)}(0)\,\Delta t^6 - \tfrac{1}{24}A^3u^{(3)}(0)\,\Delta t^6 \\ & -\tfrac{389}{3360}u^{(7)}(0)\,\Delta t^7 - \tfrac{109}{1440}Au^{(6)}(0)\,\Delta t^7 - \tfrac{73}{480}A^2u^{(5)}(0)\,\Delta t^7 \\ & -\tfrac{3}{32}A^3u^{(4)}(0)\,\Delta t^7 - \tfrac{1}{8}A^4u^{(3)}(0)\,\Delta t^7\,. \end{aligned}$$

The error $e^{4,n} = u(t_n) - u^{4,n}$ satisfies the following equation.

$$e^{4,n} = (3I - 2\Delta tA)^{-1}\left(4e^{4,n-1} - e^{4,n-2} + 2\Delta t g^n\right)\,, \quad n = 2, \ldots\,,$$

with

$$\begin{aligned} g^n \;=\; & -\tfrac{1}{3}\Delta t^2 D_+D_-D_0e^{2,n} + \tfrac{1}{4}\Delta t^3\left(D_+D_-\right)^2 e^{2,n} \\ & +\tfrac{3}{2}D_+u(t_{n-1}) - \tfrac{1}{2}D_-u(t_{n-1}) \\ & +\tfrac{1}{3}\Delta t^2 D_+D_-D_0u(t_n) - \tfrac{1}{4}\Delta t^3\left(D_+D_-\right)^2 u(t_n) - u'(t_n)\,. \end{aligned}$$

We can calculate that the choice

$$\begin{aligned} e^{4,1} \;=\; & \tfrac{1}{9}Au^{(4)}(0)\,\Delta t^5 + \tfrac{7}{90}u^{(5)}(0)\,\Delta t^5 + \tfrac{1}{9}A^2u^{(3)}(0)\,\Delta t^5 \\ & -\tfrac{23}{180}u^{(6)}(0)\,\Delta t^6 - \tfrac{13}{180}Au^{(5)}(0)\,\Delta t^6 \\ & -\tfrac{1}{18}A^2u^{(4)}(0)\,\Delta t^6 + \tfrac{1}{36}A^3u^{(3)}(0)\,\Delta t^6 \end{aligned}$$

guarantees (45). □

From this, we can calculate choices for $u^{2,1}$ and $u^{4,1}$ to guarantee the smoothness of the initial steps. Derivatives of $u(0)$ are not known, but one can use the

differential equation (3) to obtain them. Now, from $e^{2,1}$ we can prescribe $u^{2,1} = u(\Delta t) - e^{2,1}$. Taylor expansion gives

$$\begin{aligned} u^{2,1} &= u(\Delta t) - e^{2,1} = u(0) + u'(0)\Delta t + u^{(2)}(0)\tfrac{\Delta t^2}{2} \\ &\quad + u^{(3)}(0)\tfrac{\Delta t^3}{6} + u^{(4)}(0)\tfrac{\Delta t^4}{24} - e^{2,1} + \mathcal{O}\left(\Delta t^5\right). \end{aligned}$$

B Discrete Fourier transformation of (39)

In this section, we perform the discrete Fourier transformation of the IMR (39). We will see that the approximate solution has two timescales present, even though the exact solution is such that only a slow scale is present. Performing a discrete Fourier transformation of (39), where we assume

$$u_j^{2,n} = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{u}^{2,n}(\omega)\, e^{i\omega x_j},$$

yields

$$D_+\hat{u}^{2,n}(\omega) = -ai\bar{\omega}E_+\hat{u}^{2,n}(\omega) + \hat{f}(\omega)\, e^{-i\omega t_{n+\frac{1}{2}}}. \tag{46}$$

Here $\bar{\omega} = \bar{\omega}(\omega, \Delta x)$ is an approximation of ω with $Qe^{i\omega\mathbf{x}} = i\bar{\omega}e^{i\omega\mathbf{x}}$.

The solution to (46) is

$$\begin{aligned} \hat{u}^{2,n} &= E^n\hat{u}^{2,0} + \tfrac{\Delta t\hat{f}}{1+\frac{\Delta t}{2}ai\bar{\omega}} \textstyle\sum_{k=0}^{n-1} E^{n-1-k} e^{-i\omega t_{k+\frac{1}{2}}} \\ &= E^n\left(\hat{u}^{2,0} + \tfrac{\Delta t\hat{f}e^{-i\omega\frac{\Delta t}{2}}}{1-\frac{\Delta t}{2}ai\bar{\omega}} \textstyle\sum_{k=0}^{n-1} E^{-k}\left(e^{-i\omega\Delta t}\right)^k\right) \\ &= E^n\left(\hat{u}^{2,0} + \tfrac{\Delta t e^{-i\omega\frac{\Delta t}{2}}}{1-\frac{\Delta t}{2}ai\bar{\omega}}\hat{f}\,\tfrac{1-E^{-n}\left(e^{-i\omega\Delta t}\right)^n}{1-E^{-1}(e^{-i\omega\Delta t})}\right) \\ &= E^n\left(\hat{u}^{2,0} + \tfrac{\Delta t e^{-i\omega\frac{\Delta t}{2}}}{1-\frac{\Delta t}{2}ai\bar{\omega}}\bar{f}\,\tfrac{1}{1-E^{-1}(e^{-i\omega\Delta t})}\right) \\ &\quad + e^{-i\omega t_n}\tfrac{\Delta t e^{-i\omega\frac{\Delta t}{2}}}{1-\frac{\Delta t}{2}ai\bar{\omega}}\hat{f}\,\tfrac{1}{1-E^{-1}(e^{-i\omega\Delta t})}, \end{aligned}$$

where $E = \frac{1-\frac{\Delta t}{2}ai\bar{\omega}}{1+\frac{\Delta t}{2}ai\bar{\omega}} \approx e^{-i\omega a\Delta t}$ corresponds to a fast timescale for large a. Thus, again, two timescales are present in the approximate solution. In contrast to the exact solution, we see that the approximate solution has both the fast and the slow term present.

References

[1] M. Carpenter, D. Gottlieb, S. Abarbanel, and W.-H. Don, *The theoretical accuracy of Runge-Kutta time discretizations for the initial boundary value problem: a study of the boundary error*, SIAM J. Sci. Comput., 16 (1995), pp. 1241–1252.

[2] S. Chang, *A critical analysis of the modified equation technique of Warming and Hyett*, J. Comput. Phys., 86 (1990), pp. 107–126.

[3] G. Cohen and P. Joly, *Construction and analysis of fourth-order finite difference schemes for the acoustic wave equation in nonhomogeneous media*, SIAM J. Numer. Anal., 33 (1996), pp. 1266–1302.

[4] M. V. Daele, T. V. Hecke, G. V. Berghe, and H. D. Meyer, *Deferred correction with mono-implicit Runge-Kutta methods for first-order IVPs*, J. Comput. Appl. Math., (1999).

[5] J. Daniel, V. Pereyra, and L. Schumaker, *Iterated deferred corrections for initial value problems*, Acta Cient. Venezolana, 19 (1968), pp. 128–135.

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

[7] B. Fornberg and M. Ghrist, *Spatial finite difference approximation for wave-type equations*, SIAM J. Numer. Anal., 37 (1999), pp. 105–130.

[8] L. Fox, *Some improvements in the use of relaxation methods for the solution of ordinary and partial differential equations*, Proc. Roy. Soc. London. Ser. A., 190 (1947), pp. 31–59.

[9] J. Goodman and A. Majda, *The validity of the modified equation for nonlinear shockwaves*, J. Comput. Phys., 58 (1985), pp. 336–348.

[10] D. Griffiths and J. Sanz-Serna, *On the scope of the method of modified equations*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 994–1008.

[11] B. Gustafsson and L. Hemmingsson-Frändén, *Deferred correction in space and time*, J. Sci. Comput., 17 (2002), pp. 541–550.

[12] B. Gustafsson and W. Kress, *Deferred correction methods for initial value problems*, BIT, 41 (2001), pp. 986–995.

[13] E. Hairer and G. Wanner, *Solving ordinary differential equations II*, Springer Verlag, 2nd rev. ed., 1996.

[14] G. Hedstrom, *Models of difference schemes for $u_t + u_x = 0$ by partial differential equations*, Math. Comput., 29 (1975), pp. 969–977.

[15] H. B. Keller and V. Pereyra, *Difference methods and deferred corrections for ordinary boundary value problems*, SIAM J. Numer. Anal., 16 (1979), pp. 241–259.

[16] H.-O. Kreiss, *Problems with different time scales for ordinary differential equations*, SIAM J. Numer. Anal., 16 (1979), pp. 980–998.

[17] W. Kress, *A compact fourth order time discretization method for the wave equation*, Tech. Rep. 2003-041, Dept. of Information Technology, Uppsala University, 2003.

[18] W. Kress and B. Gustafsson, *Deferred correction methods for initial value problems*, J. Sci. Comput., 17 (2002), pp. 241–252.

[19] S. K. Lele, *Compact finite difference schemes with spectral-like resolution*, J. Comput. Phys., 103 (1992), pp. 16–42.

[20] E. Mossberg, *Higher order finite difference methods for wave propagation problems.* Lic. Thesis, Uppsala University, 2002.

[21] V. Pereyra, *Iterated deferred corrections for nonlinear operator equations*, Numer. Math., 10 (1967), pp. 316–323.

[22] ——, *Iterated deferred corrections for nonlinear operator equations*, Numer. Math., 11 (1968), pp. 111–125.

[23] ——, *Highly accurate numerical solution of quasilinear elliptic boundary-value problems in n dimensions*, Math. Comput., 11 (1970), pp. 771–783.

[24] V. Pereyra, W. Proskurowski, and O. Widlund, *High order fast Laplace solvers for the Dirichlet problem on general regions*, Math. Comput., 31 (1977), pp. 1–16.

[25] R. Skeel, *A theoretical framework for proving accuracy results for deferred corrections*, SIAM J. Numer. Anal., 19 (1981), pp. 171–196.

[26] J. Tuomela, *On the construction of arbitrary order schemes for the many dimensional wave equation*, BIT, 36 (1996), pp. 158–165.

[27] F. Villatoro and J. Ramos, *On the method of modified equations I-V*, Appl. Math. and Comput., 103 (1999), pp. 111–285.