Arbitrary high order A-stable time-stepping methods via deferred correction: Application to reaction-diffusion equations

Saint-Cyr Elvi Rodrigue Koyaguerebo-Imé

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctorate in Philosophy Mathematics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The Ph.D. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

In this thesis we investigate time-stepping methods having both high order of accuracy and a good stability, for the numerical analysis of reaction-diffusion equations. The approach consists in a generalization and improvement of a time-stepping method introduced by Gustafson & Kress (2002). The time-stepping schemes from Gustafsson and Kress are built via a deferred correction (DC) strategy consisting in a successive correction (perturbation) of the trapezoidal rule, leading to a scheme of order $2j+2$ of accuracy at the stage $j = 0, 1, 2, \cdots$ of the correction. However, this method addresses only linear initial-value problems (IVP) satisfying a monotonicity condition while it has an issue for the starting procedure, and it does not take advantage of an exhaustive convergence analysis for its applicability to stiff problems. Our approach is executed into three essential steps, leading to three submitted articles. First, we introduce general formulae to derive suitable arbitrary high order finite difference approximations of analytic functions. New forms of finite difference formulae suited to various approaches of DC time-stepping schemes and the computation of their starting values, complying with the high order requirement, are constructed. Second, we introduce a general idea for the construction of different DC schemes, and we present our time-stepping method. The time-stepping method consists in a sequence $\{DC2j\}_j$ of self-starting schemes built recursively from the implicit midpoint rule via the DC strategy. A complete analysis of convergence of the method, in the case of general ordinary differential equations (ODE), is given using a deferred correction condition which guarantees an improvement by two of the order of accuracy while each scheme $DC2j$ is corrected to get the scheme $DC(2j+2)$. We prove that each $DC2j$ is A-stable. Finally, we apply our DC method to an initial boundary value problem (IBVP) related to a large class of reaction-diffusion system. The IBVP is first discretized in the time variable via the DC method, and it follows a discretization in space by the Galerkin finite element method. We prove that the resulting schemes are unconditionally and strongly stable with order $2j+2$ of accuracy in time (at the stage j of the correction). The order of accuracy in space is at least equal to the degree of the finite element used. All the theories, for ODEs and IBVP, are supported by numerical tests on various standard problems with the schemes $DC2, \cdots, DC10$. The numerical experiments comply with the theory and show that the theoretical orders of accuracy are always achieved together with a satisfactory stability.

Résumé

Cette thèse porte sur la recherche de méthodes de résolution en temps d'ordre de convergence élevé ayant de bonnes propriétés de stabilité pour la résolution numérique des équations de réaction-diffusion. L'approche adoptée consiste à améliorer et étendre les schémas en temps d'ordre arbitraire introduits par Gustafsson & Kress (2001). Les schémas de Gustafsson et Kress sont construits via une technique de correction différée appliquée à la méthode du trapèze et s'adressent uniquement aux problèmes de Cauchy linéaires satisfaisant une condition de positivité. L'analyse des propriétés de convergence de ces schémas n'est pas exhaustive, pour des applications à des problèmes d'évolution plus généraux. Notre étude pour l'amélioration et la généralisation des schémas de Gustafsson et Kress est faite en trois étapes, correspondant à trois articles de journaux. Premièrement, nous avons introduit une méthode générale pour la construction de diverses formules de différence finie pour des approximations d'ordre quelconque des dérivées des fonctions analytiques. Des formules de différence finie sont construites, lesquelles sont favorables à diverses variantes de schémas en temps par la méthode de correction différée et leur procédure de démarrage. Deuxièmement, nous avons introduit notre méthode de résolution en temps qui donne lieu à une suite récurrente , $\{DC2j\}_{j\geq 1}$, de schémas en temps à procédure de démarrage automatique dont le schéma initial $DC2$ est la règle du point milieu implicite. L'analyse complète de la convergence de ces schémas, pour le cas des problèmes de Cauchy, est basée sur une condition de correction différée qui assure un incrément de 2 de l'ordre de convergence d'un schéma $DC2j$ corrigé pour obtenir $DC(2j+2)$. Nous avons prouvé que chaque schéma $DC2j$ est A-stable. Enfin, nous avons appliqué notre méthode de résolution en temps à un problème de Cauchy-Dirichlet associé à une large classe d'équations de réaction-diffusion. Le problème est tout d'abord discrétisé en temps par la méthode de correction différée pour donner une famille de problèmes elliptiques. Chacun des problèmes elliptiques est ensuite discrétisé en espace par la méthode des éléments finis. Nous avons prouvé que les schémas totalement discrétisés sont inconditionnellement stables. L'ordre de convergence en temps est $2j+2$ (pour l'étage j de la correction) tandis que l'ordre de convergence en espace est au moins égal au degré des éléments finis considérés. La théorie est appuyée de façon satisfaisante par des résultats numériques sur des problèmes standards avec les schémas $DC2, \cdots, DC10$.

Acknowledgement

Je remercie le Seigneur, DIEU d'Amour et de Miséricorde, en qui l'espérance ma beaucoup fortifié pour ce travail.

Je remercie fortement mon Directeur de thèse, Professeur Yves Bourgault, pour son habileté qui a conduit au bon accomplissement de cette thèse. Je tiens à rendre hommage à ses qualités d'homme et d'homme de sciences.

Mes reconnaissances vont aussi à l'endroit des évaluateurs de ma thèse, les Professeurs Abbas Momeni, Abdelaziz Behljadid, Martin Weiser et Thierry Giordano, qui, par leurs commentaires, ont contribué à l'amélioration de cette thèse.

Je réitère ma gratitude au Professeur Benoît Dione, Directeur des programmes de deuxième et troisième cycle au Département de mathématiques et de statistique de l'université d'Ottawa, pour son assistance dans les difficultés inhérentes à cette thèse.

Ma pensée va également à l'endroit de tous mes camarades de l'équipe de recherche du Prof. Yves Bourgaut, Diane Fokoué, Jane Shaw MacDonald, Jacob Pearce-Lance, Sana Keita, Kak Choon Loy, et al., pour les bons moments passés ensemble. Je remercie aussi fortement le personnel du Département de mathématiques et de statisque pour leur convivialité.

Mes vifs remerciements aux enseignants du Département de mathématiques et informatique de l'université de Bangui qui m'ont beaucoup encouragé pour des recherches poussées, plus particulièrement le Pr. Rainaldy Sioké qui a toujours assuré un bon suivi de mes parcours estudiantins.

Mes remerciements vont aussi à l'endroit de mon Encadreur de Master 2, Pr. Marcel Dossa, qui m'a donné le goût de la recherche, et toute l'équipe d'AIMS-Sénegal qui m'avait facilité l'accès à cette formation doctorale.

Je ne saurais terminer sans adresser une pensée très affectueuse à mes parents, frères et soeurs dont le soutien est généreux et sans conditions depuis toujours.

Contents

General Introduction

Reaction-diffusion equations model various physical phenomena. The theory emerged in the first half of the XXth century with the work of physicists and chemists Semenov and Frank-Kamenetskii describing the temperature evolution in a closed vessel with a reacting gas, Kolmogorov-Petrovskii-Piskunov (KPP) and Fisher about the propagation of dominant genes, and Alan Turing (in 1952) modelling morphogenesis (see [19, 28] and references therein). Since the second half of the last century, these equations have attracted much interest due to applications in combustion, chemical reactions, population dynamics and biomedical science (cancer modeling and other physiological processes).

In the simplest models, the reaction-diffusion equations take the form

$$u' = M\Delta u + f(u), \qquad x \in \Omega \subset \mathbb{R}^d,\ t > 0 \tag{0.0.1}$$

where $u : \Omega \times [0, +\infty[\to \mathbb{R}^J$ is the unknown, M is an $J \times J$ matrix, and $f = f(u)$ is a smooth function [24]. Among examples of scalar reaction-diffusion equations, (0.0.1) for $f(u) = \alpha u(1-u)$, where α is a constant, gives the Fisher's equation originally used to describe the spreading of biological populations, for $f(u) = \alpha u(1-u)^2$ we have the KPP equations for planar model of advance of advantageous genes, and, for $f(u) = -au(u-\theta)(u-1)$, with $a > 0$ and $\theta \in [0, 1/2]$, we have a reduction to one variable of the FitzHugh-Nagumo model for the propagation of the depolarisation front through a nerve axon. The reference [20] reviews important models which are described by reaction-diffusion equations.

As particular cases of semi-linear parabolic equations, the mathematical analysis of the existence of solutions for reaction-diffusion equations is widely investigated. In [15, 24, 28], for example, the existence of local and global solution for particular cases of reaction-diffusion equations is proven using fixed point theorems, the notion of invariant regions and maximum principles (an explicit solution of Fisher equation is constructed in [1]).

On the other hand, the numerical analysis of reaction-diffusion equations takes advantage of many results available from the numerical analysis of semi-linear parabolic partial differential equations (PDEs). The method of lines (MOL) is commonly used. By this method the PDE is first discretized in space by finite element or finite difference methods, leading to a system of ordinary differential equations (ODEs). The

resulting system of ODEs is then discretized by fully implicit or implicit-explicit (IMEX) time-stepping methods (see for instance [2, 3, 4, 12, 14, 21, 27, 30]). In [2, 3, 4], linear implicit-explicit multistep methods in time together with finite element methods in space are analysed for a class of abstract semi-linear parabolic equations that includes a large class of reaction-diffusion systems. The approaches in [2, 3, 4] are the same. The authors investigate approximate solutions expected to be in a tube around the exact solution. They proceeded by induction by adapting the time step k and the space step h and established that if k and $k^{-1}h^{2r}$, $r \geq 2$, are small enough then the global error of the scheme is of order p ($p = 1, 2, ..., 5$) in time and r in space. IMEX schemes using finite difference in space and Runge-Kutta methods of order 1 and 2 in time are also analysed in [13, 5] for a class of reaction-diffusion systems. Otherwise, references [18, 27, 30] introduced fully implicit numerical methods for reaction-diffusion equations with restrictive conditions on the nonlinear term, combining finite elements in space and backward Euler, Crank-Nicolson or fractional-step θ methods in time. The resulting schemes are unconditionally stable (the time step is independent from the space step) with order 1 or 2 of accuracy in time.

In practice, the space-discretization of time-evolution PDEs leads to stiff initial value problems (IVP) of large dimension (we recall that a stiff problem is a problem extremely hard to solve by standard explicit step-by-step methods [25]). To avoid overly small time steps, accurate approximate solutions for these IVP require high order time-stepping methods having good stability properties (A-stable methods are of great interest). Backward differentiation formulae (BDF) of order 1 and 2 are commonly used according to their A-stability. However, BDF methods of order 3 and higher lack stability properties (e.g. for systems with complex eigenvalues). Moreover, Runge-Kutta methods applied to such IVPs have an order of convergence reduced to 1 or 2 (see, e.g., [22]) and are inefficient when the IVPs are stiffer.

The aim of this thesis is to investigate high order time-stepping methods with satisfactory stability properties for the numerical approximation of an initial boundary value problems (IBVP) related to the reaction-diffusion equation (0.0.1). The general form of the problem is,

$$\begin{cases} u' - M\Delta u + f(u) = S \text{ in } \Omega \times (0,T) \\ \qquad\qquad\qquad u = 0 \text{ on } \partial\Omega \times (0,T) \\ \qquad\qquad u(.,0) = u_0 \text{ in } \Omega. \end{cases} \tag{0.0.2}$$

Here Ω is a bounded domain with smooth boundary $\partial\Omega$, and $S : \Omega \times (0,T) \to \mathbb{R}^J$ is a given smooth function called source term. We suppose that M is positive definite with constant coefficients, and the function f satisfies the following two monotonicity conditions

$$(f(x) - f(y), x - y) \geq \alpha|x - y|^q + \tau(y)|x - y|^2, \forall x, y \in \mathbb{R}^J, \text{for some } \alpha \geq 0, q \geq 1, \tag{0.0.3}$$

and

$$(df(x)y)\cdot y \geq -\mu_0|y|^2, \quad \forall x, y \in \mathbb{R}^J, \tag{0.0.4}$$

where μ_0 is a nonnegative real constant, and τ is an arbitrary continuous real-valued function. These conditions guarantee the existence of a solution of problem (0.0.2) in $L^2(0,T;H_0^1(\Omega)\cap H^2(\Omega))$ (see for instance [9, 17, 26]), and uniqueness and high order regularity can be deduced. The conditions (0.0.3)-(0.0.4) are at least satisfied by any polynomial of odd degree with positive leading coefficient, and the Dirichlet boundary condition can be substituted by Neumann boundary conditions. Since efficient classical time-stepping methods for problems (0.0.2) have order of convergence limited to 1 or 2, we are interested in applying deferred correction method to build high order time-stepping schemes.

The deferred correction (DC) method is used to improve the order of accuracy of numerical methods of lower order. This method is explored by many authors, e.g. [6, 7, 8, 10, 11, 16, 23, 29]. The method in [7] is an application of iterative deferred correction (IDC). The authors proved that an asymptotic improvement of order p can be accomplished, from a scheme of order p, at each step of the IDC procedure, provided suitable finite difference operators are employed. Numerical experiments are performed with the IDC applied to the trapezoidal rule, Taylor-2 and Adams-Bashforth of order 2. The results are promising even though they point out some difficulties of the proposed algorithms: inacuracy for "large" time step and no asymptotic improvement for high levels of correction. The approaches in [6, 8, 10, 11, 16, 29] are quite similar and consist in a linear perturbation of a low order scheme. However, these methods are not suitable for stiff problems. For example, the method in [16], concerning a highly accurate solver for stiff ODEs and reaching order up to 14, requires sufficiently small time steps for moderately stiff problems while convergence is reduced to order 2 for "very stiff" problems. The method in [10, 14] addresses linear IVP for which a monotonicity condition is enforced, that is an IVP taking the form

$$\begin{cases} w' = Aw + F \text{ in } [0,T] \\ w(0) = w_0 \end{cases} \tag{0.0.5}$$

where the unknown w is defined from $[0,T]$ into an Hilbert space H with inner product $(\ .,\ .)$, A is a square matrix satisfying

$$(Aw, w) \leq 0, \text{ for any } w \in H,$$

and $F : [0,T] \to H$ and $w_0 \in H$ are given. The method consists in a successive correction (perturbation) of the trapezoidal rule (Crank-Nicholson) via asymptotic expansions of the linear IVP by central finite difference approximations. The order of accuracy increases by two per stage of the correction. Numerical experiments with one-dimensional linear parabolic and hyperbolic equations were performed and showed that the method is effective (orders 2, 4 and 6 of accuracy are achieved).

Our approach for the numerical analysis of the reaction-diffusion problem (0.0.2) consists in an extension and improvement of the deferred correction strategy from [10, 14] which concerns only linear IVP satisfying a monotonicity condition. The idea is motivated by the following observations:

1. The schemes from [10, 14] have an issue for the starting values when order 4 and higher are investigated. Indeed, even though these schemes can be considered as one step, the centered finite difference approximations employed lead to the computation of approximate solutions for $t < t_0 = 0$ to make a correction. This procedure is impossible for reaction-diffusion equations, as for number of IVPs and IBVPs, since the exact solutions exist only for $t \geq t_0$. The alternatives proposed in [10, 14] are the use of Runge-Kutta time-stepping methods of high order or a forward/backward finite difference approximation to compute starting values. However, high order Runge-Kutta methods as other standard time-stepping methods are inefficient for stiff problems, and both forward and backward finite difference approximations are not stable when high order approximation is needed. To overcome this difficulty new centered finite difference approximations able to compute approximate solutions inside the solution interval $[0, T]$ are needed.

2. The monotonicity condition enforced on the linear IVP analysed in [10, 14] implies that the exact solution w for this IVP is bounded independently of the operator A. The approximate solution of this problem by a trapezoidal rule preserves this property which guarantees the $\mathcal{A}$-stability of the corresponding scheme (there is a proof of $\mathcal{A}$-stability in [10] for the initial stage of the correction). Therefore, the extension of the deferred correction to more general nonlinear PDEs satisfying a monotonicity condition should be compatible with this monotonicity condition so that the stability of the schemes is guaranteed together with an optimal a priori error estimate. Unfortunately, we remark that the trapezoidal rule is not compatible with monotonicity conditions in the case of nonlinear problems. Therefore, a starting scheme preserving monotonicity conditions in the case of nonlinear IVPs is needed for building efficient DC schemes.

3. The analysis of convergence for the DC method in [10, 14], even though it concerns only linear IVPs, is not sufficient to guarantee an unconditional stability when this DC method is applied, via the MOL, to obtain a full discretization of a time-evolution PDE. In fact, if the IVP (0.0.5) results from a MOL then the matrix A is equivalent to a stiffness matrix. Therefore, A is in norm proportional to h^{-2}, where h is the space step. Since, from the convergence analysis in [10, 14], the global error constant for an approximate solution of order $2j+2$ of the problem (0.0.5) depends on A^j, an unconditional stability result can not

be obtained from this analysis. More generally, a semi-discrete approximation in space for a time-evolution PDE, which is a function $u_h : [0, T] \to V_h$, is not regular enough in space since the approximation spaces V_h are often generated by functions in the Sobolev space H_0^1. As a consequence, it is difficult to bound high order derivatives of u_h with respect to time independently from h^{-1}, and then the MOL can not lead to unconditional stability when sufficiently high order is investigated. Therefore, a complete analysis of an extension of the deferred correction from [10, 14] to high order unconditionally stable time-stepping methods for more general nonlinear PDEs requires original arguments for the proofs.

These observations lead to three submitted articles which constitute chapters 1, 2 and 3 of this thesis. The thesis is organized into three chapters as follows:

- **Chapter 1**: Corresponding to the paper "Finite difference and numerical differentiation: General formulae from deferred corrections", this chapter introduces a new approach to derive various finite difference formulae of arbitrary high order. We start by recalling basic finite difference operators and prove their main properties for the numerical analysis of DC schemes. Formulae for first and second order approximations of derivatives of analytic functions are given with error terms explicitly expanded in terms of Taylor series. These lower order approximations are successively improved by one or two (order two for centered formulae) to arbitrary high order finite difference formulae. A general theorem showing how to build arbitrary high order finite difference approximation of the derivative of any order of analytic functions is proven. Among examples of finite difference formulae constructed (for new variants of DC schemes), a new form of centered finite difference formula of arbitrary high order is given and can be used for the computation of starting values of high order time-stepping methods via DC method. The new approach recovers the standard centered, forward and backward finite difference formulae that were originally obtained in an heuristic way as formal power series of finite difference operators.

- **Chapter 2:** This chapter corresponds to the paper "Arbitrary order A-stable method for ordinary differential equations via deferred correction". It presents a general idea for the construction of DC schemes. For our DC time-stepping method, we choose the implicit midpoint rule as starting scheme. The implicit midpoint rule is successively corrected (perturbed) to obtain, at the stage $j = 0, 1, 2, \cdots$ of the correction, a self-starting scheme $DC(2j + 2)$, expected to be of order $2j + 2$ of accuracy. The analysis is restricted to the case of ODEs, in order to show the properties of the numerical method. We introduce a deferred correction condition (DCC) which guarantees the improvement of the order of accuracy by two from a scheme $DC2j$, $j = 1, 2, \cdots$, to a scheme $DC(2j + 2)$,

and we prove that each scheme $DC2j$ inherits the $\mathcal{A}$-stable property of the implicit midpoint rule. Numerical tests from standard stiff and non-stiff IVPs are performed with the schemes $DC2$, $DC4$ ..., DC10.

- **Chapter 3:** This chapter consists in an application of the DC method to the IBVP (0.0.2), and corresponds to the paper "Arbitrary high-order unconditionally stable methods for reaction-diffusion equations via Deferred Correction: Case of the implicit midpoint rule". We do not use the method of line. The IBVP is first discretized with respect to the time variable via the DC method, leading to a family of time-stepping schemes. Each semi-discrete scheme in time, corresponding to the stage $j = 0, 1, 2, \cdots$ of the correction, gives an elliptic boundary value problem (BVP) for which the existence of a solution is proven using the Schaefer fixed point theorem. The elliptic BVP is in turn discretized in space by the Galerkin finite element method, leading to a fully discrete scheme for an approximate solution of (0.0.2). We prove that each fully discretized scheme, corresponding to the stage j of the correction, is unconditionally stable and converges with order $2j + 2$ of accuracy in time and an order of accuracy in space at least equal to the degree of the finite element used. The improvement of the order of accuracy in time is, as in the case of IVPs, guaranteed by a DCC. The theory is supported by a numerical test on a bistable reaction-diffusion equation having a strong stiffness ratio.

Chapter 1

Finite difference and numerical differentiation via deferred correction: Case of uniformly spaced points

This chapter is presented in terms of a journal article that will be shortly submitted. Please see the attached paper for the content.

Finite difference and numerical differentiation via deferred correction: Case of uniformly spaced points[☆]

Saint-Cyr E.R. Koyaguerebo-Imé, Yves Bourgault

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

Abstract

This paper provides a new approach to derive explicitly different arbitrary high order finite difference formulae for the numerical differentiation of analytic functions on uniformly spaced grid points. With this approach, various first and second order formulae for the numerical differentiation of analytic functions are given with error terms explicitly expanded as Taylor series of the analytic function. These lower order approximations are successively improved by one or two (two order improvement for centered formulae) to give finite difference formulae of arbitrary high order or simply a discrete Taylor serie. The new approach allows to recover all the existing finite difference formulae on uniformly spaced grid points, and the standard backward, forward, and central finite difference formulae which are usually only given heuristically in terms of formal power series of finite difference operators. Examples of new formulae suited for deferred correction methods are given.

Keywords: finite difference formulae, numerical differentiation

1. Introduction

Finite differences (FD) are commonly used for discrete approximations of derivatives. Large classess of schemes for the numerical approximation of ordinary differential equations (ODEs) and partial differential equations (PDEs) are derived from finite differences. Formulae for numerical differentiations are generally deduced from a finite combination of Taylor series, which leads to solving a system of linear equations, or a derivative of interpolating polynomials (for instance see [1, 2, 3, 4, 5]). References [6, 7, 8] give a number of finite difference formulae, for high order approximation of derivatives, in term of formal power series of finite difference operators. The approaches in [1, 2, 3, 5] are similar. Given a set of arbitrary spaced grid points $t_0 < t_1 < \cdots < t_n$ and a function $u \in C^{n+1}([0, t_n])$, it provides a unique approximation of order $n - j + 1$ for the derivative $u^{(j)}(t_i)$ on the interpolating points $t_0 < t_1 < \cdots < t_n$, by solving the system of n equations

$$u(t_j) - u(t_i) = \sum_{m=1}^{n} \frac{(t_j - t_i)^m}{m!} u^{(m)}(t_i) + O(|t_j - t_i|^{n+1}), \quad j \neq i,$$

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

with n unknowns $u'(t_i), \cdots, u^{(n)}(t_i)$. This approach is introduced by the authors in [1, 2, 3] which consider only the case of backward, forward and central approximations on uniformly spaced grid points while reference [5] treats the general case. Reference [9] provides a closed form of the finite difference formulae from [5] in term of Lagrangian numerical differentiation formula, and the approach in [10] differs from [5] only by the treatment of error terms. The reference [11] proposes simplified analytical expressions for the numerical differentiation formulae in [5] and provides an advanced analysis of the errors for the difference formulae. Reference [12] proposes an algorithm for the finite difference approximation of derivatives that avoids dealing with a Vandermonde determinant.

Although the results in [5] adress only the numerical approximation of order $n - j + 1$ for the j-th derivative of a smooth function u at a point t_i taken from a set of distinct interpolating points $t_0 < t_1 < \cdots < t_n$, the method can be extended to the case where the point t_i is replaced by a point t_* outside the interpolating points $t_0 < t_1 < \cdots < t_n$. The later approach gives a FD formula for the approximation of $u^{(j)}(t_*)$, $j = 0, 1, 2, \cdots n$. However, this method, as other FD methods, is not suited for deferred correction (DC) methods which, for instance, is a wonderful approach to derive high order and stable time-stepping schemes for the numerical approximation of ODEs and time-evolution PDEs (see, e.g., [13, 14, 15, 16]). Indeed, finite difference formulae for DC methods should be able to provide an explicit approximation of linear combinations of derivatives with error terms explicitly expressed as Taylor series [13, 14, 17]. The errors terms for the finite difference in [5], even if it can be written as a Lagrange remainder, is implicit and cannot satisfy requirements for building DC methods.

The purpose of this paper is to provide some basic results on finite difference approximations, which results are required for the numerical analysis of higher order time-stepping schemes for ODEs and PDEs. We introduce a new approach to derive arbitrary high order finite difference formulae which avoids the need for solving a system of linear equations. We provide various formulae for the discrete approximation of any order p derivative of an analytic function u at a point t_* using $p + 1$ arbitrary points $t_0 < t_1 < \cdots < t_p$ evenly spread around t_*. These discrete approximations are of order 1 or 2 (order 2 for centred formulae), with error terms explicitly expanded in terms of Taylor series with the derivatives $u^{(p+i)}(t_*)$, $i = 1, 2, \cdots$. Substituting successively $u^{(p+1)}(t_*), u^{(p+2)}(t_*), \cdots$ by their finite difference approximations in the error term for the discrete approximation of $u^{(p)}(t_*)$, we improve successively by 1 or 2 the order of the discrete approximation of $u^{(p)}(t_*)$. An efficient choice of the discrete points minimizes the number of points needed for a given order of accuracy of the discrete approximation of $u^{(p)}(t_*)$. We give a general theorem for the derivation of finite difference approximations of any derivative of analytic functions in term of discrete Taylor series. Our approach recovers all possible finite difference formulae resulting from the method introduced by Li [5], for uniformly spaced grids points, and the standard backward, forward, and centered finite difference formulae which are given in terms of formal power series of finite difference operators. Moreover, it gives rise to various new FD formulae and constitutes a useful tool for developing new stable time-stepping methods and quadrature rules . We give three new finite difference formulae which are useful for the construction of high order time-stepping schemes and their efficient starting procedures via DC strategy. In fact, the use of standard backward and central finite differences in building high order time-stepping schemes via the DC method leads to the computation of starting values for these schemes outside the solution interval while the standard forward finite difference formula leads to unstable schemes (see, e.g., [17, 15, 16, 13, 14]).

The paper is organized as follows: in section 2 we recall the main finite difference operators and prove some of their main properties; section 3 presents general first and second order ap-

proximations of derivatives with error terms explicitly expressed as Taylor series; section 4 gives many results for arbitrary high order finite difference approximations and show how to recovers standard existing finite difference formulae from the new approach, and section 5 deals with a numerical test.

2. Properties of finite difference operators

In this section we recall the standard finite difference operators and provided some of their useful properties.

For a given spacing $k > 0$ and a real $t_0 \in \mathbb{R}$, we denote $t_n = t_0 + nk$ and $t_{n+1/2} = t_0 + (n+1/2)k$, for each integer n. The centered, forward and backward difference operators D, D_+ and D_-, respectively, related to k, and applied to a function u from $\mathbb{R}$ into a Banach space X, are defined as follows:

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

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

and

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

The average operator is denoted by E:

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

The composition of D_+ and D_- are defined recursively. They commute, that is

$$(D_+D_-)u(t_n) = (D_-D_+)u(t_n) = D_-D_+u(t_n),$$

and satisfy the identities

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

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

and

$$D_+^{m_1} D_-^{m_2} u(t_n) = k^{-m_1-m_2} \sum_{j=0}^{m_1+m_2} (-1)^j \binom{m_1+m_2}{j} u(t_{n+m_1-j}), \tag{3}$$

for each nonnegative integer m, m_1, and m_2 such that these sums exist. Formulae (1)-(3) can be proven by a straightforward induction argument.

We introduce the double index $\alpha^m = (\alpha_1^m, \alpha_2^m) \in \{0, 1, ..., m\} \times \{0, 1, ..., m\}$ such that

$$D^{\alpha^m} u(t_n) = D_+^{\alpha_1^m} D_-^{\alpha_2^m} u(t_n). \tag{4}$$

Remark 1. *If* $|\alpha^m| = \alpha_1^m + \alpha_2^m$ *is even, then we have*

$$D^{\alpha^m} u(t_n) = (D_+ D_-)^{|\alpha^m|/2} u(t_{m'}), \tag{5}$$

for some integer m'*. For example,*

$$D_+ D_-^3 u(t_n) = (D_+ D_-)^2 u(t_{n-1}),$$

and

$$D_-^4 u(t_n) = (D_+ D_-)^2 u(t_{n-2}).$$

Theorem 1 (Finite difference approximation of a product)**.** *Suppose that X is a Banach algebra. Then, for any functions* $f, g : \mathbb{R} \to X$*, we have*

$$D_-(fg)(t_n) = D_- f(t_n) g(t_n) + f(t_n) D_- g(t_n) - k D_- f(t_n) D_- g(t_n), \tag{6}$$

$$D_+(fg)(t_n) = D_+ f(t_n) g(t_n) + f(t_n) D_+ g(t_n) + k D_+ f(t_n) D_+ g(t_n), \tag{7}$$

and

$$\begin{aligned} D_+ D_-(fg)(t_n) =& D_+ D_- f(t_n) g(t_n) + f(t_n) D_+ D_- g(t_n) + D_+ f(t_n) D_- g(t_n) + D_- f(t_n) D_+ g(t_n) \\ &+ k^2 D_+ D_- f(t_n) D_+ D_- g(t_n). \end{aligned} \tag{8}$$

More generally, for each integer $m = 1, 2, ...,$ *such that* $(D_+ D_-)^m (fg)(t_n)$ *exists, we have the formula*

$$(D_+ D_-)^m (fg)(t_n) = \sum_{j=0}^{m} \binom{m}{j} k^{2j} \sum_{\alpha^m + \beta^m = (m+j, m+j)} D^{\alpha^m} f(t_n) D^{\beta^m} g(t_n). \tag{9}$$

Proof. The formulae (6)-(8) can be obtained by a straightforward calculation, so we just need to establish (9). We proceed by induction on the positive integer m. From the index notation introduced in (4), we can write

$$\begin{aligned} D_+ D_- f(t_n) g(t_n) + f(t_n) D_+ D_- g(t_n) + D_+ f(t_n) D_- g(t_n) + D_- f(t_n) D_+ g(t_n) \\ = \sum_{\alpha^1 + \beta^1 = (1,1)} D^{\alpha^1} f(t_n) D^{\beta^1} g(t_n), \end{aligned}$$

and

$$D_+ D_- f(t_n) D_+ D_- g(t_n) = D^{\alpha^1} f(t_n) D^{\beta^1} g(t_n), \quad \text{with } \alpha^1 + \beta^1 = (2, 2).$$

These two identities combined with (8) yield

$$D_+ D_-(fg)(t_n) = \sum_{j=0}^{1} \binom{1}{j} k^{2j} \sum_{\alpha^1 + \beta^1 = (1+j, 1+j)} D^{\alpha^1} f(t_n) D^{\beta^1} g(t_n),$$

that is formula (9) holds for $m = 1$. Now suppose that (9) holds until some rank $m \geq 1$. We are going to show that it remains true for $m + 1$. By the induction hypothesis, we can write

$$(D_+ D_-)^{m+1} (fg)(t_n) = \sum_{j=0}^{m} \binom{m}{j} k^{2j} \sum_{\alpha^m + \beta^m = (m+j, m+j)} D_+ D_- [D^{\alpha^m} f(t_n) D^{\beta^m} g(t_n)]. \tag{10}$$

Expanding $D_+D_-[D^{\alpha^m}f(t_n)D^{\beta^m}g(t_n)]$ as in the formula (8), we deduce that

$$\sum_{\alpha^m+\beta^m=(m+j,m+j)} D_+D_-[D^{\alpha^m}f(t_n)D^{\beta^m}g(t_n)] = S(j) + k^2S(j+1), \tag{11}$$

where

$$S(j) = \sum_{\alpha^{m+1}+\beta^{m+1}=(m+1+j,m+1+j)} D^{\alpha^{m+1}}f(t_n)D^{\beta^{m+1}}g(t_n).$$

We have

$$\sum_{j=0}^{m}\binom{m}{j}k^{2j}[S(j)+k^2S(j+1)] = S(0) + \sum_{j=1}^{m}k^{2j}\left[\binom{m}{j-1}+\binom{m}{j}\right]S(j) + k^{2m+2}S(m+1),$$

and deduce from (10), (11) and the identity $\binom{m}{j}+\binom{m}{j-1} = \binom{m+1}{j}$ that the formula (9) holds for $m+1$. Finally, we conclude by induction that this formula is true for each suitable positive integer m. □

Theorem 2 (Finite difference approximation of a composite)**.** *Consider two functions f and u with values into Banach spaces such that f is differentiable, and the composite $f \circ u$ is defined on $\mathbb{R}$. Then*

$$D_-f(u(t_n)) = \int_0^1 df\left(u(t_{n-1}) + \tau k D_-u(t_n)\right)(D_-u(t_n))d\tau \tag{12}$$

and

$$D_+f(u(t_n)) = \int_0^1 df\left(u(t) + \Delta t D_+u(t)\tau\right)(D_+u(t))d\tau \tag{13}$$

Proof. As in standard mean value theorem. □

3. First and second order discrete approximation of derivatives

In this section we provide various formulae for the finite difference approximation of arbitrary high order derivatives of analytic functions. The approximations are of order one or two, and the error terms are explicitly expanded in terms of Taylor series. We need the following lemma whose proof is an easy induction.

Lemma 1. *For positive integers m and p and for any real r, we have*

$$\sum_{j=0}^{m}(-1)^j\binom{m}{j}(m+r-j)^p = \begin{cases} 0, & \text{if } 1 \le p < m, \\ m!, & \text{if } p = m. \end{cases} \tag{14}$$

In particular, for any nonnegative integer p, we have

$$\sum_{j=0}^{2m}(-1)^j\binom{2m}{j}(m-j)^{2p+1} = 0, \tag{15}$$

$$\sum_{j=0}^{2m+1}(-1)^j\binom{2m+1}{j}(m-j+1/2)^{2p} = 0, \tag{16}$$

and

$$\sum_{j=0}^{2m}(-1)^j\binom{2m}{j}\Big[(m-j+1/2)^{2p+1}+(m-j-1/2)^{2p+1}\Big]=0. \tag{17}$$

Theorem 3. *Suppose that the function $u:[0,T]\to X$ is analytic on an open interval contaning $[0,T]$. Let $0=t_0<t_1<...<t_N=T$, $t_n=nk$, be a partition of the interval $[0,T]$. For each positive integer m, we have*

$$D_+^m u(t_n)=u^{(m)}(t_n)+\sum_{i=m+1}^{\infty}\frac{k^{i-m}}{i!}u^{(i)}(t_n)\sum_{j=0}^{m}(-1)^j\binom{m}{j}(m-j)^i, \tag{18}$$

$$D_-^m u(t_n)=u^{(m)}(t_n)+\sum_{i=m+1}^{\infty}\frac{k^{i-m}}{i!}u^{(i)}(t_n)\sum_{j=0}^{m}(m-1)^j\binom{m}{j}(-j)^i, \tag{19}$$

$$\begin{aligned}D_-(D_+D_-)^m u(t_n)&=u^{(2m+1)}(t_n)\\&+\sum_{i=2m+2}^{\infty}\frac{k^{i-2m-1}}{i!}u^{(i)}(t_n)\sum_{j=0}^{2m+1}(-1)^j\binom{2m+1}{j}(m-j)^i,\end{aligned} \tag{20}$$

$$(D_+D_-)^m u(t_n)=u^{(2m)}(t_n)+\sum_{i=m+1}^{\infty}\frac{k^{2i-2m}}{(2i)!}u^{(2i)}(t_n)\sum_{j=0}^{2m}(-1)^j\binom{2m}{j}(m-j)^{2i}, \tag{21}$$

$$\begin{aligned}D(D_+D_-)^m u(t_{n+1/2})&=u^{(2m+1)}(t_{n+1/2})\\&+\sum_{i=m+1}^{\infty}\frac{k^{2i-2m}}{(2i+1)!}u^{(2i+1)}(t_{n+1/2})\sum_{j=0}^{2m+1}(-1)^j\binom{2m+1}{j}(m-j-1/2)^{2i+1},\end{aligned} \tag{22}$$

and

$$(D_+D_-)^m Eu(t_{n+1/2})=u^{(2m)}(t_{n+1/2})+\sum_{i=m+1}^{\infty}a_{mi}\frac{k^{2i-2m}}{(2i)!}u^{(2i)}(t_{n+1/2}), \tag{23}$$

where

$$a_{mi}=\frac{1}{2}\sum_{j=0}^{2m}(-1)^j\binom{2m}{j}\Big[(m-j+1/2)^{2i}+(m-j-1/2)^{2i}\Big].$$

Proof. We only prove formula (22). The other formulae can be proven similarly. By Taylor expansion series we have

$$u(t_{n+m-j})=u(t_{n+s})+\sum_{i=1}^{\infty}\frac{k^i}{i!}(m-s-j)^i u^{(i)}(t_{n+s}).$$

Choosing $s=1/2$ in this formula, we deduce from (2) that

$$\begin{aligned}D(D_+D_-)^m u(t_{n+1/2})=&k^{-2m-1}\sum_{j=0}^{2m+1}(-1)^j\binom{2m+1}{j}u(t_{n+m-j})\\&=k^{-2m-1}\sum_{i=1}^{\infty}\frac{k^i}{i!}u^{(i)}(t_{n+1/2})\sum_{j=0}^{2m+1}(-1)^j\binom{2m+1}{j}(m-j-1/2)^i,\end{aligned}$$

and (22) follows from (14) and (16). □

Theorem 4. *Let u be $C^m([0,T],X)$, $m = 1, 2, ...$, and $0 = t_0 < t_1 < ... < t_N = T$, $t_n = nk$, be a partition $[0,T]$. Let m_1 and m_2 be two positive integers such that $m_1 + m_2 \leq m$. Then, for each integer n such that $m_2 \leq n \leq N - m_1$, $D_+^{m_1} D_-^{m_2} u(t_n)$ is bounded independently of n, and we have the estimate*

$$\left\| D_+^{m_1} D_-^{m_2} u(t_n) \right\| \leq C \max_{t_{n-m2} \leq t \leq t_{n+m_1}} \left\| u^{(m_1+m_2)}(t) \right\|,$$

where C is a constant depending only on the integer m.

Proof. According to Remark 1, it is enough to just prove the theorem for $(D_+D_-)^p f(t_n)$ or $D_-(D_+D_-)^p f(t_n)$, for suitable positive integer p (the case $p = 0$ is trivial). As in the previous proof, Taylor expansion of order $(2p-1)$ with integral remainder together with formulae (1) and (14) yields

$$(D_+D_-)^p u(t_n) = \sum_{j=0}^{2p} \frac{(-1)^j}{(2p-1)!} \binom{2p}{j} (p-j)^{2p} \int_0^1 (1-s)^{2p-1} u^{(2p)}(t_n + (p-j)ks) ds.$$

It follows that

$$\|(D_+D_-)^p u(t_n)\| \leq \frac{1}{(2p)!} \sum_{j=0}^{2p} \binom{2p}{j} (p-j)^{2p} \max_{t_{n-p} \leq t \leq t_{n+p}} \left\| u^{(2p)}(t) \right\|.$$

Similar reasoning can be applied in the case of $D_-(D_+D_-)^p u(t_n)$.

□

4. Arbitrary high order finite difference approximations

Theorem 5. *There exists a sequence $\{c_i\}_{i \geq 2}$ of real numbers such that for any function $u \in C^{2p+3}([0,T],X)$, where p is a positive integer, and a partition $0 = t_0 < t_1 < ... < t_N = T$, $t_n = nk$, of $[0,T]$, we have*

$$u'(t_{n+1/2}) = \frac{u(t_{n+1}) - u(t_n)}{k} - \sum_{i=1}^{p} c_{2i+1} k^{2i} D(D_+D_-)^i u(t_{n+1/2}) + O(k^{2p+2}), \tag{24}$$

and

$$u(t_{n+1/2}) = \frac{u(t_{n+1}) + u(t_n)}{2} - \sum_{i=1}^{p} c_{2i} k^{2i} (D_+D_-)^i E u(t_{n+1/2}) + O(k^{2p+2}), \tag{25}$$

for $p \leq n \leq N - 1 - p$. The error constants for the formulae (24) and (25) are, respectively, c_{2p+3} and c_{2p+2}. Table 1 gives the first ten coefficients c_i.

Table 1: Ten first coefficients of central difference approximations (24) and (25)

c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}
$\frac{1}{8}$	$\frac{1}{24}$	$-\frac{18}{4!2^5}$	$-\frac{18}{5!2^5}$	$\frac{450}{6!2^7}$	$\frac{450}{7!2^7}$	$-\frac{22050}{8!2^9}$	$-\frac{22050}{9!2^9}$	$\frac{1786050}{10!2^{11}}$	$\frac{1786050}{11!2^{11}}$

Proof. By Taylor expansion we can write

$$u(t_{n+1}) = u(t_n) + ku'(t_{n+1/2}) + \sum_{i=1}^{p} \frac{d_{1,2i+1}}{(2i+1)!} k^{2i+1} u^{(2i+1)}(t_{n+1/2}) + O(k^{2p+3}) \tag{26}$$

and

$$u(t_{n+1}) = -u(t_n) + 2u(t_{n+1/2}) + \sum_{i=1}^{p} \frac{d_{1,2i}}{(2i)!} k^{2i} u^{(2i)}(t_{n+1/2}) + O(k^{2p+2}), \tag{27}$$

with $d_{1,i} = 2^{1-i}$, for $i = 2, 3, ..., 2p+1$. Therefore, substituting successively the derivatives $u^{(3)}(u_{n+1/2})$, $u^{(5)}(t_{n+1/2})$, ... and $u^{(2)}(t_{n+1/2})$, $u^{(4)}(t_{n+1/2})$, ... by their expansion given by the formulae (22) and (23), respectively, into (26) and (27), we deduce the identities

$$\begin{aligned} u(t_{n+1}) &= u(t_n) + ku'(t_{n+\frac{1}{2}}) + \frac{d_{1,3}}{3!} k^3 DD_+D_- u(t_{n+\frac{1}{2}}) + ... \\ &+ \frac{d_{q,2q+1}}{(2q+1)!} k^{2q+1} D(D_+D_-)^q u(t_{n+\frac{1}{2}}) + \sum_{i=q+1}^{p} \frac{d_{q+1,2i+1}}{(2i+1)!} k^{2i+1} u^{(2i+1)}(t_{n+\frac{1}{2}}) + O(k^{2p+3}) \end{aligned}$$

and

$$\begin{aligned} u(t_{n+1}) &= -u(t_n) + 2u(t_{n+1/2}) + \frac{d_{1,2}}{2!} k^2 D_+D_- Eu(t_{n+1/2}) + ... \\ &+ \frac{d_{q,2q}}{(2q)!} k^{2q} (D_+D_-)^q Eu(t_{n+1/2}) + \sum_{i=q+1}^{p} \frac{d_{q+1,2i}}{(2i)!} k^{2i} u^{(2i)}(t_{n+1/2}) + O(k^{2p+2}) \end{aligned}$$

where, for $q = 1, ..., p-1$, and $i = q+1, q+2, ..., p$, we have

$$d_{q+1,2i+1} = d_{q,2i+1} - \frac{d_{q,2q+1}}{(2q+1)!} \sum_{j=0}^{2q+1} (-1)^j \binom{2q+1}{j} (q - j - 1/2)^{2i+1},$$

and

$$d_{q+1,2i} = d_{q,2i} - \frac{d_{q,2q}}{(2q)! \times 2} \sum_{j=0}^{2q} (-1)^j \binom{2q}{j} [(q - j - 1/2)^{2i} + (q - j - 3/2)^{2i}].$$

Finally, the identities (24) and (25) follow by setting $c_{2i} = d_{i,2i}/((2i)! \times 2)$ and $c_{2i+1} = d_{i,2i+1}/(2i+1)!$, for $i = 1, 2, ..., p$. □

Remark 2. *The approximations (24) and (25) are, from the coefficients c_i computed in Table 1, equivalent to the central-difference approximation of the first derivative and the centered Bessel's formulae (see [6, p.142 & p.183] or [7, 8]).*

Remark 3. *Formula (24) gives the finite difference approximations in [2], writing*

$$u'(t_n) = \frac{u(t_{n+1/2}) - u(t_{n-1/2})}{k} - \sum_{i=1}^{p} c_{2i+1} k^{2i} D(D_+D_-)^i u(t_n) + O(k^{2p+2}), \tag{28}$$

where

$$\sum_{i=1}^{p} c_{2i+1} k^{2i} D(D_+D_-)^i u(t_n) = k^{-1} \sum_{i=1}^{p} \left[c_{2i+1} \sum_{j=0}^{2i+1} (-1)^j \binom{2i+1}{j} u(t_{n+i-j+1/2}) \right].$$

- *For $p = 1$ we have*

$$\begin{aligned} u'(t_n) &= \frac{u(t_{n+1/2}) - u(t_{n-1/2})}{k} - \frac{1}{24}k^2 D(D_+D_-)u(t_n) + O(k^4) \\ &= \frac{u(t_{n+1/2}) - u(t_{n-1/2})}{k} - \frac{u(t_{n+3/2}) - 3u(t_{n+1/2}) + 3u(t_{n-1/2}) - u(t_{n-3/2})}{24k} \\ &\quad + O(k^4). \end{aligned}$$

- *For $p = 2$ we have*

$$\begin{aligned} u'(t_n) &= \frac{u(t_{n+1/2}) - u(t_{n-1/2})}{k} - \frac{1}{24}k^2 D(D_+D_-)u(t_n) + \frac{18}{2^5 5!}k^4 D(D_+D_-)^2 u(t_n) \\ &\quad + O(k^6), \end{aligned}$$

and then

$$u'(t_n) = \frac{u(t_{n+1/2}) - u(t_{n-1/2})}{k} + \frac{1}{1920k}\begin{bmatrix} 9 & -125 & 330 & -330 & 125 & -9 \end{bmatrix} U_{n,5}^T + O(k^6),$$

where $U_{n,5}^T$ is the transpose of the vector

$$U_{n,5} = \begin{bmatrix} u(t_{n+5/2}) & u(t_{n+3/2}) & u(t_{n+1/2}) & u(t_{n-1/2}) & u(t_{n-3/2}) & u(t_{n-5/2}) \end{bmatrix}.$$

The following theorem gives a new form of centered finite difference formulae which is useful for efficient starting procedures of high order time-stepping schemes via deferred correction strategy [13, 14].

Theorem 6 (Interior centered approximations)**.** *Let $u \in C^{2p+3}([a,b], X)$, where p is a positive integer and $[a,b]$, $a < b$, is a real interval. Given a uniform partition $a = \tau_0 < \tau_1 < ... < \tau_{2p+1} = b$ of $[a,b]$, that is $\tau_n = a + nk$ with $k = (b-a)/(2p+1)$, and $\tau_{p+1/2} = (a+b)/2$, there exist reals $c_2^p, c_3^p, \cdots, c_{2p+1}^p$ such that*

$$u'(\tau_{p+1/2}) = \frac{u(b) - u(a)}{b-a} - \frac{1}{b-a}\sum_{i=1}^{p} c_{2i+1}^p k^{2i+1} D(D_+D_-)^i u(\tau_{p+1/2}) + O(k^{2p+2}). \tag{29}$$

and

$$u(\tau_{p+1/2}) = \frac{u(b) + u(a)}{2} - \sum_{i=1}^{p} c_{2i}^p k^{2i} (D_+D_-)^i E u(\tau_{p+1/2}) + O(k^{2p+2}), \tag{30}$$

Table 2 gives the coefficients c_i^p for $p = 1, 2, 3, 4$.

Proof. By Taylor expansion we have

$$u(b) = u(a) - (b-a)u'(\tau_{p+1/2}) + \sum_{i=1}^{p} \frac{(b-a)^{2i+1}}{2^{2i}(2i+1)!} u^{(2i+1)}(\tau_{p+1/2}), + O((b-a)^{2p+3}),$$

and

$$u(b) = -u(a) + 2u(\tau_{p+1/2}) + \sum_{i=1}^{p} \frac{(b-a)^{2i}}{2^{2i-1}(2i)!} u^{(2i)}(\tau_{p+1/2}) + O((b-a)^{2p+2}).$$

Table 2: Coefficients of the approximations (29)-(30) for $p = 1, 2, 3, 4$

p	c_2^p	c_3^p	c_4^p	c_5^p	c_6^p	c_7^p	c_8^p	c_9^p
1	$\frac{9}{8}$	$\frac{9}{8}$						
2	$\frac{25}{8}$	$\frac{125}{24}$	$\frac{125}{128}$	$\frac{125}{128}$				
3	$\frac{49}{8}$	$\frac{343}{24}$	$\frac{637}{128}$	$\frac{13377}{1920}$	$\frac{1029}{1024}$	$\frac{1029}{1024}$		
4	$\frac{81}{8}$	$\frac{243}{8}$	$\frac{1917}{128}$	$\frac{17253}{640}$	$\frac{7173}{1024}$	$\frac{64557}{7168}$	$\frac{32733}{32768}$	$\frac{32733}{32768}$

Substituting $b - a$ by $(2p + 1)k$ in the summations, we deduce that

$$u(b) = u(a) + (b - a)u'(\tau_{p+1/2}) + \sum_{i=1}^{p} \frac{d_{1,2i+1}^p}{(2i + 1)!} k^{2i+1} u^{(2i+1)}(\tau_{p+1/2}) + O(k^{2p+3}),$$

and

$$u(b) = -u(a) + 2u(\tau_{p+1/2}) + \sum_{i=1}^{p} \frac{d_{1,2i}^p}{(2i)!} k^{2i} u^{(2i)}(t_{p+1/2}) + O(k^{2p+2}),$$

where

$$d_{1,i}^p = 2^{1-i}(2p + 1)^i, \text{ for } i = 1, \cdots, 2p + 1.$$

Proceeding exactly as in Theorem 5, we obtain the real $d_{q,i}^p$ such that, for $q = 1, ..., p - 1$ and $i = q + 1, q + 2, ..., p$, we have

$$d_{q+1,2i+1}^p = d_{q,2i+1}^p - \frac{d_{q,2q+1}^p}{(2q + 1)!} \sum_{j=0}^{2q+1} (-1)^j \binom{2q + 1}{j} (q - j - 1/2)^{2i+1},$$

and

$$d_{q+1,2i}^p = d_{q,2i}^p - \frac{d_{q,2q}^p}{(2q)! \times 2} \sum_{j=0}^{2q} (-1)^j \binom{2q}{j} \left[(q - j - 1/2)^{2i} + (q - j + 1/2)^{2i} \right].$$

Finally, $c_{2i}^p = d_{i,2i}^p/((2i)! \times 2)$ and $c_{2i+1}^p = d_{i,2i+1}^p/(2i + 1)!$, for $i = 1, 2, ..., p$. □

The following finite difference formulae are useful for the construction of new time-stepping methods by applying the deferred correction method to first order backward or forward schemes. These formulae agree with those given in [5] but differ by their special form suited for DC methods.

Theorem 7. *(Forward-centered and backward-centered approximations)*
There exist two sequences $\{a_i\}_{i\geq 2}$ and $\{b_i\}_{i\geq 2}$ of real numbers such that, for any function $u \in C^{p+1}([0, T], X)$ and a partition $0 = t_0 < t_1 < ... < t_N = T$, $t_n = nk$, of $[0, T]$, we have

$$u'(t_n) = \frac{u(t_{n+1}) - u(t_n)}{k} - \sum_{i=2}^{p} a_i k^{i-1} D_-^{\tau(i)} (D_+ D_-)^{\mu(i)} u(t_n) + O(k^p), \tag{31}$$

and

$$u'(t_{n+1}) = \frac{u(t_{n+1}) - u(t_n)}{k} + \sum_{i=2}^{p} b_i k^{i-1} D_-^{\tau(i)} (D_+ D_-)^{\mu(i)} u(t_{n+1}) + O(k^p), \tag{32}$$

for $\mu(p)+\tau(p) \leq n \leq N-\mu(p)$, where $\mu(i)$ and $\tau(i)$ are, respectively, the quotient and the remainder of the Euclidean division of the integer i by 2, that is $i = 2\mu(i)+\tau(i)$, $\tau(i) = 0$ or 1. The errors constants for the finite differences approximations (31)-(32) are a_{p+1} and b_{p+1}, respectively, and we have the relation $a_2 = b_2$, and $a_i = -b_i$, for $i = 3, 4, \cdots$.

Table 3 gives the coefficients a_i, for $i = 2, 3, \cdots, 11$.

Table 3: Table of coefficients, for differed correction backward Euler method.

a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
$\frac{1}{2}$	$\frac{1}{3!}$	$\frac{2}{4!}$	$-\frac{4}{5!}$	$-\frac{12}{6!}$	$\frac{36}{7!}$	$\frac{144}{8!}$	$-\frac{576}{9!}$	$-\frac{2880}{10!}$	$\frac{14400}{11!}$

Proof. Taylor expansion of the function u at order p around $t = t_n$ gives

$$u(t_{n+1}) = u(t_n) + A_{1,1}ku'(t_n) + \sum_{i=2}^{p} A_{1,i}\frac{k^i}{i!}u^{(i)}(t_n) + O(k^{p+1}), \tag{33}$$

where $A_{1,i} = 1$, for $i = 1, 2, 3, ..., p$. Suppose that

$$\begin{aligned} u(t_{n+1}) &= u(t_n) + A_{1,1}ku'(t_n) + A_{2,2}k^2D_+D_-u(t_n) + A_{3,3}k^3D_-(D_+D_-)u(t_n) + ... \\ &+ A_{q-1,q-1}k^{q-1}D_-^{\tau(q-1)}(D_+D_-)^{\mu(q-1)}u(t_n) + \sum_{i=q}^{p} A_{q-1,i}k^iu^{(i)}(t_n) + O(k^{p+1}), \end{aligned} \tag{34}$$

for an arbitrary integer $q \geq 2$, where (33) is the formula for $q = 2$. From (20)-(21) and (15) we have

$$u^{(q)}(t_n) = D_-^{\tau(q)}(D_+D_-)^{\mu(q)}u(t_n) - \sum_{i=q+1}^{\infty} \frac{k^{i-q}}{i!}u^{(i)}(t_n)\sum_{j=0}^{q}(-1)^j\binom{q}{j}(\mu(q)-j)^i,$$

and it follows that

$$\begin{aligned} \sum_{i=q}^{p} A_{q-1,i}k^iu^{(i)}(t_n) &= A_{q-1,q}\frac{k^q}{q!}u^{(q)}(t_n) + \sum_{i=q+1}^{p} A_{q-1,i}\frac{k^i}{i!}u^{(i)}(t_n) = A_{q-1,q}\frac{k^q}{q!}D_-^{\tau(q)}(D_+D_-)^{\mu(q)}u(t_n) \\ &+ \sum_{i=q+1}^{p}\left(A_{q-1,i} - \frac{A_{q-1,q}}{q!}\sum_{j=0}^{q}(-1)^j\binom{q}{j}(\mu(q)-j)^i\right)\frac{k^i}{i!}u^{(i)}(t_n) + O(k^{p+1}). \end{aligned}$$

Substituting the last identity in (34), we deduce that

$$\begin{aligned} u(t_{n+1}) &= u(t_n) + ku'(t_n) + A_{2,2}k^2D_+D_-u(t_n) + A_{3,3}k^3D_-(D_+D_-)u(t_n) + ... \\ &+ A_{q,q}k^qD_-^{\tau(q)}(D_+D_-)^{\mu(q)}u(t_n) + \sum_{i=q+1}^{p} A_{q,i}k^iu^{(i)}(t_n) + O(k^{p+1}), \end{aligned}$$

where, for $q = 2, 3, \cdots, p$ we have

$$A_{q,q} = A_{q-1,q}$$

and

$$A_{q,i} = A_{q-1,i} - \frac{A_{q,q}}{q!}\sum_{j=0}^{q}(-1)^j\binom{q}{j}(\mu(q)-j)^i\,, \text{for } i = q+1, q+2, ..., p.$$

We can then deduce by induction on q that formula (31) holds with $a_i = A_{i,i}$, for $i = 2, ..., p$. The sequence $\{b_i\}_{i\geq 2}$ can be obtained similarly.

□

Remark 4. *The standard forward formula writes*

$$u'(t_n) = \frac{u(t_{n+1}) - u(t_n)}{k} - \sum_{i=2}^{p}\frac{(-1)^i}{i}k^{i-1}D_+^i u(t_n) + O(k^p). \tag{35}$$

It can be obtained by substituting successively the derivative $u^{(2)}(t_n)$, $u^{(3)}(t_n)$, ..., in (33) by the expansion (18), and the standard backward formula writes

$$u'(t_{n+1}) = \frac{u(t_{n+1}) - u(t_n)}{k} + \sum_{i=2}^{p}\frac{1}{i}k^{i-1}D_-^i u(t_{n+1}) + O(k^p), \tag{36}$$

and can be obtained from (19). The errors constants in the new forward-centered and backward-centered formulae are smaller than for the standard forward and backward formulae (35) and (36), respectively. For example, the error constant for an approximation of order 10 for $u'(t_n)$ by the formulae (35)-(36) is $1/11$ *while the corresponding error constant for (31)-(32) is* $14400/11!$.

More generally, we have the following result:

Theorem 8 (General finite difference formulae)**.** *For an analytic function $u : \mathbb{R} \longrightarrow X$, given an integer m and a real $k > 0$, we can write, for any integer $p \geq m$ and a real t,*

$$u^{(m)}(t) = k^{-m}\sum_{i=m}^{p}\sum_{|\alpha^i|=i} C_{\alpha^i}(k_i)^i D^{\alpha^i}u(t) + O(k^{p+1-m}), \tag{37}$$

where C_{α^i} are constants, $k_m = k$, $k_i = \varepsilon_i k$ (for $i \geq m+1$, where $\varepsilon_i > 0$ is arbitrarily chosen), and each finite difference operator D^{α^i} is related to k_i in the sense that

$$(k_i)^i D^{\alpha^i}u(t) = \sum_{j=0}^{i}(-1)^j\binom{i}{j}u\left(t + (\alpha_1^i - j)k_i\right), \text{ for } |\alpha^i| = i. \tag{38}$$

Proof. For a double index $\alpha^i = (\alpha_1^i, \alpha_2^i)$ such that $|\alpha^i| = i$ and a spacing $k_i > 0$, since D^{α^i} is related to $k_i > 0$, we deduce from (38) and Theorem 3 that

$$u^{(i)}(t) = D^{\alpha^i}u(t) - \sum_{l=i+1}^{\infty}\frac{(k_i)^{l-i}}{l!}u^{(l)}(t)\sum_{j=0}^{i}(-1)^j\binom{i}{j}\left(\alpha_1^i - j\right)^l. \tag{39}$$

Therefore, we can choose one double index α^m such that $|\alpha^m| = m$ and deduce that

$$k^m u^{(m)}(t) = k^m D^{\alpha^m}u(t) - \sum_{l=m+1}^{\infty}\frac{k^l}{l!}u^{(l)}(t)\sum_{j=0}^{m}(-1)^j\binom{m}{j}\left(\alpha_1^m - j\right)^l.$$

This identity can be written

$$k^m u^{(m)}(t) = k^m D^{\alpha^m} u(t) + \sum_{l=m+1}^{\infty} C_{m+1,l} \frac{(k_{m+1})^l}{l!} u^{(l)}(t), \tag{40}$$

where $k_{m+1} = \varepsilon_{m+1} k$, for a real $\varepsilon_{m+1} > 0$ arbitrarily chosen, and

$$C_{m+1,l} = -(\varepsilon_{m+1})^{-l} \sum_{j=0}^{m} (-1)^j \binom{m}{j} \left(\alpha_1^m - j\right)^l, \text{ for } l \geq m+1.$$

Next, we choose one double index α^{m+1} such that $|\alpha^{m+1}| = m+1$ and substitute the identity (39) for $i = m+1$ into (40) to obtain

$$k^m u^{(m)}(t) = k^m D^{\alpha^m} u(t) + C_{m+1,m+1} (k_{m+1})^{m+1} D^{\alpha^{m+1}} u(t) + \sum_{l=m+2}^{\infty} C_{m+2,l} \frac{(k_{m+2})^l}{l!} u^{(l)}(t), \tag{41}$$

where $k_{m+2} = \varepsilon_{m+2} k_{m+1}$, for a real $\varepsilon_{m+2} > 0$ arbitrarily chosen, and, for $l \geq m+2$,

$$C_{m+2,l} = (\varepsilon_{m+2})^{-l} \left(C_{m+1,l} - \frac{C_{m+1,m+1}}{(m+1)!} \sum_{j=0}^{m+1} (-1)^j \binom{m+1}{j} \left(\alpha_1^{m+1} - j\right)^l \right).$$

This procedure is repeated until obtaining the expected order of accuracy. □

Remark 5. *As a simple application of Theorem 8, the standard central difference for the second derivative (see, e.g., [7, Formulae (3.3.10)-(3.3.11)]) can be obtained as follows: We choose* $m = 1$ *in formula (21) and obtain*

$$k^2 u''(t_n) = k^2 (D_+ D_-) u(t_n) - 2 \sum_{i=2}^{\infty} \frac{k^{2i}}{(2i)!} u^{(2i)}(t_n), \tag{42}$$

which is the second order approximation of $u''(t_n)$ *with error constant* $K_2 = -1/12$. *The same formula for* $m = 2$ *gives*

$$k^4 u^{(4)}(t_n) = k^4 (D_+ D_-)^2 u(t_n) - \sum_{i=3}^{\infty} \frac{k^{2i}}{(2i)!} u^{(2i)}(t_n) \sum_{j=0}^{4} (-1)^j \binom{4}{j} (2-j)^{2i}.$$

Substituting the last identity in (42), we deduce that

$$k^2 u''(t_n) = k^2 (D_+ D_-) u(t_n) - \frac{2k^4}{4!} (D_+ D_-)^2 u(t_n) + \sum_{i=3}^{\infty} \left(-2 + \frac{2}{4!} \sum_{j=0}^{4} (-1)^j \binom{4}{j} (2-j)^{2i} \right) \frac{k^{2i}}{(2i)!} u^{(2i)}(t_n).$$

The last formula gives the approximation of order 4 for $u''(t_n)$ *with error constant*

$$K_4 = \left(-2 + \frac{2}{4!} \sum_{j=0}^{4} (-1)^j \binom{4}{j} (2-j)^6 \right) \frac{1}{6!} = \frac{1}{90}.$$

The arbitrary high order central difference can be obtained by continuing the procedure.

Remark 6. *The finite differences introduced by Li (see [5]), for uniformly spaced grid points $t_0 < t_1 < \cdots < t_n$, $t_n = t_0 + nk$, can be recovered by the following formulae:*
- For the boundary points t_0 and t_n, the approximations for $u^{(p)}(t_n)$ and $u^{(p)}(t_0)$, $1 \leq p \leq n$, are obtained by the formulae

$$k^p D_-^p u(t_n) = k^p u^{(p)}(t_n) + \sum_{m=p+1}^{n} \alpha_{p,n,m} k^m D_-^m u(t_n) + O(k^{n+1}),$$

and

$$k^p D_+^p u(t_0) = k^p u^{(p)}(t_0) + \sum_{m=p+1}^{n} \alpha_{p,0,m} k^m D_+^m u(t_0) + O(k^{n+1}).$$

- For $1 \leq j < n$, the approximation for $u'(t_j)$ can be deduced from the formulae

$$kD_- u(t_j) = ku'(t_j) + \sum_{m=2}^{j} \alpha_{1,j,m} k^m D_-^m u(t_j) + \sum_{m=j+1}^{n} \alpha_{1,j,m} k^m D_+^{m-j} D_-^j u(t_j) + O(k^{n+1})$$

- The approximation for $u"(t_j)$, $1 \leq j < n$, can be deduced from the following formulae. For $j = 1$ we have

$$k^2 D_+ D_- u(t_1) = k^2 u"(t_1) + \sum_{m=3}^{n} \alpha_{2,1,m} k^m D_+^{m-1} D_- u(t_1) + O(k^{n+1}),$$

and for $2 \leq j < n$,

$$k^2 D_-^2 u(t_j) = k^2 u"(t_j) + \sum_{m=3}^{j} \alpha_{2,j,m} k^m D_-^m u(t_j) + \sum_{m=j+1}^{n} \alpha_{2,j,m} k^m D_+^{m-j} D_-^j u(t_j) + O(k^{n+1}).$$

The approximation for $u^{(i)}(t_j)$, $1 \leq j < n$ and $i = 3, 4, \cdots n$ can be obtained similarly, and the coefficients $\alpha_{j,i,m}$ result from Theorem 8.

Remark 7. *To obtain a FD formula for the numerical approximation $u'(t_{i+1/2})$, $0 \leq i < n$, for example, using the interpolating points $t_0 < t_1 < \cdots < t_n$, $t_n = t_0 + nk$, one can use formula (22) and obtains*

$$Du(t_{i+1/2}) = ku'(t_{i+1/2}) + \sum_{m=1}^{\infty} \frac{k^{2m+1}}{(2m+1)! \cdot 2^{2m}} u^{(2m+1)}(t_{i+1/2}).$$

Then, an application of Theorem 8 provides coefficients $\beta_3, \cdots, \beta_n$ such that

$$\begin{aligned} Du(t_{i+1/2}) = ku'(t_{i+1/2}) &+ \sum_{m=3}^{i} \beta_m k^m D_-^{m-1} Du(t_{i+1/2}) \\ &+ \sum_{m=\max\{3,i+1\}}^{n} \beta_m k^m D_+^{m-i} D_-^{i-1} Du(t_{i+1/2}) + O(k^{n+1}). \end{aligned}$$

Different forms of the last formula are possible.

5. Numerical test

This section deals with a comparison between the standard finite difference formulae and the new formulae obtained in Theorem 6 and 7. The comparisons address the numerical differentiation of the functions $u(x) = \sin(100\pi x)$ and $u(x) = \sin(1000\pi x)$ which are taken from the list of tests functions in [2]. For the classical finite difference formulae we just select the backward formulae of order 6 and 10, denoted $B6$ and $B10$, respectively. For the new finite difference formulae we choose the backward-centered formulae of order 6 and 10, denoted $BC6$ and $BC10$, respectively, and the interior-centered formulae of order 6 and 10, denoted $IC6$ and $IC10$, respectively. We drop the standard forward finite difference formula since it reaches the same accuracy as the backward formula (for a same order of approximation). The standard centered finite difference formula has the accuracy of the interior-centered formula so that we choose to not show it. Finally, the forward-centered formula reaches the same accuracy as the backward-centered formula.

Figure 1 shows that each of the finite difference formulae choosen gives a good approximate derivative of the functions considered. The accuracy of the approximations are related to both the order of accuracy of the corresponding formula and its error constant. Moreover, the new formulae are less prone to floating point error when the approximation reaches machine accuracy.

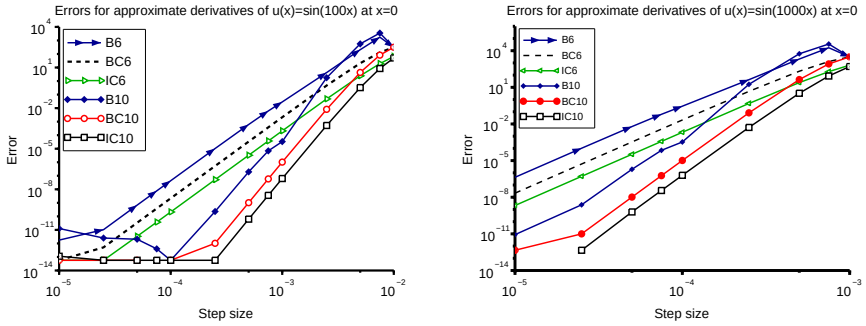

Figure 1: Graphs of absolute error for the numerical derivative of $u(x) = \sin(100\pi x)$ (left) and $u(x) = \sin(1000\pi x)$ (right) at $x = 0$ with $B6$, $B10$, $BC6$, $BC10$, $IC6$ and $IC10$.

[1] I. R. Khan, R. Ohba, Closed-form expressions for the finite difference approximations of first and higher derivatives based on Taylor series, J. Comput. Appl. Math. 107 (1999) 179–193.

[2] I. R. Khan, R. Ohba, New finite difference formulas for numerical differentiation, J. Comput. Appl. Math. 126 (2000) 269–276.

[3] I. R. Khan, R. Ohba, Taylor series based finite difference approximations of higher-degree derivatives, J. Comput. Appl. Math. 154 (2003) 115–124.

[4] A. Quarteroni, R. Sacco, F. Saleri, Numerical mathematics, 2nd Edition, Vol. 37, Springer-Verlag, Berlin, 2007.

[5] J. Li, General explicit difference formulas for numerical differentiation, J. Comput. Appl. Math. 183 (2005) 29–52.

[6] F. B. Hildebrand, Introduction to Numerical Analysis, McGraw-Hill Book Co., New York-Düsseldorf-Johannesburg, 1974.

[7] T. Chung, Computational Fluid Dynamics, 2nd Edition, Cambridge university press, 2010.

[8] G. Dahlquist, A. k. Björck, Numerical methods in scientific computing. Vol. I, SIAM, Philadelphia, PA, 2008.

[9] H. Bai, A. Xu, F. Cui, Representation for the Lagrangian numerical differentiation formula involving elementary symmetric functions, J. Comput. Appl. Math. 231 (2009) 907–913.

[10] V. Dubovsky, A. Yakhot, Finite-difference approximation for the $u^{(k)}$-derivative with $o(h^{M-k+1})$ accuracy: An analytical expression, Numer. Methods Partial Differential Equations 22 (2006) 1070–1079.

[11] F. J. Hickernell, S. Yang, Simplified analytical expressions for numerical differentiation via cycle index, J. Comput. Appl. Math. 224 (2009) 433–443.
[12] H. Hassan, A. Mohamad, G. Atteia, An algorithm for the finite difference approximation of derivatives with arbitrary degree and order of accuracy, J. Comput. Appl. Math. 236 (2012) 2622–2631.
[13] S.-C. R. Koyaguerebo-Imé, Y. Bourgault, Arbitrary order A-stable methods for ordinary differential equations via deferred correction, Submitted to BIT. (2020).
[14] S.-C. R. Koyaguerebo-Imé, Y. Bourgault, Arbitrary high-order unconditionally stable methods for reaction-diffusion equations via deferred correction: Case of the implicit midpoint rule, Submitted to ESAIM Math. Model. Numer. Anal.[arXiv:2006.02962] (2020).
[15] B. Gustafsson, W. Kress, Deferred correction methods for initial value problems, BIT 41 (2001) 986–995.
[16] W. Kress, B. Gustafsson, Deferred correction methods for initial boundary value problems, J. Sci Comput. 17 (2002) 241–251.
[17] J. W. Daniel, V. Pereyra, L. L. Schumaker, Iterated deferred corrections for initial value problems, Acta Cient. Venezolana 19 (1968) 128–135.

Chapter 2

Arbitrary order A-stable methods for ordinary differential equations via deferred correction

This chapter is presented in terms of a journal article and submitted to BIT Numerical Mathematics (BIT), carrying the same title as mentioned above. Please see the attached paper for the content.

BIT manuscript No.
(will be inserted by the editor)

Arbitrary order A-stable methods for ordinary differential equations via deferred correction

Saint-Cyr E.R. Koyaguerebo-Imé,
Yves Bourgault

Received: date / Accepted: date

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

Keywords Ordinary differential equations · high order time-stepping methods · deferred correction · A-stability

Mathematics Subject Classification (2000) MSC 65B05 · 65L04 · 65L05 · 65L12 · 65L20

1 Introduction

In [9, 18], Gustafsson and Kress introduced a new version of deferred correction strategy for the numerical solution of linear systems of ODEs [9] and initial boundary value problems [18], under a monotonicity condition. Numerical experiments with one-dimensional linear parabolic and hyperbolic equations were performed and showed that the method is effective (orders 2, 4 and 6 of accuracy are achieved). We propose to extend the method from [9,18] to the time-discretization of more general time-evolution partial differential equations (PDEs). In this paper, we restrict to the

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

Saint-Cyr E.R. Koyaguerebo-Imé · Yves Bourgault

case of the initial value problem (IVP)

$$\begin{cases} \dfrac{du}{dt} = F(t,u), \quad t \in [0,T], \\ u(0) = \quad u_0, \end{cases} \tag{1}$$

where the unknown u is from $[0,T]$ into a Banach space X, u_0 is a given data and F is a sufficiently differentiable function such that u exists and is sufficiently differentiable. The main objective is to show the properties of the numerical method (consistency, stability, convergence and order of accuracy). A complete analysis of the DC method applied to reaction-diffusion equations leads to an arbitrary high order and unconditionally stable method (see [17]).

The deferred correction (DC) method is used to improve the order of accuracy of numerical methods of lower order. This method is explored by many authors, e.g. [1, 2, 6, 7, 9, 11, 19, 21]. The method in [6] is an application of iterative deferred correction (IDC). The authors proved that an asymptotic improvement of order p can be accomplished, from a scheme of order p, at each step of the IDC procedure, provided suitable finite difference operators are employed. Numerical experiments are performed with the IDC applied to the trapezoidal rule, Taylor-2 and Adams-Bashforth of order 2. The results are promising even though they point out some difficulties of the proposed algorithms: inacuracy for "large" time step and no asymptotic improvement for high levels of correction. The approaches in [1,2,7,9,11,19] are quite similar and consist in a linear perturbation of a low order scheme. However, solving stiff problems (problems extremely hard to solve by standard explicit step-by-step methods [23]) is a challenge unfavorable for these methods. In particular, the method in [19], concerning a highly accurate solver for stiff ODEs, requires sufficiently small time steps for moderately stiff problems while convergence is reduced to order 2 for "very stiff" problems.

Our schemes are based on nonlinear perturbations (corrections) of the implicit midpoint rule and inherit the A-stable property of the trapezoidal rule [5] at any stage of the correction. Starting from an approximation $\left\{u^{2,n}\right\}_{n=0}^{N}$ of the exact solution u by the implicit midpoint rule on a uniform partition $0 = t_0 < t_1 < \cdots < t_N = T$ of $[0,T]$, at the stage $j = 1, 2, \cdots$ of the correction we obtain an approximation $\left\{u^{2j+2,n}\right\}_{n=0}^{N}$ of u, expected to be of order $2j+2$ of accuracy, on the same partition. Each approximate solution $\left\{u^{2j,n}\right\}_{n=0}^{N}$ to be corrected is subject to a deferred correction condition (DCC) which guarantees the improvement of the order of accuracy. We prove that if $\left\{u^{2j,n}\right\}_{n=0}^{N}$ satisfies the DCC and its correction $\left\{u^{2j+2,n}\right\}_{n=0}^{N}$ converges to u at the discrete points $0 = t_0 < t_1 < \cdots < t_N = T$ (or is simply bounded, when X is finite dimensional) then $\left\{u^{2j+2,n}\right\}_{n=0}^{N}$ approximates u with order $2j+2$. Moreover, provided the function F is Lipschitz with respect to its second variable or satisfies a one-sided Lipschitz condition, each $\left\{u^{2j,n}\right\}_{n=0}^{N}$ satisfies the DCC and then converges with order $2j$ of accuracy, for arbitrary positive integer j. The theory is illustrated by numerical tests, for the schemes of order 2, 4, ..., 10.

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

2 Deferred correction schemes for the implicit midpoint rule

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

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

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

and

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

The average operator is denoted by E:

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

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

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

and

$$D_-(D_+D_-)^m u(t_n) = k^{-2m-1}\sum_{i=0}^{2m+1}(-1)^i\binom{2m+1}{i}u(t_{n+m-i}), \tag{3}$$

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

$$\left\|D_+^{m_1}D_-^{m_2}u(t_n)\right\| \leq \max_{0\leq t\leq T}\left\|\frac{d^{m_1+m_2}u}{dt^{m_1+m_2}}(t)\right\|, \tag{4}$$

provided $[t_{n-m_2}, t_{n+m_1}] \subset [0,T]$ and $m_1+m_2 \leq 2p+3$ (see [14, p.249] or [16]).

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

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

and

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

From the centered finite difference approximation (see [16, Thm 5] or [3, 4, 12]) we have

$$\frac{du}{dt}(t_{n+1/2}) = \frac{u(t_{n+1})-u(t_n)}{k} - \sum_{i=1}^{j}c_{2i+1}k^{2i}(D_+D_-)^iDu(t_{n+1/2})) + O(k^{2j+2}) \tag{5}$$

and

$$u(t_{n+1/2}) = \frac{u(t_{n+1}) + u(t_n)}{2} - \sum_{i=1}^{j} c_{2i} k^{2i} (D_+ D_-)^i E u(t_{n+1/2}) + O(k^{2j+2}), \quad (6)$$

for each integer $1 \le j \le p$. These approximations lead to the schemes

$$\begin{aligned} &\frac{u^{n+1} - u^n}{k} - \sum_{i=1}^{j} c_{2i+1} k^{2i} (D_+ D_-)^i D u^{n+1/2} \\ &\qquad = F\left(t_{n+1/2}, \frac{u^{n+1} + u^n}{2} - \sum_{i=1}^{j} c_{2i} k^{2i} (D_+ D_-)^i E u^{n+1/2} \right). \end{aligned} \quad (7)$$

The schemes (7) are multi-steps and prone to stability restrictions. We resort to DC method to transform them into a sequence of one step schemes as follows: For $j = 0$, we have the implicit midpoint rule

$$\frac{u^{2,n+1} - u^{2,n}}{k} = F\left(t_{n+1/2}, \frac{u^{2,n+1} + u^{2,n}}{2} \right), \quad u^{2,0} = u_0. \quad (8)$$

For $j \ge 1$,

$$\begin{aligned} &\frac{u^{2j+2,n+1} - u^{2j+2,n}}{k} - \sum_{i=1}^{j} c_{2i+1} k^{2i} (D_+ D_-)^i D u^{2j,n+1/2} \\ &= F\left(t_{n+1/2}, \frac{u^{2j+2,n+1} + u^{2j+2,n}}{2} - \sum_{i=1}^{j} c_{2i} k^{2i} (D_+ D_-)^i E u^{2j,n+1/2} \right), \end{aligned} \quad (9)$$

$$u^{2j+2,0} = u_0. \quad (10)$$

The scheme (9)-(10) has unknowns $u^{2j+2,n}$, $n = 1, 2, ..., N$, and is deduced from (7) by substituting the unknown u^n under the summation symbols by $u^{2j,n}$. The index $2j$ indicates that $\left\{ u^{2j,n} \right\}_n$ is expected to be an approximation of the exact solution u with order $2j$ of accuracy. We call the schemes (9)-(10) Deferred Correction of order $2j+2$ for the implicit midpoint rule, denoted DC(2j+2).

Remark 1 The scheme (9)-(10), for $n = 1, 2, 3, \cdots, j$, should involve unknowns $u^{2j,-1}, ..., u^{2j,-j}$ which represent approximate solutions of (1) at the discrete points $t = -k, ..., -jk$, respectively. To avoid those approximations for $t < 0$, we propose the following scheme which is efficient for the computation of $u^{2j+2,1}, ..., u^{2j+2,j}$, using only points within the solution interval $[0, T]$.

$$\begin{aligned} &\frac{u^{2j+2,n+1} - u^{2j+2,n}}{k} - k^{-1} \sum_{i=1}^{j} c_{2i+1}^j k_j^{2i+1} (D_+ D_-)^i D \bar{u}^{2j,(2j+1)n+j+1/2} \\ &= F\left(t_{n+1/2}, E u^{2j+2,n+1/2} - \sum_{i=1}^{j} c_{2i}^j k_j^{2i} (D_+ D_-)^i E \bar{u}^{2j,(2j+1)n+j+1/2} \right), \end{aligned} \quad (11)$$

$$u^{2j+2,0} = u_0. \quad (12)$$

The finite difference operator in (11) are related to the time step $k_j = k/(2j+1)$. The approximations $\left\{\overline{u}^{2j,m}\right\}_m$ and $\left\{u^{2j,n}\right\}_n$ are computed from the same scheme, (8) or (9)-(10), but for the time steps k_j and k, respectively. The scheme (11) results from the finite difference approximations

$$u'(t_{n+1/2}) = \frac{u(t_{n+1}) - u(t_n)}{k} - \frac{1}{k}\sum_{i=1}^{j} c_{2i+1}^j k_j^{2i+1} D(D_+D_-)^i u(\tau_{j+1/2}) + O(k_j^{2j+2}) \tag{13}$$

and

$$u(t_{n+1/2}) = \frac{u(t_{n+1}) + u(t_n)}{2} - \sum_{i=1}^{j} c_{2i}^j k_j^{2i} (D_+D_-)^i E u(\tau_{j+1/2}) + O(k_j^{2j+2}), \tag{14}$$

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

Table 1 Coefficients of the approximations (13)-(14) for $j = 1, 2, 3, 4$

j	c_2^j	c_3^j	c_4^j	c_5^j	c_6^j	c_7^j	c_8^j	c_9^j
1	$\frac{9}{8}$	$\frac{9}{8}$						
2	$\frac{25}{8}$	$\frac{125}{24}$	$\frac{125}{128}$	$\frac{125}{128}$				
3	$\frac{49}{8}$	$\frac{343}{24}$	$\frac{637}{128}$	$\frac{13377}{1920}$	$\frac{1029}{1024}$	$\frac{1029}{1024}$		
4	$\frac{81}{8}$	$\frac{243}{8}$	$\frac{1917}{128}$	$\frac{17253}{640}$	$\frac{7173}{1024}$	$\frac{64557}{7168}$	$\frac{32733}{32768}$	$\frac{32733}{32768}$

Remark 2 Each $u^{2j+2,n+1}$, $n \geq j$, is an iterative solution of the system

$$x - a_n^j - kF(t_{n+1/2}, 0.5x + b_n^j) = 0, \tag{15}$$

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

3 Deferred correction condition (DCC)

In this section we give a sufficient condition for the scheme (9)-(10) to achieve order $2j+2$ of accuracy. Hereafter, the letter C will denote any constant independent from k, and that can be calculated explicitly in term of known quantities. The exact value of C may change from a line to another line. We have the following definition:

Definition 1 Let u be the exact solution of the Cauchy problem (1). Given a positive integer j, a sequence $\left\{u^{2j,n}\right\}_{n=0}^N$ of approximations of u at the discrete points $0 = t_0 < \cdots < t_N = T$ is said to satisfy the Deferred Correction Condition (DCC) for

the implicit midpoint rule if $\left\{u^{2j,n}\right\}_{n=0}^{N}$ approximates u with order $2j$ of accuracy, and we have

$$\|(D_+D_-)D(u^{2j,n+1/2}-u(t_{n+1/2}))\|+\|D_+D_-(u^{2j,n+1}-u(t_{n+1}))\|\leq Ck^{2j}, \quad (16)$$

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

Remark 3 Condition (16) is equivalent to

$$\left\|\sum_{i=1}^{j}c_{2i}k^{2i}\left(D_+D_-\right)^i\left(u^{2j,n}-u(t_n)\right)\right\|\leq Ck^{2j+2}, \quad (17)$$

and

$$\left\|\sum_{i=1}^{j}(c_{2i+1}-c_{2i})k^{2i}\left(D_+D_-\right)^i D\left(u^{2j,n+1/2}-u(t_{n+1/2})\right)\right\|\leq Ck^{2j+2}, \quad (18)$$

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

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

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

We have the following result:

Theorem 1 *Let u be the exact solution of (1) and $\left\{u^{2j,n}\right\}_{n=0}^{N}$, $1\leq j\leq p$, a sequence of approximations of u satisfying DCC for the implicit midpoint rule. Let $\left\{u^{2j+2,n}\right\}_{n=0}^{N}$ be the solution of (9)-(10) built from $\left\{u^{2j,n}\right\}_{n=0}^{N}$. We suppose that $u^{2j+2,1},...,u^{2j+2,j}$ are given and satisfy*

$$\|u^{2j+2,n}-u(t_n)\|\leq Ck^{2j+2}, \quad \text{for} \quad n=1,2,...,j, \quad (19)$$

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

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

$$\|F(t,x)-F(t,y)\|\leq\mu\|x-y\|, \quad \forall(t,x,y)\in[0,T]\times X\times X. \quad (20)$$

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

$$\|u^{2j+2,n}-u(t_n)\|\leq M, \quad \text{for each} \quad n=0,1,...,N. \quad (21)$$

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

(iv) X is a Hilbert space with inner product $(.,.)$, and F satisfied the following so-called one-sided Lipschitz condition, with a constant $\mu \geq 0$,

$$(F(t,x)-F(t,y),x-y) \leq \mu\|x-y\|^2, \quad \forall (t,x,y) \in [0,T]\times X\times X. \tag{22}$$

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

$$\|u^{2j+2,n}-u(t_n)\| \leq Ck^{2j+2}, \quad \text{for each} \quad n=0,1,...,N, \tag{23}$$

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

Proof

1. First we consider the case where the function $F=F(t,x)$ is Lipschitz with respect to the second variable x. Combining (1) and (9), we obtain the identity

$$\begin{aligned} D\Theta^{2j+2,n+1/2} &= \sigma^{2j+2,n+1/2}+(\Lambda^j-\Gamma^j)D\left(u^{2j,n+1/2}-u(t_{n+1/2})\right) \\ &+F\left(t_{n+1/2},\widehat{u}^{2j+2,n+1}-\Gamma^j\widehat{u}^{2j,n+1}\right)-F\left(t_{n+1/2},\widehat{u}(t_{n+1})-\Gamma^j\widehat{u}(t_{n+1})\right), \end{aligned} \tag{24}$$

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

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

and

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

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

$$\Theta^{2j+2,n}=\left(u^{2j+2,n}-u(t_n)\right)-\Gamma^j\left(u^{2j,n}-u(t_n)\right), \tag{25}$$

and

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

From (5) we have

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

and, since F is differentiable and u is sufficiently regular, we deduce from the mean value theorem and the approximation (6) that

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

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

$$\left\|\sigma^{2j+2,n+1/2}\right\| \leq Ck^{2j+2}. \tag{26}$$

Since the sequence $\left\{u^{2j,n}\right\}_n$ satisfies DCC, from Remark 3 we have

$$\left\|\left(\Lambda^j-\Gamma^j\right)D\left(u^{2j,n+1/2}-u(t_{n+1/2})\right)\right\|\leq Ck^{2j+2}. \tag{27}$$

From the Lipschitz condition on F we have

$$\begin{aligned}&\left\|F\left(t_{n+1/2},\widehat{u}^{2j+2,n+1}-\Gamma^j\widehat{u}^{2j,n+1}\right)-F\left(t_{n+1/2},\widehat{u}(t_{n+1})-\Gamma^j\widehat{u}(t_{n+1})\right)\right\|\\&\leq\mu\|\widehat{\Theta}^{2j+2,n+1}\|.\end{aligned} \tag{28}$$

Substituting inequalities (26)-(28) in the identity (24), we deduce that

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

and it follows from the triangle inequality that

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

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

$$\|\Theta^{2j+2,n}\|\leq C\frac{1}{2-\mu k}\left(\frac{2+\mu k}{2-\mu k}\right)^{n-j-1}k^{2j+2}+\left(\frac{2+\mu k}{2-\mu k}\right)^{n-j}\|\Theta^{2j+2,j}\|. \tag{29}$$

From hypothesis (19) and the DCC we have

$$\|\Theta^{2j+2,j}\|\leq\|u^{2j+2,j}-u(t_j)\|+\left\|\Gamma^j(u^{2j,j}-u(t_j))\right\|\leq Ck^{2j+2}, \tag{30}$$

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

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

Finally, by the triangle inequality, identity (25) and DCC, we have

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

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

2. Suppose that $\left\{u^{2j+2,n}\right\}_{n=0}^N$ satisfies (21) and X is finite dimensional. We can write

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

From (21) and the DCC there exists $k_1>0$ such that $0<k\leq k_1\leq k_0$ implies

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

On the other hand, we have

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

where

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

It follows (28) for

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

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

3. If $\left\{u^{2j+2,n}\right\}_n$ converges to the exact solution u, taking the DDC and the finite difference formula (6) into account, we have

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

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

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

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

4. Here we consider the case where X is a Hilbert space and F satisfies the monotonicity condition (22). Then, taking the inner product of the identity (24) with $\widehat{\Theta}^{2j+2,n+1}$, we deduce the inequality

$$\begin{aligned} \left(D\Theta^{2j+2,n+1/2}, \widehat{\Theta}^{2j+2,n+1}\right) \leq \left(\sigma^{2j+2,n+1/2}, \widehat{\Theta}^{2j+2,n+1}\right) + \mu \|\widehat{\Theta}^{2j+2,n+1}\|^2 \\ \left((\Lambda^j - \Gamma^j) D(u^{2j,n+1/2} - u(t_{n+1/2})), \widehat{\Theta}^{2j+2,n+1}\right) \end{aligned} \tag{31}$$

since, according to (22), we have

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

Inequalities (26)-(27) together with the Cauchy-Schwartz inequality yield

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

and

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

where C is a constant depending only on j, T, a Lipschitz constant on F and the derivatives of u up to order $2j+3$. Substituting the last three inequalities into (31), we obtain

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

and we deduce from the identity

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

and the inequality

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

that

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

for $0\leq\mu k<2$. The conclusion follows from the case (i).

Remark 4 Theorem 1 shows that the correction may be applied for any other scheme satisfying DCC.

Remark 5 In practice, the estimate (23) takes the form

$$\|u^{2j+2,n}-u(t_n)\|\leq C\left(\frac{2+\mu k}{2-\mu k}\right)^{n-j-1}k^{2j+2},\tag{32}$$

where $\mu\simeq\max\limits_{0\leq t\leq T}\|d_uF(t,u(t))\|$. This inequality requires $0\leq\mu k<2$. If $\mu k>2$ the estimate does not hold, but the methods may produce accurate solutions which are prone to oscillations around the exact solution (this is the case when the eigenvalues of the Jacobian $d_uF(t,u(t))$ along the solution curve have negative real part).

4 Convergence and order of accuracy

In this section we prove the following theorem:

Theorem 2 *Let $u\in C^{2p+3}([0,T],X)$ be the exact solution of the problem (1). Suppose that one of the four conditions (i)-(iv) of Theorem 1 holds, with condition (ii) or (iii) holding for all $j=0,1,\cdots,p+1$. Then each sequence $\left\{u^{2j,n}\right\}_{n=0}^{N}$, $1\leq j\leq p+1$, solution of the scheme (8) or (9)-(10), approximates u with order $2j$ of accuracy. Furthermore, we have the estimate*

$$\|(D_+D_-)^mD(u^{2j,n+1/2}-u(t_{n+1/2}))\|+\|(D_+D_-)^m(u^{2j,n+1}-u(t_{n+1}))\|\leq Ck^{2j}\tag{33}$$

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

To prove this theorem we need Theorem 1 and the the following lemma:

Lemma 1 *Let $\left\{u^{2,n}\right\}_{n=0}^{N}$ be the solution of the scheme (8). Suppose that one of the conditions (i), (iii) or (iv) of Theorem 1 holds, or $\left\{u^{2,n}\right\}_{n=0}^{N}$ is bounded in the sense of the condition (ii) of this theorem. Then $\left\{u^{2,n}\right\}_{n=0}^{N}$ approximates u with order 2 of accuracy, and we have the inequality*

$$\|(D_+D_-)^m D(u^{2,n+1/2}-u(t_{n+1/2}))\|+\|(D_+D_-)^m(u^{2,n+1}-u(t_{n+1}))\|\leq Ck^2, \tag{34}$$

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

Proof (Proof of Lemma 1) For the sake of simplification we suppose that $F=F(x)$. The general case can be handled by transforming (1) into an autonomous system. From the hypotheses of the Lemma, Theorem 1 implies that $\left\{u^{2,n}\right\}_{n=0}^{N}$ approximates u with order two of accuracy:

$$\|u(t_n)-u^{2,n}\|\leq Ck^2, \text{ for each } n=0,1,2,\cdots,N, \tag{35}$$

where C is a constant depending only on T, F and the derivatives of u up to order 3. To establish (34) we proceed by induction on the integer $m=0,1,...,p$.

1. Inequality (34) for $m=0$.

As in Theorem 1, we combine (1) and (8) and deduce the identity

$$D\Theta^{2,n+1/2}=\left[F\left(\widehat{u}^{2,n+1}\right)-F\left(\widehat{u}(t_{n+1})\right)\right]+\sigma^{2,n+1/2}, \tag{36}$$

where

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

and

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

From Taylor's formula with integral remainder and the estimate (4), there exists a function g such that

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

with

$$\|D_+^{m_1}D_-^{m_2}g(t_{n+1})\|\leq C, \quad \text{for} \quad m_2-1\leq n\leq N-m_1-1, \tag{37}$$

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

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

where

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

The last identities substituted into (36) yield

$$D\Theta^{2,n+1/2}=\int_0^1 dF\left(K_1^{n+1}\right)\left(\widehat{\Theta}^{2,n+1}\right)d\tau_1+k^2g(t_{n+1}). \tag{38}$$

Proceeding as in Theorem 1, we deduce from (35) and the regularity of u that

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

Therefore, taking the norm on both sides of (38), we deduce by the triangle inequality and the inequalities (35) and (37), for $m_1 = m_2 = 0$, that

$$\|D\Theta^{2,n+1/2}\| \leq C\|\widehat{\Theta}^{2,n+1}\| + k^2\|g(t_{n+1})\| \leq Ck^2, \tag{39}$$

where C is a constant depending only on T and the derivatives of u and F up to order 3 and 1, respectively. The last inequality combined with (35) implies that (34) holds for $m = 0$.

2. Here we are going to prove that inequality (34) remains true for $m+1$, assuming that it holds for an arbitrary integer m such that $0 \leq m \leq p-1$.

We apply $(D_+D_-)^m D_+$ to (38) and obtain

$$(D_+D_-)^{m+1}\Theta^{2,n+1} = (D_+D_-)^m D_+ h(t_{n+1}) + k^2 (D_+D_-)^m D_+ g(t_{n+1}), \tag{40}$$

where we set

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

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

$$D_+h(t_n) = \int_0^1 dF\left(K_1^{n+1}\right)(D_+\widehat{\Theta}^{2,n}) d\tau_1 + \int_0^1\int_0^1 d^2F\left(K_2^n\right)\left(D_+K_1^n, \widehat{\Theta}^{2,n}\right) d\tau_1 d\tau_2,$$

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

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

$$K_{i+1}^n = K_i^n + \tau_{i+1}(K_i^{n+1} - K_i^n) = K_1^n + \sum_{l=1}^{i} \sum_{2\leq i_1<\cdots<i_l\leq i+1} \tau_{i_1}\cdots\tau_{i_l} k^l D_+^l K_1^n. \tag{41}$$

It follows the general formula

$$D_+^q h(t_n) = \sum_{i=1}^{q+1} \sum_{|\alpha_i|=q} L_{i,\alpha_i}^{n,q}, \text{ for } q = 1, 2, ..., 2p+1, \text{ and } n \leq N-q, \tag{42}$$

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

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

where $\beta_i=(\beta_i^1,\cdots,\beta_i^{i-1},\beta_i^i)\in\{1,2,\cdots,q\}^{i-1}\times\{0,1,\cdots,q-i+1\}$ with $\beta_i^l+\alpha_i^l\leq q-l+1$, for $l=1,\cdots,i$. From (41) and (35) we have

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

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

$$\left\|d^iF\left(K_i^n\right)\right\|\leq C_i,\text{ for }i=1,2,...,2p+2,\text{ and }0\leq n\leq N-i+1, \tag{43}$$

where C_i is a constant depending only on k_3, T, and the derivatives of u and F up to order 3 and i, respectively. From the inductions hypothesis (34) and inequality (4) we have

$$\|D_+^rK_i^n\|\leq C,\text{ for }1\leq r\leq i\leq 2m+3,1\leq n\leq N-i-r+1, \tag{44}$$

and

$$\|D_+^r\widehat{\Theta}^{2,n}\|\leq Ck^2,\text{ for }1\leq r\leq 2m+1,1\leq n\leq N-r, \tag{45}$$

where C is a constant depending only on m, T, and the derivatives of u and F up to order $r+2$ and r, respectively. Each $L_{i,\alpha_i,\beta_i}^{n,q}$ being multilinear continuous, we deduce from (43)-(45) and the relation $\beta_i^l+\alpha_i^l\leq q-l+1$, for $l=1,\cdots,i$, that

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

It follows by the triangle inequality that (42) for $q=2m+1$ yields

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

for $n=m,m+1,\cdots,N-(m+1)-1$, where C is a constant depending only on p, T, and the derivatives of u and F up to order $2m+4$ and $2m+2$, respectively . Passing to the norm in identity (40), we deduce from (37) and the last inequality that

$$\left\|\left(D_+D_-\right)^{m+1}\Theta^{2,n+1}\right\|\leq Ck^2. \tag{46}$$

Otherwise, applying D_- to (40), inequalities (43)-(45) and (46) yield

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

for $n=m,m+1,\cdots,N-(m+1)-1$, where C is a constant depending only on p, T, and the derivatives of u and F up to order $2m+5$ and $2m+3$, respectively. Therefore, passing to the norm in the identity obtained by applying D_- to (40), we deduce from (40) and the last inequality that

$$\left\|D_-\left(D_+D_-\right)^{m+1}\Theta^{2,n+1}\right\|\leq Ck^2, \tag{47}$$

for $n=m,m+1,\cdots,N-(m+1)-1$, with the constant C depending only on p, T, and the derivatives of u and F up to order $2m+5$ and $2m+3$, respectively. Inequalities (46) and (47) imply that the induction hypothesis is also true for $m+1$, and we deduce that (34) is true for each integer $m=0,1,...,p$.

Proof (Proof of Theorem 2) We proceed by induction on $j = 1, 2, ..., p+1$. The case $j = 1$ is immediate from Lemma 1. Suppose that $\left\{u^{2j,n}\right\}_n^N$ approximates u with order $2j$ of accuracy and satisfies (33), for an arbitrary j such that $j \leq p$. We are going to prove that $\left\{u^{2j+2,n}\right\}_n^N$ approximates u with order $2j+2$ of accuracy and (33) holds substituting j by $j+1$.

From the induction hypothesis, $\left\{u^{2j,n}\right\}_n$ satisfies DCC. Because $\left\{u^{2j,n}\right\}_n$ and $\left\{\overline{u}^{2j,m}\right\}_m$ are computed from the same scheme DC2j, but for different time steps, $\left\{\overline{u}^{2j,m}\right\}_m$ also satisfies DCC. Therefore, as in 29, Theorem 1 applied to the approximation $\left\{u^{2j+2,n}\right\}_{n=0}^j$, built from $\left\{\overline{u}^{2j,m}\right\}_m$, yields

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

where

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

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

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

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

$$\|u^{2j+2,n} - u(t_n)\| \leq Ck^{2j+2}, \text{ for } n = 0, 1, \cdots, j. \tag{48}$$

From the DCC on $\left\{u^{2j,n}\right\}_n$ and the inequality (48), Theorem 1 again implies that $\left\{u^{2j+2,n}\right\}_{n=0}^N$ approximates the exact solution u with order $2j+2$ of accuracy. Therefore, it is enough to establish (33) for $j+1$, $j \leq p$. To this end we rewrite identity (24) as follows

$$D\Theta^{2j+2,n+1/2} = H(t_{n+1}) + \sigma^{2j+2,n+1/2} + (\Lambda^j - \Gamma^j)D(u^{2j,n+1/2} - u(t_{n+1/2})), \tag{49}$$

with

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

where $\Theta^{2j+2,n}$ and $\sigma^{2j+2,n+1/2}$ are as in Theorem 1. Proceeding as in Lemma 1 and taking the finite difference formulae (5) and (6) into account, we can write

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

where

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

C is a constant depending only on p, T, and the derivatives of u and F. According to the inequality (33) from the induction hypothesis, we may write

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

where

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

Therefore, writing (49) as follows

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

with

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

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

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

for $m = 0, 1, ..., p - j$ and $n = m + j - 1, m + j, ..., N - j - m$, where C is a constant depending only on p, T, and the derivatives of u and F up to order $2(m + j + 1) + 1$ and $2(m + j) + 1$, respectively. Inequality (33) holds for $\left\{u^{2j+2,n}\right\}_n$ by the triangle inequality from the last inequality.

5 Absolute stability

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

$$\begin{cases} u' &= \lambda u \\ u(0) &= 1, \end{cases} \tag{50}$$

where λ is a complex number, we have the following definition (see [5, 20]):

Definition 2 A numerical method is said to be absolutely stable if the corresponding solution for the problem (50) for fixed $k > 0$ and some $Re(\lambda) < 0$ is such that

$$\lim_{n \to +\infty} |u^n| = 0. \tag{51}$$

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

$$\mathcal{A} = \{ z = \lambda k \in \mathbb{C} : (51) \text{ is satisfied } \}. \tag{52}$$

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

Before establishing absolute stability results for the deferred correction schemes (8) and (9)-(10), we recall the following result.

Lemma 2 *Let P_m be a polynomial of degree m in one variable. Then the sum $\sum_{i=0}^{n} P_m(i)$ is a polynomial of degree $m + 1$ in the variable n.*

Proof Without loss of generality we assume that $P_m(x) = x^m$ and set $F_m(n) = \sum_{p=1}^{n} p^m$. It is then enough to prove that $F_m(n)$ is a polynomial of degree $m+1$ in the variable n, for each non-negative integer m. We proceed by induction on m. The cases $m = 0, 1$ are trivial. Assume that $F_m(n)$ is a polynomial of degree $m+1$ in n, for arbitrary positive integer m. We have the identities

$$\begin{aligned}(n+1)^{m+2} - 1 &= \sum_{q=1}^{n} [(q+1)^{m+2} - q^{m+2}] = \sum_{q=1}^{n} \sum_{i=0}^{m+1} \binom{m+2}{i} q^i \\ &= \sum_{i=0}^{m+1} \binom{m+2}{i} F_i(n)\end{aligned}$$

which implies that

$$F_{m+1}(n) = \frac{1}{m+2}(n+1)^{m+2} - \frac{1}{m+2} \sum_{i=0}^{m} \binom{m+2}{i} F_i(n) - \frac{1}{m+2}.$$

According to the induction hypothesis, $\sum_{i=0}^{m} \binom{m+2}{i} F_i(n)$ is a polynomial of degree $m+1$ in the variable n. Therefore, the last identity implies that $F_{m+1}(n)$ is a polynomial of degree $m+2$ with respect to n, and we can deduce by induction that each $F_m(n)$ is a polynomial of degree $m+1$ in the variable n, for each non-negative integer m.

Lemma 3 *Suppose that $F(t,u) = \lambda u$ and $u_0 = 1$ in the initial value problem (1), where λ is a complex number with negative real part ($\lambda \in \mathbb{C}_-$). Then the corresponding approximate solutions from the schemes (8) and (9)-(10) can be written as follows*

$$u^{2j+2,n} = \left(\frac{2+\lambda k}{2-\lambda k}\right)^{n-j} P_j(n), \text{ for } j = 0, 1, 2, ..., \text{ and } n \geq j, \tag{53}$$

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

Proof We suppose that $\lambda k \neq -2$, otherwise we trivially have $u^{2j,n+1} = 0$, for $n \geq j$. Since $F(t,u) = \lambda u$, we can rewrite (9) as follows

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

where, according to formulae (2) and (3), we have

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

and

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

Combining the last three identities, we deduce that

$$u^{2j+2,n+1} = \frac{2+\lambda k}{2-\lambda k} u^{2j+2,n} + \frac{2}{2-\lambda k} \sum_{i=0}^{2j+1} \alpha_{j,i}(\lambda k) u^{2j,n+1+j-i}, \text{ for } n \geq j \geq 1, \quad (54)$$

where $\alpha_{j,i}$ is affine in λk. Under the hypothesis of the lemma, (8) matches the trapezoidal rule, and we have

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

that is (53) is true for $j = 0$. Suppose that (53) holds for an arbitrary integer $j \geq 0$. From (54) we have

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

with $n \geq j+2$, and, substituting each $u^{2j+2,n+1+j-i}$ by the formula given by the induction hypothesis (53), we deduce that

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

where

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

It follows that

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

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

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

where

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

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

Theorem 3 *Each of the deferred correction schemes (8) and (9)-(10) is A-stable.*

Proof From Lemma 3 we have, for $Re(\lambda k) < 0$,

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

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

6 Numerical experiments

In this section we evaluate the accuracy and order of convergence of the schemes $DC2, DC4, \cdots, DC10$, implemented using the Scilab programming language. The starting values are computed using the scheme (11)-(12). We choose five standard problems for the evaluation: the first problem concerns the effect of high order derivatives, the second is about long term integration, and the three others are about stiffness. For a comparison of accuracy we implement in Scilab the backward differentiation formulae (BDF) of order 2, 4 and 6 and the explicit Runge-Kutta (RK) of order 4, in the case of problems having analytic solutions. The implemented BDF are then run with exact starting values. For the problems without analytic solutions we use the functions `stiff` (implementing BDF with adaptive steps) and `rkf` (Runge-Kutta 4-5) of the solver `ode` from Scilab. For each of the problems, except the first one, we give a table of absolute error and order of accuracy for pairs of two consecutive time steps, for the approximate solutions with the DC methods, and we present the optimal absolute error for the solvers `stiff` and `rkf`. For the problems (59) and (60) that do not have an analytic solution, we consider a small time step such that the approximate solutions with $DC6, \cdots, DC10$ are almost identical (to machine precision for problem (59)), and we choose one of the approximate solutions as reference solution. We denote by k_{max} the maximal time step allowed to compute an optimal approximate solution with the solver `stiff` or `rkf`.

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

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

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

6.1 Bernoulli differential equation

$$u'(t) = F(t,u) = -0.1u(t) - 1000u^{20}(t) \ , \quad u(0) = 1, \ \ t \in [0,10]. \tag{55}$$

Figure 1 shows the graph of the absolute error with $DC2, \cdots, DC10$, BDF2, BDF4, BDF6 and RK4.

6.2 Oscillatory problem [13]

$$u' = \lambda u \cos(t) \ , \quad u(0) = 1, \ \ T = 10^6, \lambda = 10. \tag{56}$$

The exact solution is $u(t) = e^{\lambda \sin(t)}$. The original problem is set with $\lambda = 1$ in [13]. The author in [15] solved this problem with Runge-Kutta methods of orders 4 and 8, for $\lambda = 2$ and $T = 2580\pi$, to " illustrate the need of higher order methods when a long-term integration problem is considered". Table 2 gives the absolute error and the order of convergence for each pair of consecutive time steps. The maximal absolute errors with the implemented BDF 2, 4 and 6 for the time step $k = 1.5625 \times 10^{-3}$ are respectively 22026.46, 14836.76 and 5578.40. The solvers `rkf` and `stiff`, used with $k_{max} = 0.1$ and tolerances $rtol = 100 \times atol = 10^{-10}$, give the absolute errors 22026.46 and 2636.00, respectively.

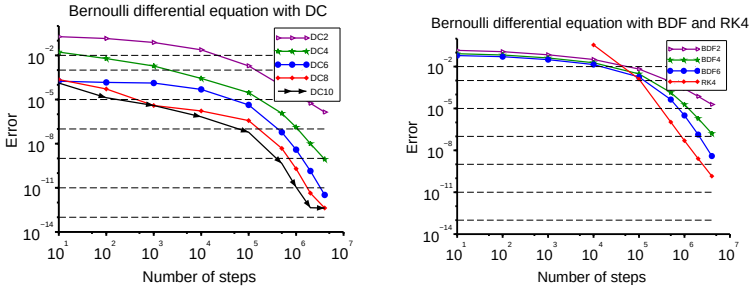

Fig. 1 Graphs of the maximal absolute error for the Bernoulli differential equation with $DC2, \cdots, DC10$ at left and BDF2, BDF4, BDF6 and RK4 at right .

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

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

6.3 Problem B5 modified [8], stiff with complex eigenvalues of negative real parts

$$y' = \begin{bmatrix} -10 & \alpha & 0 & 0 & 0 & 0 \\ -\alpha & -10 & 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.1 \end{bmatrix} y,\ y(0) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},\ \alpha = 5000,\ T = 20. \qquad (57)$$

This problem, originally set with $\alpha = 100$, is an illustration of ODEs resulting from a semidiscretization by finite element methods of parabolic PDEs [24]. We choose $\alpha = 5000$ to make the problem a little more difficult. Table 3 gives the errors for the first component of the approximate solutions which is similar for the second component. The errors for the others components quickly achieve machine precision. The maximal absolute errors with the implementing BDF 2, 4 and 6 for the time step $k = 1.25 \times 10^{-6}$ are respectively 3.38×10^{-3}, 7.94×10^{-8} and 2.3×10^{-12} (for the first two components) while, for $atol = 10 \times rtol = 10^{-15}$ and $k_{max} = 2 \times 10^{-5}$, the solvers `stiff` and `rkf` give, respectively, the absolute errors 6.6×10^{-10} and 2.36×10^{-6}.

Table 3 Error (order of convergence) for the first component of the solution for $B5$ modified

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

6.4 Problem E5 [8], stiff with complex eigenvalues of predominantly negative real parts

$$\begin{aligned} y_1' &= -7.89\times 10^{-10}y_1 - 1.1\times 10^{7}y_1y_2 \\ y_2' &= 7.89\times 10^{-10}y_1 - 1.13\times 10^{9}y_2y_3 \\ y_3' &= 7.89\times 10^{-10}y_1 - 1.1\times 10^{7}y_1y_2 + 1.13\times 10^{3}y_4 - 1.13\times 10^{9}y_2y_3 \\ y_4' &= 1.1\times 10^{7}y_1y_2 + 1.13\times 10^{3}y_4 \\ y(0) &= (1.76\times 10^{-3}, 0; 0; 0)^t, T = 1000. \end{aligned} \tag{58}$$

A reference solution is computed with $DC10$ for $k = 10^{-3}$. The solution of this problem has small magnitude in $[1.618\times 10^{-3}, 1.76\times 10^{-3}]\times[0, 1.46\times 10^{-10}]\times[0, 8.27\times 10^{-12}]\times[0, 1.38\times 10^{-10}]$ and the eigenvalues of the Jacobian matrix $dF(y)$ along the solution curve belong to the region $[-20490, 3.68\times 10^{-12}]\times[-9.17\times 10^{-5}, 9.17\times 10^{-5}]$ of the complex plane. Table 4 gives the absolute errors and order of accuracy for the four components of the approximate solutions. For the stepsize $k = 10^{-3}$, the maximal absolute errors in all the four components of the solution are equal on seventeen digits for $DC4$, $DC6$ and $DC8$. The implemented RK4 diverges to ∞ for time step $k \geq 2\times 10^{-4}$ while its absolute errors for $k = 10^{-4}$ are 2.03×10^{-16}, 2.78×10^{-22}, 1.29×10^{-18} and 1.29×10^{-18}, respectively, for the four components. The absolute errors from the solver `stiff` for $k_{max} = 10^{-3}$ and $rtol = 10^{8}\times atol = 10^{-15}$ are 1.29×10^{-16}, 7.46×10^{-22}, 4.72×10^{-23} and 7.14×10^{-22}, respectively, for the four components. The solver `rkf` is not efficient for this problem. The errors for the first component of the approximate solution for the implemented $BDF1, \cdots, BDF2$, with initial value deduced from the reference solution, for $k = 10$ are, respectively, $4.4\times 10^{-7}, 5.7\times 10^{-8}, 5.5\times 10^{-9}, 6.6\times 10^{-10}, 2.3\times 10^{-10}, 3.5\times 10^{-11}$.

6.5 Robertson (1966) [10], stiff with real negative eigenvalues

$$\begin{aligned} y_1' &= -0.04y_1 + 10^{4}y_2y_3 \\ y_2' &= 0.04y_1 - 10^{4}y_2y_3 - 3.10^{7}y_2^2 \\ y_3' &= 3.10^{7}y_2^2 \\ y(0) &= (1,0,0)^t, \quad T = 10^{5}. \end{aligned} \tag{59}$$

This is one of the three problems considered as stiffest in [10]. We compute a reference solution with DC10 for the time step $k = 1/6000$. The solution belongs to the region $[1.78\times 10^{-2}, 1.00]\times[0, 3.58\times 10^{-5}]\times[0, 0.983]$ and the eigenvalues of the Jacobian dF(y) along the solution curve belong to $[-9825.744, 0]$. Table 5 gives absolute errors

Table 4 Absolute error (order of convergence) for the problem E5

k	DC2	DC4	DC6	DC8	DC10
100	2.79e-07	5.34e-08	8.31e-09	4.26e-09	1.04e-09
	8.30e-12	9.68e-13	6.86e-14	6.14e-14	1.66e-14
	4.47e-13	5.31e-14	3.28e-15	3.40e-15	8.42e-16
	7.85e-12	9.14e-13	6.54e-14	5.81e-14	1.57e-14
12.5	4.94e-09(1.94)	5.88e-11(3.27)	1.84e-12(4.04)	4.98e-14(5.46)	4.44e-15(5.95)
	1.22e-13(2.03)	6.42e-16(3.52)	9.07e-18(4.29)	7.45e-19(5.44)	6.33e-20(5.99)
	6.71e-15(2.02)	9.18e-17(3.06)	6.08e-17(1.92)	6.44e-17(1.91)	7.03e-17(1.19)
	1.15e-13(2.03)	6.63e-16(3.48)	6.89e-17(3.29)	6.51e-17(3.27)	7.04e-17(2.60)
10	3.16e-09(1.99)	2.37e-11(4.07)	5.26e-13(5.62)	1.28e-14(6.08)	4.51e-16(10.3)
	7.77e-14(2.00)	2.79e-16(3.74)	3.02e-18(4.93)	1.15e-19(8.38)	7.28e-21(9.69)
	4.31e-15(1.98)	7.08e-17(1.16)	5.91e-17(0.13)	6.27e-17(0.12)	6.84e-17(0.12)
	7.34e-14(2.00)	3.20e-16(3.26)	6.18e-17(0.49)	6.28e-17(0.16)	6.84e-17(0.13)

and orders of accuracy for each component of the solution. For $rtol = 100 \times atol = 10^{-15}$ and $k_{max} = 1/300$, the solver `stiff` gives the absolute errors 7.28×10^{-13}, 0 and 7.02×10^{-13}, respectively for the first, second and third components of the approximate solution. The solver `rkf` is not able to solve this problem. It would require $k_{max} \simeq 10^{-4}$ (see the argument in [22] about initial time step).

Table 5 Absolute error (order of convergence) for Robertson problem

k	DC2	DC4	DC6	DC8	DC10
0.5	3.63e-5	4.46e-6	2.08e-6	2.91e-6	3.09e-6
	3.63e-5	4.46e-6	2.08e-6	2.91e-6	3.09e-6
	7.12e-5	4.37e-7	1.02e-7	4.12e-7	4.26e-7
1/300	4.7e-9 (1.8)	1.09e-9 (1.7)	4.0e-10 (1.7)	3e-10 (1.9)	2e-10 (1.9)
	7.4e-9 (1.7)	2.23e-8 (1.1)	4.16e-8 (0.8)	2.9e-8 (0.9)	2.5e-8 (0.9)
	4.7e-9 (1.9)	2.12e-8 (0.6)	4.12e-8 (0.6)	2.8e-8 (0.5)	2.5e-8 (0.6)
1/600	1.0e-9 (2.2)	1.5e-10 (2.8)	1.0e-12 (8.6)	9.9e-13 (8.)	7.5e-13(8.2)
	5e-13 (14.)	3e-14 (19.6)	2e-16 (27.7)	2e-16 (27.1)	3e-16 (26.1)
	1.0e-9 (2.2)	1.5e-10 (7.1)	1e-12 (15.3)	9.9e-13 (15)	4e-13 (15.8)
1/6000	9.24D-12	7.31D-14	1.48D-14	4.57D-14	–
	5.38D-15	0.	0.	0.	–
	9.25D-12	2.07D-13	1.36D-13	8.27D-14	–

6.6 van der Pol oscillator [8, 22], stiff, arbitrary complex eigenvalues

$$\begin{aligned} &y_1' = y_2 \\ &y_2' = \mu(1-y_1^2)y_2 - y_1 \\ &y_1(0) = 2, \quad y_2(0) = 0, T = 3000, \mu = 1000. \end{aligned} \tag{60}$$

This problem was initially proposed for $T = 1$ and $\mu = 5$ in [8]. The actual version results from a suggestion by Shampine [22]. We compute a reference solution with DC8 for $k = 1.875 \times \times 10^{-6}$. The solution belong to the region $[-2, 2.000073] \times$

$[-1323.04, 1231.35]$ of the real plan and the eigenvalues along the solution curve belong to the region $[-3000.29, 1123.17] \times [-1158.48, 1158.48]$ of the complex plan. Table 6 gives the absolute errors and orders of accuracy. For $rtol = 10atol = 10^{-16}$ and $k_{max} = 7.5 \times 10^{-5}$, the absolute errors from the solvers `rkf` are 3.54×10^{-2} and 64.76, respectively, for the first and second components of the solution while `stiff` gives 2.16×10^{-6} and 3.48×10^{-3}.

Table 6 Absolute error for the van der Pol's equation

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

6.7 Discussion of the numerical results

1. The Bernoulli equation is stiff and strongly nonlinear (the approximate solutions with the explicit forth order Runge-Kutta method diverges to $-\infty$ for time steps $k \geq 2.03 \times 10^{-3}$). The magnitude of the derivatives of $F = F(u)$ with respect to u increases exponentially with the order of the derivative, and the magnitude of the solution u is neither large nor small since $0.226 \leq u(t) \leq 1$, for $0 \leq t \leq 10$. Nevertheless, the errors for 10 steps, which corresponds to a time step $k = 1$, decrease with the level of correction and are less than 2.5×10^{-4}, for DC6, DC8 and DC10. This illustrates that high order derivatives of F do not strongly affect the quality of the approximate solutions with the DC schemes.
2. The behavior of the DC schemes on the oscillatory problem shows the ability of the DC method for long-term integration. Each DC scheme reached its theoretical order of convergence with a good accuracy way better than from the BDF and RK methods.
3. For the modified problem B5, each DC scheme converges towards machine accuracy with its theoretical order of convergence, but the time step required is somewhat small. This behaviour is not generic. For the problem E5 in [8], which also has complex eigenvalues, very accurate solutions are obtained from the DC methods even for time steps greater than 100.
4. For the Robertson problem, which corresponds to a system with real negative eigenvalues, there is no restriction on the time step k for an accurate approximate solution with the DC schemes, and high order DC methods can be avoided (DC6 is enough). The convergence is slow for $k > 1/300$, but superconvergent happens for k in the asymptotic region ($k < 1/300$).
5. The van der Pol oscillator is stiff and the solution has a large magnitude. DC6 and DC8 reached their order of accuracy. The order of convergence for DC10

is not observed because of a quick saturation of the error while DC2 and DC4 require smaller time steps.

In general, a careful assessment of the proof of Theorem 1 points out to the fact that, for a system with complex eigenvalues $\lambda = \lambda_1 + i\lambda_2$, we only need a time step k such that $kMax\{\lambda_1, |\lambda_2|\} < 2$ for a good accuracy (superconvergence happens when $-\lambda_1 >> |\lambda_2|$). However, time steps k such that $k\mu \simeq k|\lambda| < 2$, $\mu \simeq \max_{0\leq t\leq T} \|d_x F(t, u(t))\|$, is necessary for an asymptotic convergence (see Remark 3). For example, in the case of the Bernoulli equation we have $\lambda \simeq -20000.1 < 0$ and $\mu = 20000.1$. There is no restriction on the time steps for accurate approximation, but asymptotic convergences are observed only for $k\mu < 2$.

For the computational effort of the DC methods, we recall that to compute an approximate solution on discrete points $0 = t_0 < t_1 < \cdots < t_N = T$, $DC2$ solves N nonlinear systems while $DC2j$, $j \geq 2$, solves $j \times N$ systems. In the case of the Bernoulli equation, for example, $DC10$ achieves the maximal error of about 1.1×10^{-11} by solving approximately 5×10^6 nonlinear systems while the maximal absolute error for $DC2$ is about 8.9×10^{-7} for $N = 5 \times 10^6$. Since the resolution of nonlinear systems is the main burden for these methods, using high order DC methods is advantageous.

7 Conclusions

We have presented a new approach of deferred correction methods for the numerical solution of general first order ordinary differential equations. Proofs for consistency, order of convergence and stability of the method are given. The numerical experiments comply with the theory and show a high accuracy of the method and its satisfactory A-stable property. Globally, each DC scheme reaches its proper order of convergence and applies to any category of problem, providing accurate approximations for time steps not necessarily small. The accuracy of the DC schemes increases with the level of correction.

References

1. Auzinger, W.: Encyclopedia of Applied and Computational Mathematics, chap. Defect Correction Methods, pp. 323–332. Springer, Berlin, Heidelberg (2015)
2. Christlieb, A., Ong, B., Qiu, J.M.: Integral deferred correction methods constructed with high order Runge-Kutta integrators. Math. Comp. **79**, 761–783 (2010)
3. Chung, T.: Computational Fluid Dynamics, 2nd edn. Cambridge university press (2010)
4. Dahlquist, G., Björck, A.k.: Numerical methods in scientific computing. Vol. I. SIAM, Philadelphia, PA (2008)
5. Dahlquist, G.G.: A special stability problem for linear multistep methods. Nordisk Tidskr. Informationsbehandling (BIT) **3**, 27–43 (1963)
6. Daniel, J.W., Pereyra, V., Schumaker, L.L.: Iterated deferred corrections for initial value problems. Acta Cient. Venezolana **19**, 128–135 (1968)
7. Dutt, A., Greengard, L., Rokhlin, V.: Spectral deferred correction methods for ordinary differential equations. BIT **40**, 241–266 (2000)
8. Enright, W.H., Hull, T., Lindberg, B.: Comparing numerical methods for stiff systems of ODE:s. BIT **15**, 1–48 (1975)
9. Gustafsson, B., Kress, W.: Deferred correction methods for initial value problems. BIT **41**, 986–995 (2001)

10. Hairer, E., Wanner, G.: Solving ordinary differential equations. II. Stiff and differential-algebraic problems, vol. 14. Springer-Verlag, Berlin (1991)
11. Hansen, A.C., Strain, J.: On the order of deferred correction. Appl. Numer. Math. **61**, 961–973 (2011)
12. Hildebrand, F.B.: Introduction to Numerical Analysis. McGraw-Hill Book Co., New York-Düsseldorf-Johannesburg (1974)
13. Hull, T.E., Enright, W.H., Fellen, B.M., Sedgwick, A.E.: Comparing numerical methods for ordinary differential equations. SIAM J. Numer. Anal. **9**, 603–637 (1972)
14. Isaacson, E., Keller, H.B.: Analysis of numerical methods. John Wiley & Sons, Inc., New York-London-Sydney (1966)
15. Karouma, A.: A class of contractivity preserving hermite-birkhoff-taylor high order time discretization methods. Ph.D. thesis, Université d'Ottawa/University of Ottawa (2015)
16. Koyaguerebo-Imé, S.C.E., Bourgault, Y.: Finite difference and numerical differentiation: General formulae from deferred corrections. arXiv preprint arXiv:2005.11754 (2020)
17. Koyaguerebo-Imé, S.C.R., Bourgault, Y.: Arbitrary high-order unconditionally stable methods for reaction-diffusion equations via deferred correction: Case of the implicit midpoint rule. Submitted to ESAIM Math. Model. Numer. Anal.[arXiv:2006.02962] (2020)
18. Kress, W., Gustafsson, B.: Deferred correction methods for initial boundary value problems. J. Sci Comput. **17**(1-4), 241–251 (2002)
19. Kushnir, D., Rokhlin, V.: A highly accurate solver for stiff ordinary differential equations. SIAM J. Sci. Comput. **34**, A1296–A1315 (2012)
20. Quarteroni, A., Sacco, R., Saleri, F.: Numerical mathematics, vol. 37, second edn. Springer-Verlag, Berlin (2007)
21. Schild, K.H.: Gaussian collocation via defect correction. Numer. Math. **58**, 369–386 (1990)
22. Shampine, L.F.: Evaluation of a test set for stiff ODE solvers. ACM Trans. Math. Software **7**, 409–420 (1981)
23. Spijker, M.N.: Stiffness in numerical initial-value problems. J. Comput. Appl. Math. **72**, 393–406 (1996)
24. Stewart, K.: Avoiding stability-induced inefficiencies in BDF methods. J. Comput. Appl. Math. **29**, 357–367 (1990)

Chapter 3

Arbitrary high-order unconditionally stable methods for reaction-diffusion equations via Deferred Correction

This chapter is presented in terms of a journal article and submitted to Mathematical Modelling and Numerical Analysis (ESAIM Math. Model. Numer. Anal), carrying the same title as mentioned above. Please see the attached paper for the content.

Mathematical Modelling and Numerical Analysis Will be set by the publisher
Modélisation Mathématique et Analyse Numérique

ARBITRARY HIGH-ORDER UNCONDITIONALLY STABLE METHODS FOR REACTION-DIFFUSION EQUATIONS VIA DEFERRED CORRECTION: CASE OF THE IMPLICIT MIDPOINT RULE

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

Abstract. In this paper we analyse full discretizations of an initial boundary value problem (IBVP) related to reaction-diffusion equations. The IBVP is first discretized in time via the deferred correction method for the implicit midpoint rule and leads to a time-stepping scheme of order $2p+2$ of accuracy at the stage $p=0,1,2,\cdots$ of the correction. Each semi-discretized scheme results in a nonlinear elliptic equation for which the existence of a solution is proven using the Schaefer fixed point theorem. The elliptic equation corresponding to the stage p of the correction is discretized by the Galerkin finite element method and gives a full discretization of the IBVP. This fully discretized scheme is unconditionlly stable with order $2p+2$ of accuracy in time. The order of accuracy in space is equal to the degree of the finite element used when the family of meshes considered is shape-regular while an increment of one order is proven for shape-regular and quasi-uniform family of meshes. A numerical test with a bistable reaction-diffusion equation having a strong stiffness ratio is performed and shows that the orders 2,4,6,8 and 10 of accuracy in time are achieved with a very strong stability.

1991 Mathematics Subject Classification. 35K57, 35B05, 65N30, 65M12.

.

Introduction

Let Ω be a bounded domain in $\mathbb{R}^d$ $(d=1,2,3)$ with smooth boundary $\partial\Omega$ and $T>0$. Consider the following reaction-diffusion system with Cauchy-Dirichlet conditions

$$\begin{cases} u' - M\Delta u + f(u) = S \text{ in } \Omega\times(0,T) \\ \qquad\qquad u = 0 \text{ on } \partial\Omega\times(0,T) \\ \qquad u(.,0) = u_0 \text{ in } \Omega, \end{cases} \tag{1}$$

where $u:\Omega\times[0,T]\to\mathbb{R}^J$ is the unknown, for a positive integer J, M is an $J\times J$ constant matrix, $f:\mathbb{R}^J\to\mathbb{R}^J$ and $S:\Omega\times(0,T)\to\mathbb{R}^J$ are given smooth functions. This is a general form of reaction-diffusion equations (see for instance [1]) that model various phenomena in physics, combustion, chemical reactions, population dynamics and biomedical science (cancer modelling and other physiological processes) (see, e.g., [1–5]).

Keywords and phrases: time-stepping methods, deferred correction, high order methods, reaction-diffusion equations, finite elements

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

We suppose that M is positive definite and the function f satisfies the following two monotonicity conditions

$$(f(x)-f(y),x-y)\geq\alpha|x-y|^q+\tau(y)|x-y|^2,\forall x,y\in\mathbb{R}^J,\text{for some }\alpha\geq 0,q\geq 1, \tag{2}$$

and

$$(df(x)y)\cdot y\geq-\mu_0|y|^2,\ \forall x,y\in\mathbb{R}^J, \tag{3}$$

where μ_0 is a nonnegative real, and τ is an arbitrary continuous real-valued function. These conditions guarantee the existence of a solution of problem (1) in $L^2\left(0,T;H_0^1(\Omega)\cap H^2(\Omega)\right)$ (see for instance [6–8]), and uniqueness and high order regularity can be deduced. The conditions (2)-(3) are at least satisfied by any polynomial of odd degree with positive leading coefficient, and the matrix M is supposed to be constant only for the sake of simplicity. In fact, all our results remain true replacing the operator $M\Delta$ by an elliptic operator L:

$$Lu=-\sum_{i,j=1}^{J}a^{i,j}(x)u_{x_ix_j}+\sum_{i=1}^{J}b^j(x)u_{x_i}+c_0(x)u, \tag{4}$$

where the coefficients $a^{i,j}$, b^i and c_0 are smooth functions, and $a^{i,j}=a^{j,i}$ (see, e.g., [8, p.292] for a definition of elliptic operator). The analysis also remains true substituting the Dirichlet condition in (1) by Neumann conditions.

The numerical analysis of reaction-diffusion equations takes advantage of many results available from the numerical analysis of semi-linear parabolic partial differential equations (PDEs). The method of lines (MOL) is commonly used. By this method the PDE is first discretized in space by finite element or finite difference methods, leading to a system of ordinary differential equations (ODEs). The resulting system of ODEs is then discretized by fully implicit or implicit-explicit (IMEX) time-stepping methods (see for instance [9–16]). In [9–11], linear implicit-explicit multistep methods in time together with finite element methods in space are analysed for a class of abstract semi-linear parabolic equations that includes a large class of reaction-diffusion systems. The approaches in [9–11] are the same. The authors investigate approximate solutions expected to be in a tube around the exact solution. They proceeded by induction by adapting the time step k and the space step h and established that if k and $k^{-1}h^{2r}$, $r\geq 2$, are small enough then the global error of the scheme is of order p ($p=1,2,...,5$) in time and r in space. IMEX schemes with finite difference in space and Runge-Kutta of order 1 and 2 in time are also analysed in [17, 18] for a class of reaction-diffusion systems. Otherwise, in [12, 13, 19] fully implicit numerical methods for reaction-diffusion equations with restrictive conditions on the nonlinear term are introduced, combining finite elements in space and backward Euler, Crank-Nicolson or fractional-step θ methods in time. The resulting schemes are unconditionally stable (the time step is independent from the space step) with order 1 or 2 of accuracy in time. The time-stepping method in [16] is constructed via a deferred correction strategy applied to the trapezoidal rule and is of arbitrary high order. However, this method concerns only linear initial value problems (IVP) (resulting eventually from a MOL) satisfying a monotonicity condition and has an issue for the starting procedure. Furthermore, the stability analysis proposed in this paper does not guarantee unconditional stability and/or an optimal a priori error estimate, when a full discretization is considered.

In practice, the space-discretization of time-evolution PDEs leads to a stiff IVP of large dimension (we recall that a stiff problem is a problem extremely hard to solve by standard explicit step-by-step methods (see, e.g., [20]). To avoid overly small time steps, accurate approximate solutions for these IVPs require high order time-stepping methods having good stability properties (A-stable methods are of great interest). Backward differentiation formulae (BDF) of order 1 and 2 are commonly used according to their A-stability. However, BDF methods of order 3 and higher lack stability properties (e.g. for systems with complex eigenvalues). Moreover, Runge-Kutta methods applied to such IVPs have order of convergence reduced to 1 or 2 (see [21]), and are inefficient when the IVPs are stiffer.

The aim of this paper is to apply the deferred correction (DC) method introduced in [22] for the semi-discretization in time of the problem (1). The deferred correction method consists in a successive perturbation

(correction) of the implicit midpoint rule, leading to A-stable schemes of order $2p+2$ at the stage $p=1,2,\cdots$ of the correction. The order of accuracy of the DC schemes is guaranteed by a deferred correction condition (DCC). Applying the DC method to (1), the main difficulty is to prove that the resulting schemes satisfy DCC up to a certain stage p of the correction so that we obtain a time semi-discrete approximate solution with order $2p+2$ of accuracy. To overcome this difficulty, we suppose that the exact solution u of (1) is stationary in a small time interval $[0,(2p+1)k_0]$, where k_0 is a maximal time step for the time semi-discretized schemes and satisfies $k_0\mu_0<2$ (μ_0 is the constant introduced in (3)). The stationary hypothesis is a simple trick to simplify our proof. Indeed, the DCC is proven without restrictive condition in the case of IVP (see [22]), but the difficulty in the case of PDEs is related to the presence of unbounded operator. Each semi-discretized scheme in time leads to a nonlinear elliptic equation that is discretized using the Galerkin finite element method. It results an arbitrary high-order unconditionally stable methods for the numerical solution of problem (1). A numerical illustration using the bistable reaction-diffusion equation with the schemes of order 2, 4, 6, 8 and 10 in time is given.

The paper is organized as follows. We recall some algebraic property of finite difference operators in section 1. In section 2 we introduce the semi-discretized schemes in time and prove the existence of a solution. The analysis of convergence and order of accuracy of solutions for the semi-discretized schemes in time is done in section 3. The fully discretized schemes are presented and analysed in section 4, and numerical experiments are carried in section 5.

1. Finite difference operators

In this section we recall main results from finite difference (FD) approximations. Details and proofs for these results can be found in [23]. For a time step $k>0$, we denote $t_n=nk$ and $t_{n+1/2}=(n+1/2)k$, for each integer n. This implies that $t_0=0$. We consider the time steps k such that $0=t_0<t_1<\cdots<t_N=T$ is a partition of $[0,T]$, for a nonnegative integer N. The centered, forward and backward difference operators D, D_+ and D_-, respectively, related to k and applied to a function v from $[0,T]$ into a Banach space X (with norm $\|\cdot\|_X$), are defined as follows:

$$Dv(t_{n+1/2})=\frac{v(t_{n+1})-v(t_n)}{k},$$

$$D_+v(t_n)=\frac{v(t_{n+1})-v(t_n)}{k},$$

and

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

The average operator is denoted by E:

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

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

$$(D_+D_-)^mv(t_n)=k^{-2m}\sum_{i=0}^{2m}(-1)^i\binom{2m}{i}v(t_{n+m-i}),\tag{5}$$

and

$$D_-(D_+D_-)^mv(t_n)=k^{-2m-1}\sum_{i=0}^{2m+1}(-1)^i\binom{2m+1}{i}v(t_{n+m-i}),\tag{6}$$

for each integer $m \geq 1$ such that $0 \leq t_{n-m-1} \leq t_{n+m} \leq T$. If $\{v^n\}_n$ is a sequence of approximation of v at the discrete points t_n, the finite difference operators apply to $\{v^n\}$ and we define

$$Dv^{n+1/2} = D_+v^n = D_-v^{n+1} = \frac{v^{n+1}-v^n}{k}.$$

and

$$Ev^{n+1/2} = \widehat{v}^{n+1} = \frac{v^{n+1}+v^n}{2}.$$

We have the following three results:

Result 1

For nonnegative integers m_1 and m_2, provided $v \in C^{m_1+m_2}([0,T],X)$ and $m_2 \leq n \leq N-m_1$, we have

$$\left\|D_+^{m_1}D_-^{m_2}v(t_n)\right\| \leq \max_{t_{n-m_2}\leq t\leq t_{n+m_1}} \left\|\frac{d^{m_1+m_2}v}{dt^{m_1+m_2}}(t)\right\|. \tag{7}$$

Result 2 (Central finite difference approximations)

There exists a sequences $\{c_i\}_{i\geq 2}$ of real numbers such that, for all $v \in C^{2p+3}([0,T],X)$, where p is a positive integer, and $p \leq n \leq N-1-p$, we have

$$v'(t_{n+1/2}) = \frac{v(t_{n+1})-v(t_n)}{k} - \sum_{i=1}^{p} c_{2i+1}k^{2i}D(D_+D_-)^i v(t_{n+1/2}) + O(k^{2p+2}), \tag{8}$$

and

$$v(t_{n+1/2}) = \frac{v(t_{n+1})+v(t_n)}{2} - \sum_{i=1}^{p} c_{2i}k^{2i}(D_+D_-)^i Ev(t_{n+1/2}) + O(k^{2p+2}). \tag{9}$$

Table 1 gives the ten first coefficients c_i.

TABLE 1. Ten first coefficients of central difference approximations (8) and (9)

c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}
$\frac{1}{8}$	$\frac{1}{24}$	$-\frac{18}{4!2^5}$	$-\frac{18}{5!2^5}$	$\frac{450}{6!2^7}$	$\frac{450}{7!2^7}$	$-\frac{22050}{8!2^9}$	$-\frac{22050}{9!2^9}$	$\frac{1786050}{10!2^{11}}$	$\frac{1786050}{11!2^{11}}$

Result 3 (Interior central finite difference approximations)

For each positive integer p there exists reals $c_2^p, c_3^p, \cdots, c_{2p+1}^p$ such that, for each $v \in C^{2p+3}([a,b],X)$ and a uniform partition $a = \tau_0 < \tau_1 < ... < \tau_{2p+1} = b$ of the interval $[a,b]$, with $\tau_n = a+nk$, $k = (b-a)/(2p+1)$ and $\tau_{p+1/2} = (a+b)/2$, we have

$$u'(\tau_{p+1/2}) = \frac{u(b)-u(a)}{b-a} - \frac{1}{b-a}\sum_{i=1}^{p} c_{2i+1}^p k^{2i+1}D(D_+D_-)^i u(\tau_{p+1/2}) + O(k^{2p+2}), \tag{10}$$

and

$$u(\tau_{p+1/2}) = \frac{u(b)+u(a)}{2} - \sum_{i=1}^{p} c_{2i}^p k^{2i}(D_+D_-)^i Eu(\tau_{p+1/2}) + O(k^{2p+2}). \tag{11}$$

Table 2 gives the coefficients c_i^p for $p = 1,2,3,4$.

TABLE 2. Coefficients of the approximations (10)-(11) for $p=1,2,3,4$

p	c_2^p	c_3^p	c_4^p	c_5^p	c_6^p	c_7^p	c_8^p	c_9^p
1	$\frac{9}{8}$	$\frac{9}{8}$						
2	$\frac{25}{8}$	$\frac{125}{24}$	$\frac{125}{128}$	$\frac{125}{128}$				
3	$\frac{49}{8}$	$\frac{343}{24}$	$\frac{637}{128}$	$\frac{13377}{1920}$	$\frac{1029}{1024}$	$\frac{1029}{1024}$		
4	$\frac{81}{8}$	$\frac{243}{8}$	$\frac{1917}{128}$	$\frac{17253}{640}$	$\frac{7173}{1024}$	$\frac{64557}{7168}$	$\frac{32733}{32768}$	$\frac{32733}{32768}$

2. Semi-discrete schemes in time: existence of a solution

Hereafter we suppose that (1) has a unique solution $u \in C^{2p+4}\left([0,T], H^{r+1}(\Omega)\cap H_0^1(\Omega)\right)$, for some positive integers p and r. We denote by $(\cdot,\cdot)$ the inner product in $L^2(\Omega)$ and by $\|\cdot\|$ the corresponding norm. The norm in the Sobolev spaces $H^m(\Omega)$ will be noted $\|\cdot\|_m$, for each nonnegative integer m, and we note $\|\cdot\|_\infty = \|\cdot\|_{L^\infty(\Omega)}$. We use h and k to denote stepsizes for space and time discretizations, respectively. The letter C will denote any constant independent from h and k, and that can be calculated explicitly in term of known quantities. The exact value of C may change from a line to another line.

As in [22], we can apply deferred correction method to (1) and deduce the following schemes:

For $j=0$, we have the *implicit midpoint rule*

$$\begin{cases} \dfrac{u^{2,n+1}-u^{2,n}}{k} - M\Delta\left(\dfrac{u^{2,n+1}+u^{2,n}}{2}\right) + f\left(\dfrac{u^{2,n+1}+u^{2,n}}{2}\right) = s(t_{n+1/2}), \text{ in } \Omega, \\ u^{2,n} = 0 \text{ on } \partial\Omega, \\ u^{2,0} = u_0. \end{cases} \tag{12}$$

For $j \geq 1$, we have

$$\begin{cases} \dfrac{u^{2j+2,n+1}-u^{2j+2,n}}{k} - D\Lambda^j u^{2j,n+1/2} - M\Delta\left(\widehat{u}^{2j+2,n+1} - \Gamma^j E u^{2j,n+1/2}\right) \\ + f\left(\dfrac{u^{2j+2,n+1}+u^{2j+2,n}}{2} - \Gamma^j E u^{2j,n+1/2}\right) = s(t_{n+1/2}), \text{ in } \Omega, \text{ for } n \geq j+1, \\ u^{2j+2,n} = 0 \text{ on } \partial\Omega, \\ u^{2j+2,0} = u_0, \end{cases} \tag{13}$$

where Γ and Λ are finite differences operators defined for each positive integer j, and $n \geq j$, by

$$\Lambda^j u(t_n) = \sum_{i=1}^{j} c_{2i+1} k^{2i} (D_+D_-)^i u(t_n) = \sum_{i=1}^{j}\sum_{l=0}^{2i} c_{2i+1}(-1)^l \binom{2i}{l} u(t_{n+i-l}), \tag{14}$$

and

$$\Gamma^j u(t_n) = \sum_{i=1}^{j} c_{2i} k^{2i} (D_+D_-)^i u(t_n) = \sum_{i=1}^{j}\sum_{l=0}^{2i} c_{2i}(-1)^l \binom{2i}{l} u(t_{n+i-l}). \tag{15}$$

The scheme (12) has unknowns $\left\{u^{2,n}\right\}_{n=1}^N$ corresponding to approximations of $u(t_n)$, expected to be of order 2 of accuracy. For (13) the unknowns are $\left\{u^{2j+2,n}\right\}_{n=j+1}^N$, expected to be of order $2j+2$, while $\left\{u^{2j,n}\right\}_{n=j}^N$ is supposed known from the preceding stage. To avoid computing approximate solution of (1) for $t<0$, the

scheme (13) is used only for $n \geq j$. For the starting values, $0 \leq n \leq j-1$, we consider the scheme

$$\begin{cases} Du^{2j+2,n+1/2} - \dfrac{1}{2j+1}\bar{\Lambda}^j D\bar{u}^{2j,n_j+1/2} - M\Delta\left(\widehat{u}^{2j+2,n+1} - \bar{\Gamma}^j E\bar{u}^{2j,n_j+1/2}\right) \\ \qquad + f\left(\widehat{u}^{2j+2,n+1} - \bar{\Gamma}^j E\bar{u}^{2j,n_j+1/2}\right) = s(t_{n+1/2}), \\ u^{2j+2,n} = 0 \text{ on } \partial\Omega, \\ u^{2j+2,0} = u_0, \end{cases} \tag{16}$$

where we set $n_j = (2j+1)n + j$,

$$\frac{1}{2j+1}\bar{\Lambda}^j D\bar{u}^{2j,(2j+1)n+j+1/2} = k^{-1}\sum_{i=1}^{j}\sum_{l=0}^{2i+1} c^j_{2i+1}(-1)^l \binom{2i+1}{l} \bar{u}^{2j,(2j+1)n+j+i-l+1}, \tag{17}$$

and

$$\bar{\Gamma}^j \bar{u}^{2j,(2j+1)n+j} = \sum_{i=1}^{j}\sum_{l=0}^{2i} c^j_{2i}(-1)^l \binom{2i}{l} \bar{u}^{2j,(2j+1)n+j+i-l}. \tag{18}$$

This scheme is built from (10) and (11), for $a = t_n$ and $b = t_{n+1}$. $\left\{\bar{u}^{2,n}\right\}_{n=1}^{N}$ is computed from (12) with time the step $k/3$ instead of k. Similarly, $\left\{\bar{u}^{2j,n}\right\}_{n=j}^{N}$, $j \geq 2$, is computed from the scheme (13) with the time step $k/(2j+1)$ instead of k.

To prove the existence of a solution for the schemes (12) and (13), we need the following lemma.

Lemma 1. *Let $k > 0$ such that $k|\tau(0)| \leq 1/4$, and $v \in L^2(\Omega)$. Then the elliptic problem*

$$u - kM\Delta u + kf(u) = v \quad \text{in } \Omega, \tag{19}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{20}$$

has a solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfying the inequality

$$\|u\|_2 \leq C\left(|||M|||/\gamma\right)\left(k^{-1}\|v-u\| + \sqrt{2\gamma\mu_0}\|\nabla u\|\right), \tag{21}$$

where γ is the smallest eigenvalue of the positive definite matrix M, $|||M|||$ is any norm of the matrix M, and the function τ and the scalar μ_0 are defined in (2) and (3), respectively.

Proof. The existence can be deduced from the Schaefer fixed point theorem [8, p. 504]. In fact, given $u \in H^2(\Omega) \cap H_0^1(\Omega)$, the problem

$$w - kM\Delta w + kf(u) = v \quad \text{in } \Omega, \tag{22}$$

$$w = 0 \quad \text{on } \partial\Omega, \tag{23}$$

has a unique solution $w \in H^2(\Omega) \cap H_0^1(\Omega)$ (see [8, p.317]). Consider the nonlinear mapping

$$A \ : \ H^2(\Omega) \cap H_0^1(\Omega) \longrightarrow H^2(\Omega) \cap H_0^1(\Omega),$$

which maps $u \in H^2(\Omega) \cap H_0^1(\Omega)$ to the unique solution $w = A[u]$ of (22)-(23). It is enough to prove that A is continuous, compact, and that the set

$$\Sigma = \left\{ u \in H^2(\Omega) \cap H_0^1(\Omega) | u = \lambda A[u], \text{ for some } \lambda \in [0,1] \right\} \tag{24}$$

is bounded.

(i) The mapping A is continuous. Indeed, let $\{u_m\}_{m=1}^{\infty}$ in $H^2(\Omega)\cap H_0^1(\Omega)$ which converges to $u\in H^2(\Omega)\cap H_0^1(\Omega)$. For each $m=1,2,\cdots$, let $w_m=A[u_m]$ and $w=A[u]$. Then $w-w_m$ belongs to $H^2(\Omega)\cap H_0^1(\Omega)$ and satisfies the equation

$$(w-w_m)-kM\Delta(w-w_m)+k\,(f(u)-f(u_m))=0 \ \text{ in } \Omega. \tag{25}$$

The inner product of the last identity with $w-w_m$ yields

$$\|w-w_m\|^2+\gamma k\|\nabla(w-w_m)\|^2+k\,(f(u)-f(u_m),w-w_m)\le 0. \tag{26}$$

We can write,

$$f(u(x))-f(u_m(x))=\int_0^1 df\,(u(x)-\xi(u(x)-u_m(x)))(u(x)-u_m(x))\,d\xi.$$

Since $u_m\longrightarrow u$ in $H^2(\Omega)$ and $H^2(\Omega)\hookrightarrow C^0(\overline{\Omega})$, there exists a positive integer m_0 such that $m\ge m_0$ implies

$$\max_{x\in\overline{\Omega}}|u(x)-u_m(x)|\le c_2\|u-u_m\|_2\le 1, \tag{27}$$

where c_2 is the constant from the Sobolev embedding. It follows that

$$|f(u(x))-f(u_m(x))|\le\beta|u(x)-u_m(x)|, \tag{28}$$

where

$$\beta=\max_{|y|\le 1+c_2\|u\|_2}|df(y)|.$$

Therefore, by Cauchy-Schwartz inequality we have

$$k\,|(f(u)-f(u_m),w-w_m)|\le k\beta\|u-u_m\|\|w-w_m\|\le\frac{(k\beta)^2}{2}\|u-u_m\|^2+\frac{1}{2}\|w-w_m\|^2.$$

The last inequality substituted into (26) yields

$$\|w-w_m\|^2+2\gamma k\|\nabla(w-w_m)\|^2\le(k\beta)^2\|u-u_m\|^2.$$

It follows that $w_m\to w$ in $H_0^1(\Omega)$ when $m\to+\infty$. On the other hand, elliptic regularity results applied to the identity (25) yields, owing to (28) and the last inequality,

$$\|w-w_m\|_2\le C\left(k^{-1}\|w-w_m\|+\|f(u)-f(u_m)\|\right)\le 2\beta C\|u-u_m\|\to 0 \text{ as } m\to+\infty.$$

Whence $\{w_m\}_{m=1}^{+\infty}$ converges to w in $H^2(\Omega)\cap H_0^1(\Omega)$, and the continuity of the mapping A follows.

(ii) The mapping A is compact. Indeed, given a bounded sequence $\{u_m\}_{m\in\mathbb{N}}$ in $H^2(\Omega)\cap H_0^1(\Omega)$, from the compact embedding $H^2(\Omega)\hookrightarrow H_0^1(\Omega)$ we can extract a subsequence $\left\{u_{m_j}\right\}_{j\in\mathbb{N}}$ that converges to u strongly in $H_0^1(\Omega)$ and weakly in $H^2(\Omega)$. The subsequence $\left\{u_{m_j}\right\}_{j\in\mathbb{N}}$ is then bounded in $H^2(\Omega)\cap H_0^1(\Omega)$. Let

$$\kappa=\sup_{m\in\mathbb{N}}\|u_m\|_2 \ \text{ and } \ \beta'=\max_{|y|\le c_2(\kappa+\|u\|_2)}|df(y)|.$$

Therefore, proceeding exactly as in part (i), substituting m by m_j, the inequality (27) by

$$\max_{x\in\overline{\Omega}}|u_{m_j}(x)|\le c_2\sup_{m\in\mathbb{N}}\|u_{m_j}\|_2=c_2\kappa,$$

and β by β' in (28), we deduce that $w_{m_j} = A[u_{m_j}] \to w$ strongly in $H^2(\Omega) \cap H_0^1(\Omega)$. Hence A is compact.

(iii) The set Σ is bounded.

Let $u \in H^2(\Omega) \cap H_0^1(\Omega)$ such that $u = \lambda A[u]$ for some $\lambda \in (0,1]$. Then u satisfies

$$u - kM\Delta u + \lambda k f(u) = \lambda v \quad \text{in } \Omega, \tag{29}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{30}$$

By elliptic regularity results we have

$$\|u\|_2 \leq C\|k^{-1}(\lambda v - u) - \lambda f(u)\| = C\|M\Delta u\|. \tag{31}$$

The inner product of (29) with u, taking the boundary condition (30) into account, yields

$$\|u\|^2 + \gamma k\|\nabla u\|^2 + \lambda k \int_\Omega f(u)\cdot u dx = \lambda \int_\Omega v \cdot u dx.$$

Without loss of generality we suppose that $f(0) = 0$, otherwise we change f by $\tilde{f} = f - f(0)$ and v by $\tilde{v} = v - kf(0)$. Then the monotonicity condition (2) combined with the hypothesis of the lemma yields

$$\lambda k \int_\Omega f(u)\cdot u dx \geq \alpha\lambda k\|u\|^q + \lambda k\tau(0)\|u\|^2 \geq \alpha\lambda k\|u\|^q - \frac{1}{4}\|u\|^2, \ \forall \lambda \in (0,1]. \tag{32}$$

From Cauchy-Schwartz inequality and the Cauchy inequality with $\varepsilon = 1$, we have

$$\lambda \int_\Omega v \cdot u dx \leq \lambda^2\|v\|^2 + \frac{1}{4}\|u\|^2.$$

Substituting the last two inequalities in the previous identity, we deduce that

$$\|u\|^2 + 2\gamma k\|\nabla u\|^2 \leq 2\lambda^2\|v\|^2. \tag{33}$$

On the other hand, the inner product of (29) with $-\Delta u$ yields

$$\gamma\|\Delta u\|^2 \leq k^{-1}\int_\Omega (\lambda v - u)\cdot(-\Delta u)dx + \int_\Omega \lambda f(u)\cdot\Delta u dx. \tag{34}$$

We can write

$$f(u)\cdot\Delta u = \sum_{i=1}^J \nabla\cdot(f_i(u)\nabla u_i) - \sum_{i=1}^J \left(df(u)\left(\frac{\partial u}{\partial x_i}\right)\right)\cdot\frac{\partial u}{\partial x_i},$$

and deduce from (3), the boundary condition and the hypothesis $f(0) = 0$ that

$$\int_\Omega f(u)\cdot\Delta u dx = -\sum_{i=1}^J \int_\Omega \left(df(u)\left(\frac{\partial u}{\partial x_i}\right)\right)\cdot\frac{\partial u}{\partial x_i}dx \leq \mu_0\sum_{i=1}^J \left\|\frac{\partial u}{\partial x_i}\right\|^2 = \mu_0\|\nabla u\|^2.$$

By Cauchy-Schwartz inequality and the Cauchy inequality with $\varepsilon = 1/(2\gamma)$ we have

$$\left|k^{-1}\int_\Omega (\lambda v - u)\cdot(-\Delta u)dx\right| \leq k^{-1}\|\lambda v - u\|\|\Delta u\| \leq \frac{1}{2\gamma k^2}\|\lambda v - u\|^2 + \frac{\gamma}{2}\|\Delta u\|^2.$$

Substituting the last two inequalities in (34), we obtain

$$\gamma^2\|\Delta u\|^2 \le k^{-2}\|\lambda v - u\|^2 + 2\lambda\gamma\mu_0\|\nabla u\|^2.$$

Therefore,

$$\|M\Delta u\|^2 \le (|||M|||/\gamma)^2\left(k^{-2}\|\lambda v - u\|^2 + 2\gamma\mu_0\|\nabla u\|^2\right)$$

since $0 \le \lambda \le 1$, and we deduce from (31) that

$$\|u\|_2 \le C\,(|||M|||/\gamma)\left(k^{-1}\|\lambda v - u\| + \sqrt{2\gamma\mu_0}\|\nabla u\|\right). \tag{35}$$

The last inequality together with (33) yields

$$\|u\|_2 \le C\,(|||M|||/\gamma)\,k^{-1}\left(1+\sqrt{2}+\sqrt{2k\mu_0}\right)\|v\|,$$

and it follows that Σ is bounded. From (i)-(iii) we deduce by the Schaefer fixed point theorem that (19)-(20) has a solution $u \in H^2(\Omega)\cap H_0^1(\Omega)$ and (21) follows, taking $\lambda = 1$ in (35). □

The following theorem shows the existence of a solution for the schemes (12) and (13).

Theorem 1. *Suppose that $u_0 \in H^2(\Omega)\cap H_0^1(\Omega)$. Then, for each nonnegative integer n, the scheme (12) and (13) has a solution in $H^2(\Omega)\cap H_0^1(\Omega)$.*

Proof. Proceeding by induction, the proof is immediate from Lemma 1 for a suitable choice of the functions u and v. For example, multiplying the first equation in (12) by $k/2$, we deduce (19)-(20) for $u = (u^{2,n+1}+u^{2,n})/2$, $v = ks(t_{n+1/2})/2 + u^{2,n}$ and k substituted by $k/2$. □

Hereafter we suppose that $u^{2j,n} \in H^{r+1}(\Omega)\cap H_0^1(\Omega)$, for $1 \le j \le p+1$ and each $n = 0, 1\cdots, N$. Convergence results for these semi-discrete solutions are proven in section 3.

3. Convergence and order of accuracy of the semi-discrete solution

The deferred correction condition (DCC) defined in [22] for ODEs applies to PDEs.

Definition 1. *Let u be the exact solution of (1). For a positive integer j, a sequence $\left\{u^{2j,n}\right\}_n \subset H_0^1(\Omega)$ of approximations of u on the uniform partition $0 = t_0 < t_1 < \cdots < t_N = T$, $t_n = nk$, is said to satisfy the Deferred Correction Condition (DCC) for the implicit midpoint rule if $\left\{u^{2j,n}\right\}_n$ approximates $u(t_n)$ with order $2j$ of accuracy in time, and for $n = 1, 2, ..., N-2$ we have*

$$\|(D_+D_-)D(u^{2j,n+1/2} - u(t_{n+1/2}))\| + \|D_+D_-(u^{2j,n+1} - u(t_{n+1}))\| \le Ck^{2j}, \tag{36}$$

for each time steps $k \le k_1$, where $k_1 > 0$ is fixed and C is a constant independent from k.

Remark 1. *Condition (36) is equivalent to*

$$\left\|\Gamma^j\left(u^{2j,n} - u(t_n)\right)\right\| \le Ck^{2j+2}, \tag{37}$$

and

$$\left\|(\Lambda^j - \Gamma^j)D\left(u^{2j,n+1/2} - u(t_{n+1/2})\right)\right\| \le Ck^{2j+2}, \tag{38}$$

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

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

and

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

The following theorem gives a sufficient condition for the semi-discrete schemes in time to converge with the expected order of accuracy.

Theorem 2. *Let j be a positive integer and $\left\{u^{2j,n}\right\}_n \subset H_0^1(\Omega)$ a sequence of approximations of u, on the discrete points $t_0=0<t_1<\cdots<t_N=T$, satisfying DCC for the implicit midpoint rule. Suppose that $k<k_1$, and that $u^{2j+2,1},...,u^{2j+2,j}$ are given and satisfy*

$$\|u^{2j+2,n}-u(t_n)\| \le Ck^{2j+2}, \quad \textit{for} \quad n=0,1,...,j. \tag{39}$$

Then the sequence $\left\{u^{2j+2,n}\right\}_{n\ge j}$, solution of the scheme (13) built from $\left\{u^{2j,n}\right\}_n$, approximates u with order $2j+2$ of accuracy in time, and we have, for $n=0,1,\cdots,N$,

$$\|u^{2j+2,n}-u(t_n)\|+\left(\gamma k\sum_{i=j}^{n}\|\nabla\widehat{\Theta}^{2j+2,i}\|^2\right)^{\frac{1}{2}} \le Ck^{2j+2}, \tag{40}$$

where

$$\Theta^{2j+2,n}=\left(u^{2j+2,n}-u(t_n)\right)-\Gamma^j\left(u^{2j,n}-u(t_n)\right), \tag{41}$$

and C is a constant depending only on j, T, M, $u\in C^{2j+3}\left([0,T],H^2(\Omega)\right)$, a Lipschitz constant on f and the DCC constant.

Proof. Combining (13) and (1), we obtain the identity

$$\begin{aligned} &D\Theta^{2j+2,n+1/2}+f\left(\widehat{u}^{2j+2,n+1}-\Gamma^j\widehat{u}^{2j,n+1}\right)-f\left(\widehat{u}(t_{n+1})-\Gamma^j\widehat{u}(t_{n+1})\right)\\ &-M\Delta\widehat{\Theta}^{2j+2,n+1}=\sigma^{2j+2,n+1/2}+(\Lambda^j-\Gamma^j)D(u^{2j,n+1/2}-u(t_{n+1/2})), \end{aligned} \tag{42}$$

where

$$\begin{aligned} \sigma^{2j+2,n+1/2}&=u'(t_{n+1/2})-Du(t_{n+1/2})+\Lambda^jDu(t_{n+1/2})+f(u(t_{n+1/2}))\\ &-f\left(\widehat{u}(t_{n+1})-\Gamma^j\widehat{u}(t_{n+1/2})\right)-M\Delta\left(u(t_{n+1/2})-\widehat{u}(t_{n+1})+\Gamma^j\widehat{u}(t_{n+1/2})\right). \end{aligned}$$

The inner product of (42) with $\widehat{\Theta}^{2j+2,n+1}$, taking into account the monotonicity condition (2) and the fact that $\widehat{\Theta}^{2j+2,n+1}=0$ on $\partial\Omega$, yields

$$\begin{aligned} &(D\Theta^{2j+2,n+\frac{1}{2}},\widehat{\Theta}^{2j+2,n+1})+\gamma\|\nabla\widehat{\Theta}^{2j+2,n+1}\|^2\le\tau(\widehat{u}(t_{n+1})-\Gamma^j\widehat{u}(t_{n+1}))\|\widehat{\Theta}^{2j+2,n+1}\|^2\\ &+\left(\sigma^{2j+2,n+1/2}+(\Lambda^j-\Gamma^j)D(u^{2j,n+1/2}-u(t_{n+1/2})),\widehat{\Theta}^{2j+2,n+1}\right). \end{aligned} \tag{43}$$

From the central finite differences (8)-(9) and the mean value theorem we have

$$\|\sigma^{2j+2,n+1/2}\|\le Ck^{2j+2},$$

where C is a constant depending only on a Lipschitz condition on f and the norm of u as element of $C^{2j+3}\left([0,T],H^2(\Omega)\right)$, and there exists $0<k_2\le k_1$ such that $k\le k_2$ implies that

$$\|\widehat{u}(t_{n+1})-\Gamma^j\widehat{u}(t_{n+1})\|_\infty\le\|u(t_{n+1/2})-\widehat{u}(t_{n+1})+\Gamma^j\widehat{u}(t_{n+1})\|_\infty+\|u(t_{n+1/2})\|_\infty\le 1+\|u\|_{L^\infty(Q_T)},$$

where $Q_T = \Omega \times (0,T)$. It follows that, for $k \leq k_2$,

$$\left\|\tau(\widehat{u}(t_{n+1}) - \Gamma^j \widehat{u}(t_{n+1}))\right\|_\infty \leq \max_{|y| \leq 1+\|u\|_{L^\infty(Q_T)}} |\tau(y)| =: \mu.$$

On the other hand, from the DCC we immediately have

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

Substituting the last inequalities in (43), taking into account the identity

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

we deduce that

$$\|\Theta^{2j+2,n+1}\|^2 - \|\Theta^{2j+2,n}\|^2 + 2k\gamma\|\nabla\widehat{\Theta}^{2j+2,n+1}\|^2 \leq Ck^{2j+3}\|\widehat{\Theta}^{2j+2,n+1}\| + 2k\mu\|\widehat{\Theta}^{2j+2,n+1}\|^2. \tag{44}$$

This inequality yields

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

and, for $\mu k < 2$, we deduce from the inequality

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

that

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

It follows by induction on n that

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

From the hypothesis (39) and the DCC we have

$$\left\|\Theta^{2j+2,j}\right\| \leq \left\|u^{2j+2,j} - u(t_j)\right\| + \left\|\Gamma^j\left(u^{2j,j} - u(t_j)\right)\right\| \leq Ck^{2j+2}, \tag{45}$$

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

$$\|\Theta^{2j+2,n}\| \leq Ck^{2j+2}. \tag{46}$$

Finally, by the triangle inequality, the identity (41) and the DCC, we have

$$\|u^{2j+2,n} - u(t_n)\| \leq Ck^{2j+2} + \left\|\Gamma^j(u^{2j,n} - u(t_n))\right\| \leq Ck^{2j+2}, \tag{47}$$

where C is a constant depending only on j, T, μ, M, a Lipschitz constant on f and u as element of $C^{2j+3}\left([0,T], H^2(\Omega)\right)$. Substituting (46) in (44), we have

$$\|\Theta^{2j+2,n+1}\|^2 - \|\Theta^{2j+2,n}\|^2 + 2k\gamma\|\nabla\widehat{\Theta}^{2j+2,n+1}\|^2 \leq Ck^{4j+5},$$

and it follows by induction, taking (45) into account, that

$$\|\Theta^{2j+2,n+1}\|^2+2k\gamma\sum_{i=j}^{n}\|\nabla\widehat{\Theta}^{2j+2,i}\|^2\leq Ck^{4j+4}.$$

Inequality (40) follows from (47) and the last inequality. □

To prove DCC for the schemes (12) and (13) we need the following lemma:

Lemma 2. *The sequence $\left\{u^{2,n}\right\}_n$ from the scheme (12) approximates u, the exact solution of (1), with order 2 of accuracy. Furthermore, if $u(.,t)=u_0$ for all $t\in[0,(2p+1)k_0]$, where k_0 is the initial time step defined in the introduction ($k_0\mu_0<2$), then we have*

$$\begin{aligned}\|D_-(D_+D_-)^m\Theta^{2,n+1}\|+\left\|(D_+D_-)^m\Theta^{2,n+1}\right\|+\left\|(D_+D_-)^m\widehat{\Theta}^{2,n+1}\right\|_2\\ +\left(\gamma k\sum_{i=m}^{n}\|\nabla(D_+D_-)^mD\widehat{\Theta}^{2,i+1/2}\|^2\right)^{1/2}\leq Ck^2,\end{aligned}\tag{48}$$

for $m=0,1,2,...,p$, $n=m,m+1,\cdots,N-m$, and $k\leq k_0$, where $\Theta^{2,n}=u^{2,n}-u(t_n)$, for $n=0,1,2,\cdots,N$, μ_0 is from (3), and C is a constant depending only on T, Ω, μ_0, k_0, M, the continuity of the source term S, the derivatives of f up to order $2m+2$, and the derivatives of u with respect to the time variable t up to order $2m+4$.

Proof. According to Theorem 2, it is immediate that the sequence $\left\{u^{2,n}\right\}_n$ from the scheme (12) approximates u with order 2 of accuracy in time, and

$$\|\Theta^{2,n}\|^2+\gamma k\sum_{i=0}^{n}\|\nabla\widehat{\Theta}^{2,i}\|^2\leq Ck^4,\ \text{for } n=0,1,\cdots,N,\tag{49}$$

where C is a constant depending only on T, Ω, a Lipschitz constant on f and the derivatives of $u\in C^3\left([0,T],H^2(\Omega)\right)$. To prove (48) we proceed by induction on the integer m.

1) The case $m=0$.

Combining (1) and (12), we obtain the identity

$$D\Theta^{2,n+1/2}-M\Delta\widehat{\Theta}^{2,n+1}+h(t_{n+1})=w^{2,n+1/2},\tag{50}$$

where

$$h(t_n)=f(\widehat{u}^{2,n})-f(\widehat{u}(t_n))=\int_0^1 df\left(K_1^n\right)(\widehat{\Theta}^{2,n})d\tau_1,$$

with

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

and

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

Applying D_+ to (50), we obtain

$$DD_+\Theta^{2,n+1/2}-M\Delta D_+\widehat{\Theta}^{2,n+1}+D_+h(t_{n+1})=D_+w^{2,n+1/2},$$

and the inner product of this identity with $D_+\widehat{\Theta}^{2,n+1}$ yields

$$\|D_+\Theta^{2,n+1}\|^2 - \|D_+\Theta^{2,n}\|^2 + 2\gamma k\|\nabla D_+\widehat{\Theta}^{2,n+1}\|^2 \leq 2k\left(-D_+h(t_{n+1}) + D_+w^{2,n+1/2}, D_+\widehat{\Theta}^{2,n+1}\right). \tag{51}$$

We can write

$$D_+h(t_n) = \int_0^1 df\left(K_1^{n+1}\right)(D_+\widehat{\Theta}^{2,n})d\tau_1 + \int_0^1\int_0^1 d^2f\left(K_2^n\right)\left(D_+K_1^n, \widehat{\Theta}^{2,n}\right)d\tau_1 d\tau_2, \tag{52}$$

where, for $n+i \leq N$, we have

$$K_{i+1}^n = K_i^n + \tau_{i+1}(K_i^{n+1} - K_i^n) = K_1^n + \sum_{l=1}^{i}\sum_{2\leq i_1<\cdots<i_l\leq i+1} \tau_{i_1}\cdots\tau_{i_l}k^l D_+^l K_1^n. \tag{53}$$

The scheme (12) can be transformed into equations (19)-(20), substituting k by $k/2$ and choosing $u = \widehat{u}^{2,n+1}$ and $v = (k/2)S(t_{n+1/2}) + u^{2,n}$. It follows from (21) and the triangle inequality that

$$\|\widehat{u}^{2,n}\|_2 \leq C\left(\|S(t_{n-\frac{1}{2}})\| + \|D_-\Theta^{2,n}\| + \|\nabla\widehat{\Theta}^{2,n}\| + \|D_-u(t_n)\| + \|\nabla\widehat{u}(t_n)\|\right),$$

where C is a constant depending only on Ω, the matrix M and μ_0. From inequalities (7), (49) and the Sobolev embedding $H^2(\Omega) \hookrightarrow C^0(\overline{\Omega})$, the last inequality implies the existence of a real $R > 0$, depending only on T, Ω, the regularity of S, the first derivative of f, and the second derivative of u with respect to t, such that

$$\|K_i^n\|_\infty \leq R, \text{ for } i = 1, 2, \cdots, 2p+1. \tag{54}$$

From the condition (3) we have

$$\left(df\left(K_1^n\right)(D_+\widehat{\Theta}^{2,n}), D_+\widehat{\Theta}^{2,n}\right) \geq -\mu_0\|D_+\widehat{\Theta}^{2,n}\|^2. \tag{55}$$

From (54) and (7) we have, for almost every $x \in \Omega$,

$$\begin{aligned}\left|d^2f\left(K_2^n\right)\left(D_+K_1^n, \widehat{\Theta}^{2,n+1}\right)(x)\right| &\leq \max_{|y|\leq R}\left|d^2f(y)\right| |D_+K_1^n(x)||\widehat{\Theta}^{2,n+1}(x)| \\ &\leq C\left(|\widehat{\Theta}^{2,n+1}(x)| + |D_+\widehat{\Theta}^{2,n+1}(x)||\widehat{\Theta}^{2,n+1}(x)|\right).\end{aligned}$$

Therefore,

$$\left\|d^2f\left(K_2^n\right)\left(D_+K_1^n, \widehat{\Theta}^{2,n+1}\right)\right\| \leq C\left(\|\widehat{\Theta}^{2,n+1}\| + \|D_+\widehat{\Theta}^{2,n+1}\|_{L^4(\Omega)}\|\widehat{\Theta}^{2,n+1}\|_{L^4(\Omega)}\right),$$

and we deduce from the Sobolev embedding $H_0^1(\Omega) \hookrightarrow L^4(\Omega)$ and the Poincaré inequality that

$$\left\|d^2f\left(K_2^n\right)\left(D_+K_1^n, \widehat{\Theta}^{2,n+1}\right)\right\| \leq C\left(\|\widehat{\Theta}^{2,n+1}\| + \|\nabla D_+\widehat{\Theta}^{2,n+1}\|\|\nabla\widehat{\Theta}^{2,n+1}\|\right). \tag{56}$$

This inequality and (55) together with the Cauchy-Schwartz inequality yield

$$\begin{aligned}&-k\left(D_+h(t_{n+1}), D_+\widehat{\Theta}^{2,n+1}\right) \leq k\mu_0\|D_+\widehat{\Theta}^{2,n+1}\|^2 + \frac{1}{2}\gamma k\|\nabla D_+\widehat{\Theta}^{2,n+1}\|^2 \\ &+ Ck\|D_+\widehat{\Theta}^{2,n+1}\|\left(\|\widehat{\Theta}^{2,n+1}\| + \|\nabla\widehat{\Theta}^{2,n+1}\|^2\|D_+\widehat{\Theta}^{2,n+1}\|\right),\end{aligned} \tag{57}$$

where we have used the Cauchy inequality with $\varepsilon = \gamma/2$:

$$\|\nabla D_+\widehat{\Theta}^{2,n+1}\|\|\nabla\widehat{\Theta}^{2,n+1}\|\|D_+\widehat{\Theta}^{2,n+1}\| \leq \frac{\gamma}{2}\|\nabla D_+\widehat{\Theta}^{2,n+1}\|^2 + \frac{1}{2\gamma}\|\nabla\widehat{\Theta}^{2,n+1}\|^2\|D_+\widehat{\Theta}^{2,n+1}\|^2.$$

According to (49), we have

$$\|\nabla\widehat{\Theta}^{2,n+1}\|^2\|D_+\widehat{\Theta}^{2,n+1}\| \leq k^{-1}\|\nabla\widehat{\Theta}^{2,n+1}\|^2\left(\|\widehat{\Theta}^{2,n+2}\| + \|\widehat{\Theta}^{2,n+1}\|\right) \leq Ck^4. \tag{58}$$

From Taylor's formula with integral remainder we can write

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

where, according to (7), we have

$$\|D_+^{m_1}D_-^{m_2}g(t_n)\| \leq C, \quad \text{for} \quad m_2 \leq n \leq N - m_1, \tag{59}$$

for each nonnegative integers m_1 and m_2 such that $m_1 + m_2 \leq 2p+1$. C is a constant depending only on T, the derivatives of f up to order $m_1 + m_2 + 1$, and the norm of u in $C^{m_1+m_2+3}\big([0,T], H^2(\Omega)\big)$. It follows from Cauchy-Schwartz inequality that

$$\left|\left(kD_+w^{2,n+1/2}, D_+\widehat{\Theta}^{2,n+1}\right)\right| \leq Ck^3\|D_+\widehat{\Theta}^{2,n+1}\|.$$

Substituting the last inequality and the inequality (57) in (51), taking (49) and (58) into account, we deduce that

$$\|D_+\Theta^{2,n+1}\|^2 - \|D_+\Theta^{2,n}\|^2 + \gamma k\|\nabla D_+\widehat{\Theta}^{2,n+1}\|^2 \quad \leq 2k\mu_0\|D_+\widehat{\Theta}^{2,n+1}\|^2 + Ck^3\|D_+\widehat{\Theta}^{2,n+1}\|, \tag{60}$$

where C is a constant depending only on T, Ω, S, the second derivative of f and $u \in C^4([0,T], H^2(\Omega))$. This inequality yields

$$\|D_+\Theta^{2,n+1}\| - \|D_+\Theta^{2,n}\| \leq k\mu_0\|D_+\widehat{\Theta}^{2,n+1}\| + Ck^3.$$

Since $k\mu_0 \leq k_0\mu_0 < 2$, it follows by induction that

$$\|D_+\Theta^{2,n}\| \leq Ck^2\left(\frac{2+k\mu_0}{2-k\mu_0}\right)^n + \left(\frac{2+k\mu_0}{2-k\mu_0}\right)^n\|D_+\Theta^{2,1}\|.$$

The condition $u(t_n) = u_0$, for $0 \leq t_n \leq (2p+1)k_0$, implies $\|D_+\Theta^{2,1}\| = 0$. Whence

$$\|D_-\Theta^{2,n}\| = \|D_+\Theta^{2,n-1}\| \leq Ck^2, \text{ for } n = 1, 2, \cdots, N. \tag{61}$$

Substituting (61) in the right hand side of (60), we deduce that

$$\|D_-\Theta^{2,n}\|^2 + \gamma k\sum_{l=0}^{n}\|\nabla D_+\widehat{\Theta}^{2,l}\|^2 \leq Ck^4. \tag{62}$$

On the other hand, by the elliptic regularity results applied to (50), we deduce from (54), (59) for $m_1 = m_2 = 0$, and (61) that

$$\|\widehat{\Theta}^{2,n+1}\|_2 \leq C\left(\|D_-\Theta^{2,n+1}\| + \|h(t_{n+1})\| + \|w^{2,n+1/2}\|\right) \leq Ck^2.$$

Inequality (48) for $m = 0$ holds from (49), (62) and the last inequality.

2) Inequality (48) for $m+1$, assuming that it holds for arbitrary $m \leq p-1$.

We apply $(D_+D_-)^{m+1}$ to the identity (50) and take the inner product of the resulting identity with $(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}$ to obtain, as in (51),

$$\begin{aligned}\|(D_+D_-)^{m+1}\Theta^{2,n+1}\|^2-\|(D_+D_-)^{m+1}\Theta^{2,n}\|^2+2\gamma k\|\nabla(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}\|^2\leq\\ 2k\left(-(D_+D_-)^{m+1}h(t_{n+1})+(D_+D_-)^{m+1}w^{2,n+1/2},(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}\right).\end{aligned}\tag{63}$$

As in [22] we can write

$$D_+^s h(t_n)=\sum_{i=1}^{s+1}\sum_{|\alpha_i|=s}L_{i,\alpha_i}^{n,s},\ \text{for } s=1,2,...,2p+1,\ \text{and } n\leq N-s,\tag{64}$$

where $\alpha_i=(\alpha_i^1,\cdots,\alpha_i^{i-1},\alpha_i^i)\in\{1,2,\cdots,s\}^{i-1}\times\{0,1,\cdots,s-i+1\}$. $L_{i,\alpha_i}^{n,s}$ is a linear combination of the quantities

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

where $\beta_i=(\beta_i^1,\cdots,\beta_i^{i-1},\beta_i^i)\in\{1,2,\cdots,s\}^{i-1}\times\{0,1,\cdots,s-i+1\}$ with $\beta_i^l+\alpha_i^l\leq s-l+1$, for $l=1,\cdots,i$, and $d\tau^i=d\tau_1\cdots d\tau_i$. From (54) and the regularity of f we have

$$\left\|d^if\left(K_i^n\right)\right\|_\infty\leq C_i,\quad\text{for } i=1,2,...,2p+1,0\leq n\leq N-i+1,\tag{65}$$

where C_i is a constant depending only on T, the i-th derivative of f and the second derivative of u. From the induction hypothesis (48), the Sobolev embedding $H^2(\Omega)\hookrightarrow L^\infty(\Omega)$, and inequality (7), we have

$$\|D_+^lK_i^n\|_\infty\leq C,\ \text{for } 1\leq l\leq 2m+2,0\leq n\leq N-i-l+1,\tag{66}$$

and

$$\|D_+^l\widehat{\Theta}^{2,n}\|\leq Ck^2,\ \text{for } 1\leq l\leq 2m+1,0\leq n\leq N-l.\tag{67}$$

- For $i=1$ we have

$$L_{1,\alpha_1}^{n,s}=\int_0^1 df(K_1^{n+s})\left(D_+^s\widehat{\Theta}^{2,n}\right)d\tau,$$

and, by taking $s=2m+2$, it follows from (3) that

$$\left(L_{1,\alpha_1}^{n-m,2m+2},(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}\right)\geq-\mu_0\|(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}\|^2\tag{68}$$

since

$$D_+^{2m+2}\widehat{\Theta}^{2,n-m}=(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}.$$

- For $i=2$ and $|\alpha_2|\leq 2m+2$, we have $1\leq\alpha_2^1\leq 2m+2$ and $0\leq\alpha_2^2\leq 2m+1$. It follows by the triangle inequality, the inequalities (7) and (65)-(67) that

$$\|L_{2,\alpha_2,\beta_2}^{n,s^*}\|\leq\left\|d^2f\left(K_2^{n+s^*-1}\right)\right\|_\infty\|D_+^{\alpha_2^1}K_1^{n+\beta_2^1}\|_\infty\|D_+^{\alpha_2^2}\widehat{\Theta}^{2,n}\|\leq Ck^2,\ \text{for } s^*\leq 2m+2.\tag{69}$$

- For $i\geq 3$ and $|\alpha_i|\leq 2m+3$, we have $1\leq\alpha_i^l\leq 2m+2$, for $l=1,2,\cdots,i-1$, and $0\leq\alpha_i^i\leq 2m+1$. It follows by the triangle inequality, the inequalities (7) and (65)-(67) that, for $s^*\leq 2m+3$,

$$\|L_{i,\alpha_i,\beta_i}^{n,s^*}\|\leq\|d^if(K_i^{n+s^*+1-i})\|_\infty\|D_+^{\alpha_i^i}\widehat{\Theta}^{2,n+\beta_i^i}\|\prod_{l=1}^{i-1}\|D_+^{\alpha_i^l}K_l^{n+\beta_i^l}\|_\infty\leq Ck^2.\tag{70}$$

From the identity (64), inequalities (68)-(70) yield

$$\left(-(D_+D_-)^{m+1}h(t_{n+1}),(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}\right)\leq \mu_0\|(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}\|^2+Ck^2\|(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}\|. \tag{71}$$

From inequality (59) we have

$$\|(D_+D_-)^{m+1}w^{2,n+1/2}\|\leq Ck^2. \tag{72}$$

Substituting (71) and (72) in (63), we obtain

$$\begin{aligned}&\|(D_+D_-)^{m+1}\Theta^{2,n+1}\|^2-\|(D_+D_-)^{m+1}\Theta^{2,n}\|^2+2\gamma k\|\nabla(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}\|^2\\&\leq 2k\mu_0\|(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}\|^2+Ck^3\|(D_+D_-)^{m+1}\widehat{\Theta}^{2,n+1}\|.\end{aligned} \tag{73}$$

Proceeding as in (60), we deduce by induction that

$$\|(D_+D_-)^{m+1}\Theta^{2,n}\|\leq \left(Ck^2+\|(D_+D_-)^{m+1}\Theta^{2,m+1}\|\right)\left(\frac{2+k\mu_0}{2-k\mu_0}\right)^{n-m-1}.$$

Since $u(t_n)=u_0$ for $0\leq t_n\leq (2p+1)k_0$, we have $\|(D_+D_-)^{m+1}\Theta^{2,m+1}\|=0$, for $m\leq p-1$. Whence

$$\|(D_+D_-)^{m+1}\Theta^{2,n}\|\leq Ck^2,\ \text{for } n=m+1,m+2,\cdots,N-m-1. \tag{74}$$

Substituting (74) in the right hand side of (73), we deduce by induction that

$$\|(D_+D_-)^{m+1}\Theta^{2,n}\|^2+2\gamma k\sum_{i=m+1}^{n}\|\nabla(D_+D_-)^{m+1}\widehat{\Theta}^{2,i}\|^2\leq Ck^4.$$

It is immediate from (65)-(67) that

$$\|L_{1,\alpha_1}^{n,2m+1}\|\leq \|df(K_1^{n+2m+1})\|_\infty\|D_+^{2m+1}\widehat{\Theta}^{2,n}\|\leq Ck^2.$$

Therefore, applying $D_-(D_+D_-)^m$ to (50), we deduce from the elliptic regularity inequality, the identity (64), the last inequality, the inequalities (69)-(70), (74) and (59) that

$$\|D_-(D_+D_-)^m\widehat{\Theta}^{2,n+1}\|_2\leq \|D_-(D_+D_-)^m\left(D\Theta^{2,n+1/2}+h(t_{n+1})+w^{2,n+1}\right)\|\leq Ck^2.$$

It follows that

$$\|(D_+D_-)^{m+1}\Theta^{2,n+1}\|+\left(\gamma k\sum_{i=m+1}^{n}\|\nabla(D_+D_-)^{m+1}\widehat{\Theta}^{2,i}\|^2\right)^{1/2}+\|D_-(D_+D_-)^m\widehat{\Theta}^{2,n+1}\|_2\leq Ck^2. \tag{75}$$

Otherwise, applying $D_+(D_+D_-)^{m+1}$ to (50), the same reasoning, taking the induction hypothesis and the inequality (75) into account, yields (48) for $m+1$. Finally, we deduce by induction that Lemma 2 is true for each $m=0,1,\cdots,p$. □

The following theorem shows DCC for the schemes (12) and (13) .

Theorem 3. *Suppose that the exact solution u of (1) satisfies $u(.,t)=u_0$ for each $t\in[0,(2p+1)k_0]$, where $k_0>0$ is a fixed real such that $k_0\mu_0<2$. Then, for $k\leq k_0$, each sequence $\left\{u^{2j,n}\right\}_n$, $j=1,2,...,p+1$, from the*

schemes (12) or (13) approximates u with order $2j$ of accuracy in time and we have the estimate

$$\begin{aligned}&\left\|(D_+D_-)^m\left(\widehat{u}^{2j,n+1}-\widehat{u}(t_{n+1})\right)\right\|_2+\sqrt{k\sum_{i=m}^{n}\|\nabla D_-(D_+D_-)^m\left(\widehat{u}^{2j,i}-\widehat{u}(t_i)\right)\|^2}\\&+\|D_-(D_+D_-)^m\left(u^{2j,n+1}-u(t_{n+1})\right)\|+\left\|(D_+D_-)^m\left(u^{2j,n+1}-u(t_{n+1})\right)\right\|\leq Ck^{2j}.\end{aligned}\tag{76}$$

for $m=0,1,...,p-j$ and $n=m+j-1,m+j,...,N-j-m$, where μ_0 is from (3), and C is a constant depending only on m, T, μ_0, k_0, M, the function S, and the derivatives of f and $u=u(t)$ up to order $2m+2j$ and $2m+2j+2$, respectively.

Proof. We proceed by induction on $j=1,2,...,p+1$, and the case $j=1$ results from Lemma 2. Suppose that $\left\{u^{2j,n}\right\}_n$ satisfies (76) up to an arbitrary order $j\leq p$. Let us prove that the theorem is still true for $j+1$.

Since $\left\{u^{2j,n}\right\}_n$ satisfies (76), it also satisfies DCC, and then Theorem 2 together with the condition $u(.,t)=u_0$ in $[0,(2p+1)k_0]$ implies that $\left\{u^{2j+2,n}\right\}_n$ approximates u with order $2j+2$ of accuracy in time. Therefore, it is enough to establish (76) for $j+1$. We can rewrite the identity (42) as follows

$$D\Theta^{2j+2,n+1/2}-M\Delta\widehat{\Theta}^{2j+2,n+1}+H(t_{n+1})=w^{2j+2,n+1/2},\tag{77}$$

where

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

and

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

Here $\Theta^{2j+2,n+1}$ and $\sigma^{2j+2,n+1/2}$ are as in Theorem 2. From the central finite difference (8)-(9) and the regularity of u with respect to t, we can write

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

where

$$\|D_+^{m_1}D_-^{m_2}G(t_n)\|\leq C,\ \text{ for }\ m_2\leq n\leq N-m_1,$$

for each nonnegative integers m_1 and m_2 such that $m_1+m_2\leq 2p-2j+1$. C is a constant depending only on T, the derivatives of f up to order m_1+m_2+2j+1 and the norm of u in $C^{m_1+m_2+2j+3}\left([0,T],H^2(\Omega)\right)$. On the other hand, from the induction hypothesis and Remark 1, we immediately have

$$\|D_-(D_+D_-)^m(\Lambda^j-\Gamma^j)(u^{2j,n}-u(t_n))\|\leq Ck^{2j+2},\ \text{for } m=0,1,...,p-(j+1).$$

The last two inequalities implies that

$$\|D_+^{m_1}D_-^{m_2}w^{2j+2,n+1/2}\|\leq Ck^{2j+2},\ \text{for } m_1+m_2\leq 2p-2j-2,$$

and $m_2+j\leq n\leq N-m_1-j-1$. Therefore, the reasoning from Lemma 2, substituting the functions h by H, $w^{2,n+1/2}$ by $w^{2j+2,n+1/2}$, $\widehat{\Theta}^{2,n+1}$ by $\widehat{\Theta}^{2j+2,n+1}$ and k^2 by k^{2j+2}, yields

$$\begin{aligned}\|D_-(D_+D_-)^m\Theta^{2j+2,n+1}\|+\left\|(D_+D_-)^m\Theta^{2j+2,n+1}\right\|+\left\|(D_+D_-)^m\widehat{\Theta}^{2j+2,n+1}\right\|_2\\+\left(k\sum_{i=m}^{n}\|\nabla D(D_+D_-)^m\widehat{\Theta}^{2j+2,i+1/2}\|^2\right)^{1/2}\leq Ck^{2j+2},\end{aligned}$$

for $m=0,1,...,p-(j+1)$, and (76) for $j+1$ follows by the triangle inequality. Inequality (76) then holds for arbitrary integer $j\leq p+1$. □

4. Fully discretized schemes and convergence results

Let S_h be a finite dimensional subspace of $H_0^1(\Omega)$ and $\{\phi_i\}_{i=1}^{N_h}$ a basis for S_h consisting in continuous piecewise polynomials of degree $r\geq 1$ (see for instance [24] for an introduction to finite element subspaces S_h; the integer r is related to the regularity of the exact solution of (1) in space). We suppose that there exist an interpolating operator I_h^r from $H_0^1(\Omega)$ onto S_h and a constant $c>0$ such that $0\leq l\leq r$ implies

$$\|v-I_h^r v\|+h\|\nabla\left(v-I_h^r v\right)\|\leq ch^{l+1}|v|_{l+1,2,\Omega},\ \forall v\in H^{l+1}(\Omega)\cap H_0^1(\Omega), \tag{78}$$

and

$$\|v-I_h^r v\|_{L^4(\Omega)}+h\|\nabla\left(v-I_h^r v\right)\|_{L^4(\Omega)}\leq ch^{l+1}|v|_{l+1,4,\Omega},\ \forall v\in W^{l+1,4}(\Omega)\cap H_0^1(\Omega), \tag{79}$$

where $|\cdot|_{l+1,\rho,\Omega}$ is the following seminorm in $W^{l+1,\rho}(\Omega)$:

$$|v|_{l+1,\rho,\Omega}=\sum_{|\alpha|=l+1}|\partial^\alpha v|_{L^\rho(\Omega)}.$$

We say that S_h satisfies the inverse inequality if

$$\|v_h\|_\infty\leq ch^{m-d/2}\|v_h\|_m,\ \forall v_h\in S_h,\ \text{and } m=0,1. \tag{80}$$

The estimates (78) and (79) hold when S_h is obtained from a shape-regular family of meshes $\{\mathcal{T}_h\}_{h>0}$ [24, Corollary 1.109 & 1.110] while (80) is due to [25, Theorem 3.2.6] or [24, Lemma 1.142] for a family of meshes $\{\mathcal{T}_h\}_{h>0}$ that is shape-regular and quasi-uniform. We consider the elliptic operator R_h, orthogonal projection of $H_0^1(\Omega)$ onto S_h with respect to the inner product $(v,w)\mapsto(M\nabla v,\nabla w)$. Proceeding as in [13, Theorem 1.1], we deduce from (78) that

$$\|R_h v-v\|+h\|\nabla(R_h v-v)\|\leq ch^{l+1}\|v\|_{H^{l+1}(\Omega)},\ \forall v\in H_0^1(\Omega)\cap H^{l+1}(\Omega), 0\leq l\leq r. \tag{81}$$

Furthermore, if S_h satisfies the inverse inequality (80), we deduce from (81) and (78) for $l=1$, and (79) for $l=0$ together with the continuous embedding $H^2(\Omega)\hookrightarrow W^{1,4}(\Omega)\hookrightarrow L^\infty(\Omega)$, that

$$\|R_h v\|_\infty\leq\|R_h v-I_h^r v\|_\infty+\|v-I_h^r v\|_\infty+\|v\|_\infty\leq ch^{1/2}\|v\|_2+C\|v\|_2, \tag{82}$$

for each $v\in H^2(\Omega)\cap H_0^1(\Omega)$.

For $j=0,1,2,\cdots,p$ and each positive integer $n\leq N$, we look for a function $u_h^{2j+2,n}\in H_0^1(\Omega)$ of the form

$$u_h^{2j+2,n}=\sum_{l=1}^{N_h}U_l^{2j+2,n}\phi_l, \tag{83}$$

satisfying

$$\begin{aligned}&\left(Du_h^{2j+2,n+1/2}-\Lambda^j Du_h^{2j,n+1/2},\phi\right)+\left(M\nabla\left(Eu_h^{2j+2,n+1/2}-\Gamma^j Eu_h^{2j,n+1/2}\right),\nabla\phi\right)\\&+\left(f\left(Eu_h^{2j+2,n+1/2}-\Gamma^j Eu_h^{2j,n+1/2}\right),\phi\right)=\left(s(t_{n+1/2}),\phi\right),\forall\phi\in S_h,\ \text{and } n\geq j\end{aligned} \tag{84}$$

$$u_h^{2j+2,0}=R_h u_0, \tag{85}$$

where $\Lambda^j D u_h^{2j,n+1/2} = \Gamma^j \widehat{u}_h^{2j,n+1/2} = 0$ if $j = 0$. The scheme (84)-(85), denoted DC(2j+2), constitutes a full discretization of the problem (1) with deferred correction in time, at the discrete points $0 = t_0 < t_1 < \cdots < t_N = T$, $t_n = nk$, and finite element in space. For the starting values in (84)-(85), $0 \le n \le j-1$, we consider the following scheme which is deduced from (16):

$$\begin{aligned} &\left(D u_h^{2j+2,n+1/2} - \frac{1}{2j+1} \bar{\Lambda}^j D \bar{u}_h^{2j,n_j+1/2} + f(\widehat{u}_h^{2j+2,n+1/2} - \bar{\Gamma}^j E \bar{u}_h^{2j,n_j+1/2}), \phi \right) \\ &+ \left(M \nabla \left(\widehat{u}_h^{2j+2,n+1/2} - \bar{\Gamma}^j E \bar{u}_h^{2j,n_j+1/2} \right), \nabla \phi \right) = \left(s(t_{n+1/2}), \phi \right), \forall \phi \in S_h, \end{aligned} \tag{86}$$

$$u_h^{2j+2,0} = R_h u_0, \tag{87}$$

The following theorem proves the existence of a solution for the schemes (84)-(85).

Theorem 4 (Existence of a solution for the fully discretized scheme)**.** *We suppose that $k|\tau(0)| < 2$. Then, for each $j = 1, 2, \cdots$, there exists a sequence $\left\{ u_h^{2j,n} \right\}_{n=0}^{N}$ of elements of the form (83) satisfying (84)-(85).*

To prove this theorem we need the following lemma which is an adaptation of the lemma on zeros of a vector field [8, p.493].

Lemma 3. *Let m be a positive integer and $v : \mathbb{R}^m \to \mathbb{R}^m$ a continuous function satisfying*

$$v(z) \cdot z \ge 0 \quad \text{if } \|z\|_* = R, \tag{88}$$

for a positive real R, where $\|.\|_*$ is an arbitrary norm on $\mathbb{R}^m$. Then there exists a point z in the closed ball

$$\overline{B}(0, R) = \{ z \in \mathbb{R}^m : \|z\|_* \le R \}$$

such that $v(z) = 0$.

Proof of Lemma 3. Suppose that $v(z) \neq 0$ for each $z \in \overline{B}(0, R)$. The mapping

$$\varphi \ : \ \overline{B}(0, R) \to \overline{B}(0, R)$$

defined by

$$\varphi(z) = -\frac{R}{\| v(z) \|_*} v(z)$$

is continuous. Since $\overline{B}(0, R)$ is a compact and convex subset of $\mathbb{R}^m$, we deduce from Schauder's fixed-point theorem [8, p.502] that φ has a fixed point $z \in \overline{B}(0, R)$. Therefore, $\|z\|_* = R$, and this leads to the contradiction

$$0 < |z|^2 = \varphi(z) \cdot z = -\frac{R}{\| v(z) \|_*} v(z) \cdot z \le 0.$$

□

Proof of Theorem 4. We proceed by double induction on $j = 1, 2, \cdots$ and $n = 0, 1, \cdots, N$, using Lemma 3 for the function $v \ : \mathbb{R}^{N_h} \to \mathbb{R}^{N_h}$ defined by

$$v^l(z) = \left(\frac{2z_h - 2a_h}{k}, \phi_l \right) + (M \nabla z_h, \nabla \phi_l) + \left(f(z_h) - s(t_{n+1/2}), \phi_l \right), \tag{89}$$

for $l = 1, 2 \cdots, N_h$, where $a_h \in S_h$ is fixed and z_h is the unique element of S_h associated to $z \in \mathbb{R}^{N_h}$ and defined by

$$z_h = \sum_{l=1}^{N_h} z_l \phi_l.$$

We take $\|z\|_* = \|z_h\|$. The function v is continuous. For $j = 1$, we have $u_h^{2,0} = R_h u_0$ and, supposing that $u_h^{2,n}$ exists for an arbitrary integer $n < N$ and taking $a_h = u_h^{2,n}$ in (89), we have

$$\begin{aligned} v(z) \cdot z &= \left(\frac{2z_h - 2u_h^{2,n}}{k}, z_h \right) + (M \nabla z_h, \nabla z_h) + \big(f(z_h) - s(t_{n+1/2}), z_h \big) \\ &\geq \frac{\|z_h\|}{k} \Big[(2 + k\tau(0)) \|z_h\| - 2\|u_h^{2,n}\| - k \left(\|f(0)\| + \|s(t_{n+1/2})\| \right) \Big] \\ &\geq 0, \end{aligned} \tag{90}$$

for

$$\|z\|_* = \frac{1}{2 + k\tau(0)} \left(1 + 2\|u_h^{2,n}\| + k\|s(t_{n+1/2})\| + k\|f(0)\| \right) := R.$$

Then, from Lemma 3, there exists a point z in the closed ball $\overline{B}(0, R)$ of $\left(\mathbb{R}^{N_h}, \|\cdot\|_* \right)$ such that $v(z) = 0$. Taking

$$U^{2,n+1} = \left(U_1^{2,n+1}, \cdots, U_{N_h}^{2,n+1} \right) = 2z - U^{2,n},$$

we have

$$v\left(\frac{U^{2,n+1} + U^{2,n}}{2} \right) \cdot e_l = 0,$$

for each e_l in the standard basis of $\mathbb{R}^{N_h}$. The last identity implies the existence of $u_h^{2,n+1}$ of the form (83) satisfying (84)-(85). Moreover, if $\left\{ u_h^{2j,n} \right\}_{n=0}^{N}$ exists and satisfies (84)-(85), for an arbitrary integer $j \geq 1$, then we have $u_h^{2j+2,0} = R_h u_0$, and the existence of $u_h^{2j+2,n+1}$ is immediate from the existence of $u_h^{2j+2,n}$, proceeding as in the case $j = 1$, taking $a_h = u_h^{2j+2,n} - \Gamma^j \widehat{u}^{2j,n+1} + 0.5 k \Lambda^j D u^{2j,n+1/2}$ in (89). □

The following theorem shows the convergence and order of accuracy of the fully discretized schemes.

Theorem 5 (Order of convergence of the fully discretized schemes)**.** *Suppose that the exact solution u of (1) is $C^{2p+4}\left([0,T], H^{r+1}(\Omega) \cap H_0^1(\Omega)\right)$ and satisfies $u(.,t) = u_0$ for $t \in [0, (2p+1)k_0]$, where p is a positive integer and $k_0 > 0$ is a real such that $k_0 \max\{\mu_0, \tau(0)\} < 2$, μ_0 and τ are defined in (2)-(3). In addition, suppose that S_h satisfies the inverse inequality (80). Then, for $j = 1, 2, \cdots, p+1$, the solution $\left\{ u_h^{2j,n} \right\}_{n=0}^{N}$ of the scheme (84)-(85) approximates u with order $2j$ of accuracy in time and order $r+1$ in space, that is*

$$\|u_h^{2j,n} - u(t_n)\| + h \left\| \nabla \left(u_h^{2j,n} - u(t_n) \right) \right\| \leq C(k^{2j} + h^{r+1}), \tag{91}$$

for $k < k_0$. Furthermore, we have the estimate

$$\|u_h^{2j,n} - R_h u^{2j,n}\|_1^2 + k \sum_{i=0}^{n} \|D(u_h^{2j,i+1/2} - R_h u^{2j,i+1/2})\|^2 + 2\alpha k \sum_{i=0}^{n} \|u_h^{2j,i} - R_h u^{2j,i}\|_{L^q(\Omega)}^q \leq C h^{2r+2}, \tag{92}$$

where C is a constant depending only on j, T, Ω, M, k_0, μ_0 and the derivatives of S, f and u.

Proof. Inequality (91) is immediate from (92) by quadruple triangle inequality, writing

$$u_h^{2j,n}-u(t_n)=\left(u_h^{2j,n}-R_hu^{2j,n}\right)-[u(t_n)-u^{2j,n}]-[u(t_n)-R_hu(t_n)]+\left[u(t_n)-u^{2j,n}-R_h(u(t_n)-u^{2j,n})\right],$$

and taking (81) and (76) into account. Therefore, we just need to establish (92). We proceed by induction on $j=1,2,\cdots,p+1$. For this purpose, we need the following claim which proof is a straightforward application of the mean value theorem, the triangle inequality, and inequalities (76), (81)-(82).

Claim 1. *There exist $0<k_3\le k_0$ and $h_1>0$ such that $k\le k_3$ and $h\le h_1$ imply,*

$$\|R_h(\widehat{u}^{2j+2,n+1}-\Gamma^j\widehat{u}^{2j,n+1})\|_\infty\le 1+C\|u\|_{L^\infty(0,T,H^2(\Omega))},\tag{93}$$

and

$$\|w_h^{2j+2,n+1/2}\|\le Ch^{r+1},\tag{94}$$

for each $j=0,1,\cdots,p$, and $n=0,1,\cdots,N$, where we define

$$\begin{aligned}w_h^{2j+2,n+1/2}&=f\left(\widehat{u}^{2j+2,n+1}-\Gamma^j\widehat{u}^{2j,n+1}\right)-f\left(R_h(\widehat{u}^{2j+2,n+1}-\Gamma^j\widehat{u}^{2j,n+1})\right)\\&+D\left(u^{2j+2,n+1/2}-\Lambda^ju^{2j,n+1/2}\right)-R_hD\left(u^{2j+2,n+1/2}-\Lambda^ju^{2j,n+1/2}\right),\end{aligned}\tag{95}$$

and we set $u^{0,n}=0$.

1. The case $j=1$. We proceed in two steps:

(i) First, we are going to prove the inequality

$$\|u_h^{2,n}-R_hu^{2,n}\|^2+2\gamma k\sum_{i=0}^n\|\nabla E(u_h^{2,i+1/2}-R_hu^{2,i+1/2})\|^2+2\alpha k\sum_{i=0}^n\|E(u_h^{2,i+1/2}-R_hu^{2,i+1/2})\|_{L^q(\Omega)}^q\le Ch^{2r+2}.\tag{96}$$

The scheme (12) yields

$$\left(Du^{2,n+1/2},\phi\right)+\left(M\nabla\widehat{u}^{2,n+1},\nabla\phi\right)+\int_\Omega f(\widehat{u}^{2,n+1})\phi dx=\left(s(t_{n+1/2}),\phi\right),\ \forall\phi\in S_h.$$

Therefore, combining this identity and (84), for $j=0$, we deduce that

$$\begin{aligned}&\left(D\Theta_h^{2,n+1/2},\phi\right)+\left(M\nabla\widehat{\Theta}_h^{2,n+1},\nabla\phi\right)+\int_\Omega\left(f(\widehat{u}_h^{2,n+1})-f(R_h\widehat{u}^{2,n+1})\right)\phi dx\\&=\left(w_h^{2,n+1/2},\phi\right)+\left(M\nabla\left(\widehat{u}^{2,n+1}-R_h\widehat{u}^{2,n+1}\right),\nabla\phi\right),\ \forall\phi\in S_h,\end{aligned}\tag{97}$$

where

$$\Theta_h^{2,n}=u_h^{2,n}-R_hu^{2,n},$$

and $w_h^{2,n+1/2}$ is defined in (95). Hypothesis (2) and inequality (93) yield

$$\int_\Omega\left(f(\widehat{u}_h^{2,n+1})-f(R_h\widehat{u}^{2,n+1})\right)\widehat{\Theta}_h^{2,n+1}dx\ge\alpha\|\widehat{\Theta}_h^{2,n+1}\|_{L^q(\Omega)}^q-\mu\|\widehat{\Theta}_h^{2,n+1}\|^2,\tag{98}$$

where

$$\mu=\max_{|y|\le 1+\|u\|_{L^\infty(0,T;H^2(\Omega))}}|\tau(y)|.$$

From the properties of orthogonal projection we have,

$$\left(M\nabla\left(\widehat{u}^{2,n+1}-R_h\widehat{u}^{2,n+1}\right),\nabla\phi\right)=0,\forall\phi\in S_h. \tag{99}$$

Therefore, choosing $\phi=\widehat{\Theta}_h^{2,n+1}$ in (97), we deduce from the Cauchy-Schwartz inequality and the inequalities (94) and (98) that

$$\left(D\Theta_h^{2,n+1/2},\widehat{\Theta}_h^{2,n+1}\right)+\gamma\|\nabla\widehat{\Theta}_h^{2,n+1}\|^2+\alpha\|\widehat{\Theta}_h^{2,n+1}\|_{L^q(\Omega)}^q\leq Ch^{r+1}\|\widehat{\Theta}_h^{2,n+1}\|+\mu\|\widehat{\Theta}_h^{2,n+1}\|^2, \tag{100}$$

for $0<k\leq k_3$ and $0<h\leq h_1$. This inequality yields

$$\left(D\Theta_h^{2,n+1/2},\widehat{\Theta}_h^{2,n+1}\right)\leq Ch^{r+1}\left\|\widehat{\Theta}_h^{2,n+1}\right\|+\mu\left\|\widehat{\Theta}_h^{2,n+1}\right\|^2,$$

and it follows for $0<k\mu\leq k_3\mu<2$ that

$$\left\|\Theta_h^{2,n+1}\right\|\leq C\frac{k}{2-k\mu}h^{r+1}+\frac{2+k\mu}{2-k\mu}\left\|\Theta_h^{2,n}\right\|.$$

Proceeding by induction as in Theorem 2, the last inequality yields

$$\left\|\Theta_h^{2,n}\right\|\leq\left(nkCh^{r+1}+\left\|\Theta_h^{2,0}\right\|\right)\left(\frac{2+k\mu}{2-k\mu}\right)^n\leq Ch^{r+1} \tag{101}$$

since $nk\leq T$ and $\Theta_h^{2,0}=0$. Inequality (96) follows by substituting (101) in (100).

(ii) Now we are going to prove the inequality

$$k\sum_{i=0}^{n}\left\|D\Theta_h^{2,n+1/2}\right\|^2+\gamma\|\nabla\Theta_h^{2,n+1}\|^2\leq Ch^{2r+2}. \tag{102}$$

We choose $\phi=D\Theta_h^{2,n+1/2}$ in (97) and obtain

$$\begin{aligned}&\int_\Omega\left(f(\widehat{u}_h^{2,n+1})-f(R_h\widehat{u}^{2,n+1})\right)D\Theta_h^{2,n+1/2}dx+\left(M\nabla\widehat{\Theta}_h^{2,n+1},\nabla D\Theta_h^{2,n+1/2}\right)\\&+\left\|D\Theta_h^{2,n+1/2}\right\|^2=\left(w_h^{2,n+1/2},D\Theta_h^{2,n+1/2}\right).\end{aligned} \tag{103}$$

We can write

$$f(\widehat{u}_h^{2,n+1})-f(R_h\widehat{u}^{2,n+1})=\int_0^1 df\left(R_h\widehat{u}^{2,n+1}+\xi\widehat{\Theta}_h^{2,n+1}\right)\left(\widehat{\Theta}_h^{2,n+1}\right)d\xi.$$

From the inverse inequality (80) and the inequality (101), we have

$$\|\widehat{\Theta}_h^{2,n+1}\|_\infty\leq ch^{-3/2}\|\Theta_h^{2,n}\|\leq Ch^{r-1/2},\ r\geq 1. \tag{104}$$

This inequality together with (93) implies that there exists $0<h_2\leq h_1$ such that, for $0<h\leq h_2$, we have

$$\|R_h\widehat{u}^{2,n+1}+\xi\widehat{\Theta}_h^{2,n+1}\|_\infty\leq 2+\|u\|_{L^\infty\left(0,T;H^2(\Omega)\right)}.$$

The last identity yields

$$\left\|f(\widehat{u}_h^{2,n+1})-f(R_h\widehat{u}^{2,n+1})\right\| \leq \max_{|y|\leq 2+\|u\|_{L^\infty(0,T;H^2(\Omega))}} |df(y)| \left\|\widehat{\Theta}_h^{2,n+1}\right\| \leq C\left\|\widehat{\Theta}_h^{2,n+1}\right\|. \tag{105}$$

Substituting (105) in (103), we deduce by Cauchy-Schwartz inequality and (94) that

$$k\|D\Theta_h^{2,n+1/2}\|^2+\left(M\nabla\Theta_h^{2,n+1},\nabla\Theta_h^{2,n+1}\right)-\left(M\nabla\Theta_h^{2,n},\nabla\Theta_h^{2,n}\right)\leq Ckh^{2r+2},$$

for $n=0,1,\cdots,N-1$. It follows the inequality

$$k\sum_{i=0}^{n}\left\|D\Theta_h^{2,n+1/2}\right\|^2+\left(M\nabla\Theta_h^{2,n+1},\nabla\Theta_h^{2,n+1}\right)\leq Cnkh^{2r+2}$$

since $\Theta_h^{2,0}=0$. The last inequality gives exactly (102), where C is a constant depending only on T, Ω, k_{i+1}, h_i, $i=1,2$, and the derivatives of f and u.

Estimates (96) and (102) gives (92) for $j=1$.

2. Here we prove inequality (92) for $j+1$, assuming that it holds up to order j, $1\leq j\leq p$.

From the scheme (13) we have

$$\begin{aligned}&\left(Du^{2j+2,n+1/2}-\Lambda^j Du^{2j,n+1/2},\phi\right)+\left(M\nabla\left(\widehat{u}^{2j+2,n+1}-\Gamma^j\widehat{u}^{2j,n+1}\right),\nabla\phi\right)\\&+\int_\Omega f\left(\widehat{u}^{2j+2,n+1}-\Gamma^j\widehat{u}^{2j,n+1}\right)\phi dx=\left(s(t_{n+1/2}),\phi\right),\ \forall\phi\in S_h.\end{aligned}\tag{106}$$

Combining this identity and (84), we deduce that

$$\begin{aligned}&\left(D\Theta_h^{2j+2,n+1/2}+f\left(\widehat{u}_h^{2j+2,n+1}-\Gamma^j\widehat{u}_h^{2j,n+1}\right)-f\left(R_h(\widehat{u}^{2j+2,n+1}-\Gamma^j\widehat{u}^{2j,n+1}\right),\phi\right)\\&+\left(M\nabla\widehat{\Theta}_h^{2j+2,n+1},\nabla\phi\right)=\left(w_h^{2j+2,n+1/2}+(\Lambda^j-\Gamma^j)D(u_h^{2j,n+1/2}-R_hu^{2j,n+1/2}),\phi\right),\end{aligned}\tag{107}$$

for any $\phi\in S_h$, where we define

$$\Theta_h^{2j+2,n}=u_h^{2j+2,n}-R_hu^{2j+2,n}-\Gamma^j(u_h^{2j,n}-R_hu^{2j,n}),$$

and we use the identity

$$\left(M\nabla(Id-R_h)\left(\widehat{u}^{2j+2,n+1}-\Gamma^j\widehat{u}^{2j,n+1}\right),\nabla\phi\right)=0,\forall\phi\in S_h.\tag{108}$$

Id denotes the identity application. As in (98) we have

$$\begin{aligned}&\int_\Omega\left(f\left(\widehat{u}_h^{2j+2,n+1}-\Gamma^j\widehat{u}_h^{2j,n+1}\right)-f\left(R_h(\widehat{u}^{2j+2,n+1}-\Gamma^j\widehat{u}^{2j,n+1})\right)\right)\widehat{\Theta}_h^{2j+2,n+1}dx\\&\quad\geq\alpha\|\widehat{\Theta}_h^{2j+2,n+1}\|_{L^q(\Omega)}^q-\mu\|\widehat{\Theta}_h^{2j+2,n+1}\|^2.\end{aligned}$$

Therefore, choosing $\phi=\widehat{\Theta}_h^{2j+2,n+1}$ in (107), we deduce by the triangle inequality, the last inequality and (94) that

$$\begin{aligned}&(D\Theta_h^{2j+2,n+1/2},\widehat{\Theta}_h^{2j+2,n+1})+\gamma\|\nabla\widehat{\Theta}_h^{2j+2,n+1}\|^2+\alpha\|\widehat{\Theta}_h^{2,n+1}\|_{L^q(\Omega)}^q\leq\mu\|\widehat{\Theta}_h^{2j+2,n+1}\|^2+\\&\left(Ch^{r+1}+\|(\Lambda^j-\Gamma^j)D_-(u_h^{2j,n+1}-R_hu^{2j,n+1})\|\right)\|\widehat{\Theta}_h^{2j+2,n+1}\|\end{aligned}\tag{109}$$

This inequality implies that

$$\|\Theta_h^{2j+2,n+1}\| - \|\Theta_h^{2j+2,n}\| \le k\mu \|\widehat{\Theta}_h^{2j+2,n+1}\| + k\left(Ch^{r+1} + \left\|(\Lambda^j - \Gamma^j)D\left(u_h^{2j,n+1/2} - R_h u^{2j,n+1/2}\right)\right\|\right), \quad (110)$$

and we deduce, for $k\mu < 2$, that

$$\|\Theta_h^{2j+2,n+1}\| \le \frac{k}{2-k\mu}\left(Ch^{r+1} + \left\|(\Lambda^j - \Gamma^j)D\left(u_h^{2j,n+1/2} - R_h u^{2j,n+1/2}\right)\right\|\right) + \frac{2+k\mu}{2-k\mu}\left\|\Theta_h^{2j+2,n}\right\|.$$

It follows by induction that,

$$\begin{aligned}\|\Theta_h^{2j+2,n+1}\| \le & C\left(\frac{2+k\mu}{2-k\mu}\right)^{n-j}\left(h^{r+1} + \|\Theta_h^{2j+2,j}\|\right) \\ & + k\left(\frac{2+k\mu}{2-k\mu}\right)^{n-j}\sum_{m=j}^{n}\|(\Lambda^j - \Gamma^j)D(u_h^{2j,m+1/2} - R_h u^{2j,m+1/2})\|,\end{aligned} \quad (111)$$

for $n \ge j$, and for $0 \le n \le j-1$ we have

$$\begin{aligned}&\|\bar{\Theta}_h^{2j+2,n+1}\| \le C\left(\frac{2+k\mu}{2-k\mu}\right)^{n}\left(h^{r+1} + \|\bar{\Theta}_h^{2j+2,0}\|\right) \\ &+ k\left(\frac{2+k\mu}{2-k\mu}\right)^{n}\sum_{m=0}^{j}\left\|(\bar{\Lambda}^j - \bar{\Gamma}^j)D\left(\bar{u}_h^{2j,(2j+1)m+j+1/2} - R_h\bar{u}^{2j,(2j+1)m+j+1/2}\right)\right\|,\end{aligned} \quad (112)$$

where we define

$$\bar{\Theta}_h^{2j+2,n} = u_h^{2j+2,n} - R_h u^{2j+2,n} - \bar{\Gamma}^j\left(\bar{u}_h^{2j,(2j+1)n+j+1} - R_h\bar{u}^{2j,(2j+1)n+j+1}\right).$$

Since $\left\{u_h^{2j,n}\right\}_{n=0}^{N}$ and $\left\{\bar{u}_h^{2j,m}\right\}_{m=0}^{j}$ are obtained from the same scheme, but for different time steps k and $k_j = k/(2j+1)$, respectively, as for $\left\{u^{2j,n}\right\}_{n=0}^{N}$ and $\left\{\bar{u}^{2j,m}\right\}_{m=0}^{j}$, we deduce from the induction hypothesis and the formulae (17) and (18) that

$$\|\bar{\Theta}_h^{2j+2,0}\|_1 = \|\bar{\Gamma}^j\left(\bar{u}_h^{2j,j+1} - R_h\bar{u}^{2j,j+1}\right)\|_1 \le C\sum_{m=0}^{2j}\|\bar{u}_h^{2j,m} - R_h\bar{u}^{2j,m}\|_1 \le Ch^{r+1}, \quad (113)$$

and

$$\begin{aligned}&k\sum_{m=0}^{j}\left\|(\bar{\Lambda}^j - \bar{\Gamma}^j)D\left(\bar{u}_h^{2j,(2j+1)m+j+1/2} - R_h\bar{u}^{2j,(2j+1)m+j+1/2}\right)\right\| \\ &\le C\sqrt{k\sum_{m=0}^{2j^2+3j}\|D(\bar{u}_h^{2j,m+1/2} - R_h\bar{u}^{2j,m+1/2})\|^2} \le Ch^{r+1}.\end{aligned}$$

Substituting the last two inequalities in (112), we deduce that

$$\|\bar{\Theta}_h^{2j+2,n}\| \le Ch^{r+1}, \text{ for } 0 \le n \le j,$$

and it follows by the triangle inequality and the induction hypothesis that

$$\|u_h^{2j+2,n} - R_h u^{2j+2,n}\| \le Ch^{r+1}, \text{ for } 0 \le n \le j. \quad (114)$$

By the triangle inequality and the induction hypothesis, (114) in turn yields

$$\|\Theta_h^{2j+2,j}\| \le Ch^{r+1},$$

and we have from (14) and (15)

$$k\sum_{m=j}^{n}\|(\Lambda^j-\Gamma^j)D(u_h^{2j,m+1/2}-R_hu^{2j,m+1/2})\| \le C\sqrt{nk}\sqrt{k\sum_{m=0}^{n+j}\|D(u_h^{2j,m+1/2}-R_hu^{2j,m+1/2})\|^2} \le Ch^{r+1}.$$

The last two inequalities and (114) substituted in (111) yields

$$\|\Theta_h^{2j+2,n}\| \le Ch^{r+1}, \text{ for } j \le n \le N, \tag{115}$$

and it follows from (109) and (114) that

$$\|u_h^{2j+2,n}-R_hu^{2j+2,n}\|^2+2\alpha k\sum_{i=0}^{n}\|\widehat{u}_h^{2j+2,i}-R_h\widehat{u}^{2j+2,i}\|_{L^q(\Omega)}^q \le Ch^{2r+2}. \tag{116}$$

Otherwise, proceeding as in the step 1-(ii) of this proof, we choose $\phi = D\Theta_h^{2j+2,n+1/2}$ in (107) and deduce from (115) that

$$k\sum_{i=j}^{n}\left\|D\Theta_h^{2j+2,i+1/2}\right\|^2+\gamma\left\|\nabla\Theta_h^{2j+2,n+1}\right\|^2 \le Ch^{2r+2}+\left(M\nabla\Theta_h^{2j+2,j},\nabla\Theta_h^{2j+2,j}\right), \tag{117}$$

for $j \le n \le N$, and, for $0 \le n \le j-1$,

$$k\sum_{i=0}^{j}\left\|D\bar{\Theta}_h^{2j+2,i+1/2}\right\|^2+\gamma\left\|\nabla\bar{\Theta}_h^{2j+2,n+1}\right\|^2 \le Ch^{2r+2} \tag{118}$$

since, from Cauchy-Schwartz inequality and (113), we have

$$\left|\left(M\nabla\bar{\Theta}_h^{2j+2,0},\nabla\bar{\Theta}_h^{2j+2,0}\right)\right| \le |||M|||\nabla\bar{\Theta}_h^{2j+2,0}\|^2 \le Ch^{2r+2}.$$

By the triangle inequality and the induction hypothesis, inequality (118) for $n = j-1$ yields

$$\left|\left(M\nabla\Theta_h^{2j+2,j},\nabla\Theta_h^{2j+2,j}\right)\right| \le |||M|||\nabla\Theta_h^{2j+2,j}\|^2 \le Ch^{2r+2}.$$

Substituting the last identity in (117), we deduce from (118), the induction hypothesis, and the triangle inequality that

$$k\sum_{i=0}^{n}\|D(\widehat{u}_h^{2j+2,i+1/2}-R_h\widehat{u}^{2j+2,i+1/2})\|^2+\gamma\|\nabla\left(\widehat{u}_h^{2j+2,n}-R_h\widehat{u}^{2j+2,n}\right)\|^2 \le Ch^{2r+2}, \tag{119}$$

for $0 \le n \le N-1$, where C is a constant depending only on j, T, Ω, M, and the derivatives of f and u. Inequality (92) for the case $j+1$ follows from (116) and (119). Therefore, we can conclude by induction that the Theorem holds for $1 \le j \le p+1$. □

Corollary 1. *Under the conditions of Theorem 5, if S_h does not satisfy the inverse inequality, provided that, in addition to conditions (2) and (3), f satisfies the inequality*

$$|f(x)-f(y)| \le C\left(|x-y|+|x-y|^{q-1}\right), \text{ for each } x,y \in \mathbb{R}^J, \tag{120}$$

then the solution $\left\{u_h^{2j,n}\right\}_{n=0}^{N}$, $1 \le j \le p+1$, of the scheme (84)-(85) satisfies

$$\|u_h^{2j,n} - u(t_n)\| \le C(h^r + k^{2j}), \quad \forall n = 0,1,...,N, k < k_0. \tag{121}$$

Furthermore, we have the estimate

$$\|u_h^{2j,n} - I_h^r u^{2j,n}\|_1^2 + k\sum_{i=0}^{n}\|D(u_h^{2j,i} - I_h^r u^{2j,i})\|^2 + 2\alpha k\sum_{i=0}^{n}\|u_h^{2j,i} - I_h^r u^{2j,i}\|_{L^q(\Omega)}^q \le Ch^{2r} \tag{122}$$

where C is a constant depending only on j, T, Ω, M, k_0, μ_0, and the derivatives of S, f and u.

Proof. Inequality (122) is deduced from Theorem 5 substituting the elliptic operator R_h by the interpolating operator I_h^r. By this substitution, the corresponding Claim 1 is obtained from (78) and (79). Since (104) does not hold without inverse inequality, (105) is replaced by the inequality

$$\left|\int_\Omega \left(f(\widehat{u}_h^{2,n+1}) - f(I_h^r\widehat{u}^{2,n+1})\right)\left(\widehat{u}_h^{2,n+1} - I_h^r\widehat{u}^{2,n+1}\right)dx\right| \le C\left(\|\widehat{u}_h^{2,n+1} - I_h^r\widehat{u}^{2,n+1}\|^2 + \|\widehat{u}_h^{2,n+1} - I_h^r\widehat{u}^{2,n+1}\|_{L^q(\Omega)}^q\right),$$

owing to the hypothesis (120). The order of accuracy in space is reduced since, instead of identities (99) and (108), we have

$$\left|\left(M\nabla(Id - I_h^r)(\widehat{u}^{2j+2,n+1} - \Gamma^j\widehat{u}^{2j,n+1}), \nabla\phi\right)\right| \le C\|\nabla(Id - I_h^r)(\widehat{u}^{2j+2,n+1} - \Gamma^j\widehat{u}^{2j,n+1})\|\|\nabla\phi\| \le Ch^r\|\nabla\phi\|,$$

for each $\phi \in S_h$. □

5. Numerical experiment

For the numerical experiment we consider the bistable reaction-diffusion equation

$$\begin{aligned} u_t - u_{xx} + 10^4 u(u-1)(u-0.25) &= 0 \text{ in } \Omega\times(0,T), \\ \frac{\partial u}{\partial n} &= 0 \text{ on } \partial\Omega\times(0,T), \\ u(\cdot,0) &= e^{-100x^2} \text{ in } \Omega. \end{aligned} \tag{123}$$

We choose $\Omega = (0,1)$ and $T = 0.0295$. We are interested in the order of convergence in time. For this purpose, we simply use P_1 Lagrange finite elements in space with uniform mesh and the step $h = 10^{-3}$. We compute a reference solution using DC10 with the time step $k = 1.64\times10^{-5}$ (N=1800). Table 3 gives the maximal absolute error in time, norm $L^2(\Omega)$ in space, and the order of convergence for each pair of consecutive time steps.

For this problem, we have

$$f(u) = 10^4 u(u-1)(u-0.25),$$

and inequalities (2) and (3) hold with $\tau(0) = -1500$ and $\mu_0 = 8125/3$. Therefore, according to Theorem 5, the maximal time step to solve the problem with the DC methods is $k_0 = 6/8125 \simeq 7.38\times10^{-4}$, that is $N = 39.9479 \simeq 40$.

For the computational effort of the DC methods, we recall that to compute an approximate solution at the discrete points $0 = t_0 < t_1 < \cdots < t_N = T$, $DC2$ solves N nonlinear systems while $DC2j$, $j \ge 2$, solves

$j \times N$ systems. For the bistable reaction-diffusion, it is clear that, for $N > 180$, higher order DC method have the smallest maximal error by solving less systems of equations. For exemple, $DC10$ achieves an absolute error of about 2.48×10^{-15} by solving approximately 2250 while $DC4$ achieves almost the same accuracy by solving 3600 nonlinear systems. $DC10$, $DC8$, $DC4$ and $DC2$ solve approximately 1800 nonlinear systems, but the corresponding errors are, respectively, 8.57×10^{-14}, 2.4×10^{-14}, 5.63×10^{-13} and 6.25×10^{-9}. Since the resolution of nonlinear systems is the main burden for these methods, using high order DC methods is advantageous.

TABLE 3. Absolute error (order of convergence) for the bistable reaction-diffusion equation

N	DC2	DC4	DC6	DC8	DC10
40	0.115	4.62e-03	9.14e-04	1.97e-04	1.11e-03
90	8.48e-04(3.21)	4.59e-05(5.68)	2.05e-06(7.52)	1.55e-06(5.97)	1.45e-06(8.22)
180	5.91e-05(3.84)	2.17e-06(7.72)	5.53e-09(8.53)	4.09e-09(8.56)	1.90e-09(9.57)
360	3.87e-06(3.93)	8.59e-10(7.98)	2.57e-12(11.07)	4.51e-13(13.15)	8.57e-14(14.44)
450	1.55e-06(3.96)	1.44e-10(8.01)	2.33e-13(10.74)	2.40e-14(13.14)	2.48e-15(15.88)
900	9.97e-08(4.00)	5.63e-13(7.99)	2.67e-16(9.77)	8.62e-19(14.75)	7.36e-21(18.36)
1800	6.25e-09(3.99)	2.18e-15(8.00)	2.13e-19(10.29)	1.74e-22(12.27)	–

ACKNOWLEDGEMENTS

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

REFERENCES

[1] J. Smoller, Shock waves and reaction-diffusion equations, 2nd Edition, Vol. 258, Springer-Verlag, New York, 1994.
[2] Y. B. Pesin, A. A. Yurchenko, Some physical models described by the reaction-diffusion equation, and coupled map lattices, Uspekhi Mat. Nauk 59 (2004) 81–114.
[3] A. C. Newell, J. A. Whitehead, Finite bandwidth, finite amplitude convection, J. Fluid Mech. 38.
[4] P. A. Markowich, Applied partial differential equations: A visual approach, Springer, Berlin, 2007.
[5] V. Volpert, Elliptic partial differential equations, Vol. 104, Birkhäuser/Springer Basel AG, Basel, 2014.
[6] J.-L. Lions, Quelques méthodes de résolution des problèmes aux limites non linéaires, Dunod; Gauthier-Villars, Paris, 1969.
[7] R. Temam, Infinite-dimensional dynamical systems in mechanics and physics, 2nd Edition, Vol. 68, Springer-Verlag, New York, 1997.
[8] L. C. Evans, Partial differential equations, Vol. 19, American Mathematical Society, Providence, RI, 1998.
[9] G. Akrivis, Stability of implicit-explicit backward difference formulas for nonlinear parabolic equations, SIAM J. Numer. Anal. 53 (2015) 464–484.
[10] G. Akrivis, M. Crouzeix, Linearly implicit methods for nonlinear parabolic equations, Math. Comp. 73 (2004) 613–635.
[11] G. Akrivis, M. Crouzeix, C. Makridakis, Implicit-explicit multistep finite element methods for nonlinear parabolic problems, Math. Comp. 67 (1998) 457–477.
[12] M. Zlámal, Finite element methods for nonlinear parabolic equations, RAIRO Anal. Numér. 11 (1977) 93–107.
[13] V. Thomée, Galerkin finite element methods for parabolic problems, Vol. 25, Springer-Verlag, Berlin, 1997.
[14] D. Hoff, Stability and convergence of finite difference methods for systems of nonlinear reaction-diffusion equations, SIAM J. Numer. Anal. 15 (1978) 1161–1177.
[15] S. J. Ruuth, Implicit-explicit methods for reaction-diffusion problems in pattern formation, J. Math. Biol. 34 (2) (1995) 148–176.
[16] W. Kress, B. Gustafsson, Deferred correction methods for initial boundary value problems, J. Sci Comput. 17 (1-4) (2002) 241–251.
[17] T. Koto, IMEX Runge-Kutta schemes for reaction-diffusion equations, J. Comput. Appl. Math. 215 (1) (2008) 182–195.

[18] B. Bujanda, J. C. Jorge, Efficient linearly implicit methods for nonlinear multidimensional parabolic problems, J. Comput. Appl. Math. 164 (x) (2004) 159–174.
[19] A. Madzvamuse, A. H. W. Chung, Fully implicit time-stepping schemes and non-linear solvers for systems of reaction-diffusion equations, Appl. Math. Comput. 244 (2014) 361–374.
[20] M. N. Spijker, Stiffness in numerical initial-value problems, J. Comput. Appl. Math. 72 (2) (1996) 393–406.
[21] J. M. Sanz-Serna, J. G. Verwer, W. H. Hundsdorfer, Convergence and order reduction of Runge-Kutta schemes applied to evolutionary problems in partial differential equations, Numer. Math. 50 (4) (1987) 405–418.
[22] S.-C. R. Koyaguerebo-Imé, Y. Bourgault, Arbitrary order A-stable methods for ordinary differential equations via deferred correction, arXiv preprint arXiv:1903.02115v2.
[23] S.-C. E. Koyaguerebo-Imé, Y. Bourgault, Finite difference and numerical differentiation: General formulae from deferred corrections, arXiv preprint arXiv:2005.11754.
[24] A. Ern, J.-L. Guermond, Theory and practice of finite elements, Vol. 159, Springer-Verlag, New York, 2004.
[25] P. G. Ciarlet, The finite element method for elliptic problems, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978, studies in Mathematics and its Applications, Vol. 4.

Conclusion and perspectives

The aim of this thesis was to investigate time-stepping methods having both high order of accuracy and a good stability, for the numerical approximations of reaction-diffusion equations. The idea consists in a generalization and improvement of a family of arbitrary high order time-stepping schemes introduced by Gustafsson and Kress (2002). The time-stepping method from Gustafsson and Kress are constructed via a deferred correction (DC) strategy and adresses only linear initial value problems (IVP) satisfying a monotonicity condition while they have an issue for their starting procedures. The generalization and improvement of these time-stepping methods and their application to reaction-diffusion equations are done in three steps corresponding to three submitted articles that constitute the chapter 1, 2 and 3 of the thesis:

1) The paper in chapter 1 introduces a new approach to derive various finite differences formulae of arbitrary high order for the numerical approximation of the derivatives of any order of analytic functions. Many examples of arbitrary high order finite difference formulae suited for DC methods are given in this paper. Furthermore, the new approach recovers the standard centered, backward and forward finite difference formulae given in terms of formal power series of finite difference operators.

2) The paper in chapter 2 gives a sequence $\{DC2j\}_j$ of A-stable arbitrary high order time-stepping schemes which are self-starting, for the numerical approximation of general initial value problems. The schemes are built recursively from the implicit midpoint rule using the deferred correction strategy inspired by Gustafsson and Kress. The starting procedures are made automatic and optimal owing to an efficient centered finite difference formula introduced in chapter 1, and the complete analysis of the convergence is done using a deferred correction condition (DCC) which guarantees the improvement of the order of accuracy by two from a scheme $DC2j$ to the scheme $DC(2j+2)$. Numerical experiments on standard stiff and non-stiff IVPs are performed and showed that the DC schemes have large stability region, achieved their proper order of accuracy (case of $DC2\cdots, DC10$), and are adapted for approximate solutions on large time intervals. The step sizes used by the schemes to compute an approximate solution are not necessarily small.

3) The paper in chapter 3 constitutes the application of the new DC method intro-

duced in chapter 2 to an initial boundary value problem (IBVP) related to reaction-diffusion equations. The IBVP is first discretized in time via the DC method, followed by a space discretization using the Galerkin finite element method. It results a family of fully discrete schemes for the numerical approximation of the IBVP which is proven to be:

- of arbitrary high order both in time and space (the order of accuracy in time is $2j + 2$ at the stage $j = 0, 1, 2, ...$ of the correction while the order of accuracy in space is at least equal to the degree of the finite element used);
- unconditionally stable (convergences of the fully discrete schemes hold with time steps independent from space steps);
- and strongly stable (the method is compatible with the monotonicity condition of the reaction-diffusion equations).

A complete analysis of the method is given using a deferred correction condition, as in the case of IVPs, and a numerical test on a bistable reaction-diffusion equation having a strong stiffness ratio agrees with the theory. The higher order DC methods reach smaller error levels by solving a smaller number of nonlinear systems than lower order DC methods, for time steps not necessarily small, and the convergence is towards machine accuracy.

The existing time-stepping methods face at least one of the following challenges: lack of stability when order 3 or greater is investigated, overly small time step or/and reduction of the order of convergence for stiff problems, inefficiency for long term integration. In fact, our DC time-stepping methods have a large stability region, they are not prone to order reduction or a stability restriction even for a stiff problems (ODEs as PDEs) of large dimension and are adapted for long intervals of integration. Their convergence are toward machine accuracy with time steps not necessarily small, and the computational effort favors higher order DC methods for a given accuracy (at least in the region of asymptotic convergence). Furthermore, the discretizations in chapter 3 give efficient schemes for the numerical approximation of parabolic equations in general, and the techniques used for our proofs constitute some new tools for the analysis of the DC methods applied to more general time-evolution PDEs. Consequently, our work in this thesis constitutes a very important contribution to the literature on the numerical analysis of time-stepping methods.

The work done in this thesis gives rise to many perspectives for which the following list is far from being complete:

- Apply the DC methods constructed to more general time-evolution partial differential equations such as hyperbolic PDEs and Naviers-Stokes equations. In fact, the DC methods have a simple structure and can be easily applied, as time stepping methods, to more general ODEs and PDEs.

- Investigate a solver with adaptive time-steps for the DC schemes constructed. In fact, our convergence analysis of the DC schemes pointed out the necessity of an initial time step k_0 for a global convergence. Constant stepsize codes use time steps $k \leq k_0$ to compute accurate approximate solutions, resulting in high computational efforts. In practice, the stepsizes needed for an accurate approximate solution of an IVP by the DC methods depend locally on the Jacobian matrix of a linearized form of the IVP along its solution curve. Since the Jacobian matrix is not constant along the solution curve (except for some linear systems), a variable-stepsize code should allow the use of time steps $k \geq k_0$ in regions with mild variations and then significantly reduces the CPU time.

- Investigate the impact of starting values on standard high order time-stepping methods that require a starting procedure. In fact, we have done a careful study on starting values versus possible order reduction, which allowed us to built the efficient starting procedures for our DC schemes. However, the numerical experiments performed in chapter 2, comparing our DC schemes with respect to implemented backward differentiation formulae (BDF) using exact starting values and the BDF solver with adaptive steps, showed an impact of starting values on the BDF methods. As a consequence, a study about the impact of starting values on high order time-stepping methods in general is necessary.

- Investigate new variant of DC time-stepping methods based on the BDF and the Runge Kutta methods. In fact, DC method for BDF1 was analysed in first versions of the papers in chapter 2 and 3, but the theory was not complete (in the case of PDEs) and numerical experiments were not performed. It is possible to continue such study and investigate more general extensions. We recall that a correction on, for example, BDF4 will lead to a scheme of order 8 of accuracy with about the double of the computational effort required by BDF4.

Bibliography

[1] Mark J Ablowitz and Anthony Zeppetella. Explicit solutions of Fisher's equation for a special wave speed. *Bulletin of Mathematical Biology*, 41(6):835–840, 1979.

[2] Georgios Akrivis. Stability of implicit-explicit backward difference formulas for nonlinear parabolic equations. *SIAM J. Numer. Anal.*, 53:464–484, 2015.

[3] Georgios Akrivis and Michel Crouzeix. Linearly implicit methods for nonlinear parabolic equations. *Math. Comp.*, 73:613–635, 2004.

[4] Georgios Akrivis, Michel Crouzeix, and Charalambos Makridakis. Implicit-explicit multistep finite element methods for nonlinear parabolic problems. *Math. Comp.*, 67:457–477, 1998.

[5] B. Bujanda and J. C. Jorge. Efficient linearly implicit methods for nonlinear multidimensional parabolic problems. *J. Comput. Appl. Math.*, 164(x):159–174, 2004.

[6] Andrew Christlieb, Benjamin Ong, and Jing-Mei Qiu. Integral deferred correction methods constructed with high order Runge-Kutta integrators. *Math. Comp.*, 79:761–783, 2010.

[7] James W Daniel, Victor Pereyra, and Larry L Schumaker. Iterated deferred corrections for initial value problems. *Wisconsin Univ Madison Mathematics Research center*, 1967.

[8] Alok Dutt, Leslie Greengard, and Vladimir Rokhlin. Spectral deferred correction methods for ordinary differential equations. *BIT*, 40:241–266, 2000.

[9] Lawrence C. Evans. *Partial differential equations*, volume 19. American Mathematical Society, Providence, RI, 1998.

[10] Bertil Gustafsson and Wendy Kress. Deferred correction methods for initial value problems. *BIT*, 41:986–995, 2001.

[11] Anders C Hansen and John Strain. On the order of deferred correction. *Appl. Numer. Math.*, 61:961–973, 2011.

[12] David Hoff. Stability and convergence of finite difference methods for systems of nonlinear reaction-diffusion equations. *SIAM J. Numer. Anal.*, 15:1161–1177, 1978.

[13] Toshiyuki Koto. IMEX Runge-Kutta schemes for reaction-diffusion equations. *J. Comput. Appl. Math.*, 215(1):182–195, 2008.

[14] Wendy Kress and Bertil Gustafsson. Deferred correction methods for initial boundary value problems. *J. Sci Comput.*, 17(1-4):241–251, 2002.

[15] Kristian Kristiansen. *Reaction-diffusion models in mathematical biology.* PhD thesis, Master thesis, Technology University of Denmark, 2008.

[16] Dan Kushnir and Vladimir Rokhlin. A highly accurate solver for stiff ordinary differential equations. *SIAM J. Sci. Comput.*, 34:A1296–A1315, 2012.

[17] J.-L. Lions. *Quelques méthodes de résolution des problèmes aux limites non linéaires.* Dunod; Gauthier-Villars, Paris, 1969.

[18] Anotida Madzvamuse and Andy H. W. Chung. Fully implicit time-stepping schemes and non-linear solvers for systems of reaction-diffusion equations. *Appl. Math. Comput.*, 244:361–374, 2014.

[19] Peter A. Markowich. *Applied partial differential equations: A visual approach.* Springer, Berlin, 2007.

[20] Ya. B. Pesin and A. A. Yurchenko. Some physical models described by the reaction-diffusion equation, and coupled map lattices. *Uspekhi Mat. Nauk*, 59:81–114, 2004.

[21] Steven J. Ruuth. Implicit-explicit methods for reaction-diffusion problems in pattern formation. *J. Math. Biol.*, 34(2):148–176, 1995.

[22] J. M. Sanz-Serna, J. G. Verwer, and W. H. Hundsdorfer. Convergence and order reduction of Runge-Kutta schemes applied to evolutionary problems in partial differential equations. *Numer. Math.*, 50(4):405–418, 1987.

[23] K.-H. Schild. Gaussian collocation via defect correction. *Numer. Math.*, 58:369–386, 1990.

[24] Joel Smoller. *Shock waves and reaction-diffusion equations*, volume 258. Springer-Verlag, New York, 2nd edition, 1994.

[25] M. N. Spijker. Stiffness in numerical initial-value problems. *J. Comput. Appl. Math.*, 72(2):393–406, 1996.

[26] Roger Temam. *Infinite-dimensional dynamical systems in mechanics and physics*, volume 68. Springer-Verlag, New York, 2nd edition, 1997.

[27] Vidar Thomée. *Galerkin finite element methods for parabolic problems*, volume 25. Springer-Verlag, Berlin, 1997.

[28] Vitaly Volpert. *Elliptic partial differential equations*, volume 104. Birkhäuser/Springer Basel AG, Basel, 2014.

[29] Auzinger W. *Encyclopedia of Applied and Computational Mathematics.*, chapter Defect Correction Methods, pages 323–332. Springer, Berlin, Heidelberg, 2015.

[30] Miloš Zlámal. Finite element methods for nonlinear parabolic equations. *RAIRO Anal. Numér.*, 11:93–107, 1977.