SOLVING THE SYSTEMS OF EQUATIONS ARISING IN THE DISCRETIZATION OF SOME NONLINEAR P.D.E.'S BY IMPLICIT RUNGE–KUTTA METHODS

GEORGIOS AKRIVIS, VASSILIOS A. DOUGALIS, AND OHANNES KARAKASHIAN

Abstract. We construct and analyze iterative methods for the efficient solution of the nonlinear equations that result from the application of Implicit Runge–Kutta methods to the temporal integration of nonlinear evolution equations. Some of the schemes we consider have as starting point Newton's method and can be applied to a large class of evolution equations.

Rèsumè. On construit et analyse des méthodes itératives permettant une rśolution efficace des systèmes non linéaires issus de la discrétisation en temps d'équations d'évolution non linéaires par des méthodes de Runge–Kutta implicites. Certains schémas considérés dérivent de la méthode de Newton et s'appliquent à une large classe d'équations non linéaires.

1. Introduction

Whenever an implicit Runge–Kutta method is used to generate approximations of solutions of evolution equations, the issue of solving the resulting system of equations arises. One realises the importance of this simply by recognizing the fact that the computational work is almost entirely concentrated there.

In this work, our aim is to propose and analyze efficient solutions to this problem. With this in mind, the first issue that we needed to address was the choice of an appropriate class of evolution problems to consider. This had to be sufficiently large to encompass problems of practical interest and yet one that could be described simply. Two specific types of problems we wished to study were stiff systems of nonlinear ordinary differential equations (posed on $\mathbb{R}^m$ for some fixed m) and, mainly, large, sparse stiff systems resulting from finite element or finite difference spatial semidiscretization of initial and boundary value problems for nonlinear partial differential equations with smooth solutions. In the latter case, the size of the systems increases without bound as the spatial discretization parameters tend to zero. In order to conduct the analysis in a unified manner we chose to work in the setting of a family of finite-dimensional Hilbert spaces H_m parametrized by a positive parameter m that can take large values. This family may reduce to a single member ($\mathbb{R}^m$) in the case of a specific system of o.d.e.'s or represent, for example, a sequence of finite element spaces of increasing dimension.

The work of these authors was supported by a Greece–USA Collaboration in Research and Technology Grant INT-9107457.

In the case of a semidiscretization of a partial differential operator, the parameter m also enters the problem as a measure of the magnitude of the error of the semidiscrete approximation, through the comparability constants of several norms defined on H_m for the purpose of the error analysis and through bounds on quantities associated with the nonlinear part of the p.d.e. Since it is imperative that all the error constants be bounded independently of m, all quantities depending on the latter must be carefully monitored.

The techniques of error estimation are motivated by our previous studies of low- and high-order accurate IRK temporal discretizations (and their efficient implementation) in the context of the Korteweg–de Vries equation ([3], [8], [12], [4]) and the nonlinear Schrödinger equation ([1], [11]). In the paper at hand we work in an abstract setting and under assumptions on the nonlinear terms that permit the analysis to carry over to more general problems and to other semidiscretization of nonlinear evolution equations as well.

This paper is organized as follows. In Section 2 we introduce the problem and the attendant notation and state the basic assumptions on the solution, the operators in the differential equation and on the IRK schemes. A basic feature of our work is that the assumptions an the nonlinear part of the operator afford us a considerable generalization over the (global) monotonicity condition frequently assumed in the literature. Indeed, our methodology is designed to apply to specific classes of p.d.e.'s with spatial derivatives in the nonlinear terms. In this approach, which invokes a local monotonicity condition, one takes pains to operate in a neighborhood of (or in a tube around) a smooth solution of the evolution equation. This idea is certainly not new. Indeed it is a pervading, though not explicitly recognized theme in the works of many authors, including the present ones, who have analyzed spatial and temporal discretizations of solutions to time dependent p.d.e.'s. Its importance is beginning to be explicitly recognized, see e.g. [2], [14]. The examples contained in this work should convince the reader that the particular norm defining the tube around the solution is highly dependent on the particular nonlinearity and is much more likely to be an $L^\infty-$based Sobolev norm than the Hilbert space norm.

In Section 3 we introduce the base scheme that is obtained by applying the IRK method to the initial-value problem. For the purposes of the error analysis we found convenient to assume that the IRK schemes under consideration are algebraically stable, satisfy the usual simplifying assumptions on the order conditions and, also, a positivity property that guarantees the existence of solutions of the nonlinear system of intermediate stages [7]. We consider issues of existence and uniqueness of the solutions of the resulting discrete problems and estimate their errors. We then prove a general convergence result for the base scheme with an error estimate of optimal-rate spatial accuracy. The techniques we used are well-known and can be found, with references to the original papers in [6], [9]. Nevertheless, we also note that the analysis presented, especially in what concerns stability, uses only the local monotonicity condition alluded to above rather than the global version.

In Section 4 we consider Newton's iterative method for solving the nonlinear system of the intermediate equations. We show that it preserves the spatial and temporal orders of accuracy of the base scheme, provided it is started with sufficiently accurate initial conditions at each time step, if certain suitable mesh conditions are valid, and if sufficiently many Newton iterations are performed at each time step. The number of iterations needed depends on the accuracy of the starting values and the temporal order of accuracy of the base scheme. It is shown that under some realistic conditions, no more than one iteration is needed.

In Section 5 we study an efficient variant of Newton's method, the so-called modified Newton method. The obvious advantage that the Jacobian matrix need not be updated at every iteration is enhanced by the possibility of decoupling and simultaneous solving ("in parallel") for the intermediate stages. The modified scheme no longer converges quadratically; we show however that, with sufficiently many iterations, it preserves the spatial and temporal orders of accuracy of the base scheme. In Section 6 we analyze an even simpler iterative scheme, which is sometimes referred to as the "explicit–implicit" method as it is based on a splitting of the linear and nonlinear parts of the operator. The resulting method is very efficient in that the linear systems that need to be solved have the same coefficient matrix, i.e. a matrix that does not vary with the time stepping. However this scheme is not applicable to as a wide class of evolution equations as the modified Newton method.

Finally, in Section 7 we apply the methodology developed in the previous sections to two concrete examples corresponding to finite element semidiscretizations of the Korteweg–de Vries (KdV) and the Cubic Schrödinger equations. In addition to providing illustrative examples to the formal approach adopted in the work at hand, the results of this section supplement those in [12] and [11] by providing complete analyses of efficient linearizations of the fully discrete schemes proposed in those works. Besides establishing convergence, the following useful information is gleaned:

(i) The number of iterations required to preserve the rate of convergence of the base scheme is determined for each linearization technique.
(ii) Typically, Courant number type stability conditions between the spatial and temporal discretization parameters are required. These are explicitly exhibited.

Newton-type methods for solving the nonlinear systems resulting from IRK schemes have often been considered in the literature of stiff systems of o.d.e.'s. For a survey of the literature and a list of references we refer the reader to a recent paper of Alexander, [2]. In that work, Alexander analyzes the modified Newton method as applied to nonlinear systems resulting from the application of quite general IRK schemes to stiff systems of o.d.e.'s, that have a Jacobian of the right-hand side term which is essentially negative dominant and slowly varying. Using matrix methods he proves that the modified Newton iteration converges linearly to the locally unique solution of o.d.e.'s. In this work, we emphasize models of stiff initial-value problems that are semidiscretizations of nonlinear p.d.e.'s. In such cases, especially if higher-order semidiscretizations are used, the Jacobian may not be essentially negative dominant, or if such be the case, it may be quite difficult to establish this property given that the entries of the Jacobian must be examined.

2. Preliminaries

2.1. **The basic assumptions.**

Let $\mathscr{M}$ denote a set of positive numbers (infinite or otherwise) and for $m \in \mathscr{M}$, let $(H_m, (\cdot,\cdot)_{H_m})$ denote a corresponding family of finite dimensional (real) inner product spaces. In some applications $\mathscr{M}$ may be a set of positive integers and m may denote the dimension of H_m; in others, m may be used to denote a more general parameter for H_m.

We consider the following family of initial-value problems

$$\begin{cases} \dfrac{d\omega_m}{dt} = L_m\omega_m + \varphi_m(\omega_m) + \varepsilon_m(t), \quad 0 \le t \le T, \\ \omega_m(0) = \omega_m^0, \end{cases} \tag{2.1}$$

for some $T > 0$, where $\omega_m : [0,T] \to H_m, L_m : H_m \to H_m$ are linear operators, $\varphi_m : H_m \to H_m$ are smooth functions and $\varepsilon_m : [0,T] \to H_m$ are smooth functions satisfying

$$\max_{0\le t\le T} \|\varepsilon_m(t)\|_{H_m} \le cm^{-s}, \tag{H1}$$

for some $s > 0$ and a constant c independent of m.

One area of application of (2.1) we have in mind is that when ω_m represents a continuous-in-time approximation to the solution u of the time dependent p.d.e.

$$\begin{cases} \dfrac{du}{dt} = Lu + \varphi(u), \quad 0 \le t \le T, \\ u(0) = u^0. \end{cases}$$

In this case, the functions $\varepsilon_m(t)$ may represent semidiscretization errors, and may be unknown. In view of (H1) and other considerations to follow, it turns out that the $\varepsilon_m(t)$ will not play a major part in the time integration process.

We assume that for some constants λ, η, independent of m,

$$(L_m v, v)_{H_m} \le \lambda \|v\|_{H_m}^2, \quad \forall v \in H_m, \tag{H2}$$

$$(\varphi_m(v), v)_{H_m} \le \eta \|v\|_{H_m}^2, \quad \forall v \in H_m. \tag{H3}$$

Note that $\varphi_m(0) = 0$, as a consequence of (H3) and the continuity of φ_m. To simplify matters, we assume that $\lambda, \eta \ge 0$.

We assume that for each m, H_m is additionally equipped with norms $\|\!| \cdot |\!\|_{i,m}, i = 1,2,3,4$. These norms are obviously equivalent to $\|\cdot\|_{H_m}$. Specifically, let $c > 0$ and $s_1, s_2, s_3, s_4 \ge 0$ be constants independent of m such that

$$\|\!| v |\!\|_{i,m} \le cm^{s_i} \|v\|_{H_m}, \quad i = 1,2,3,4, \quad \forall v \in H_m. \tag{2.2}$$

For $m \in \mathscr{M}, i = 1,2,3,4$ and $\rho > 0$, we introduce the sets

$$B_{i,m}(\rho) = \{v \in H_m : \|\!| v |\!\|_{i,m} \le \rho\}.$$

Now let M, K, β, γ and δ be given positive numbers. For $m \in \mathscr{M}$, we introduce the spaces $\mathscr{F}^1_{m,K,M}, \mathscr{F}^2_{m,K,M,\beta}, \mathscr{F}^3_{m,K,M,\gamma}$ and $\mathscr{F}^4_{m,K,M,\gamma}$ by

$$\mathscr{F}^1_{m,K,M} = \{g : H_m \to H_m | \, (g(u)-g(v), u-v)_{H_m} \le K\|u-v\|^2_{H_m} \, \forall u \in B_{1,m}(M), \forall v \in H_m\},$$

$$\mathscr{F}^2_{m,K,M,\beta} = \{g : H_m \to H_m | \, (Dg(u)v, v)_{H_m} \le Km^\beta \|v\|^2_{H_m} \, \forall u \in B_{2,m}(M), \forall v \in H_m\},$$

$$\mathscr{F}^3_{m,K,M,\gamma} = \{g : H_m \to H_m | \, \|Dg(u)v\|_{H_m} \le Km^\gamma \|v\|_{H_m} \, \forall u \in B_{3,m}(M), \forall v \in H_m\},$$

$$\begin{aligned} \mathscr{F}^4_{m,K,M,\delta} = \{g : H_m \to H_m | \, & \|D^2 g(u)[v,w]\|_{H_m} \le Km^\delta \|v\|_{H_m} \|w\|_{H_m} \\ & \forall u \in B_{4,m}(M), \forall v, w \in H_m\}. \end{aligned}$$

Here, $Dg, D^2 g$ are the first and second Fréchet derivatives of g, respectively.

We assume that there exist nonnegative constants $M, K, \beta, \gamma, \delta$ such that

$$\text{(H4)} \qquad \varphi_m \in \mathscr{F}^1_{m,K,M}, \quad \forall m \in \mathscr{M},$$

$$\text{(H5)} \qquad \varphi_m \in \mathscr{F}^2_{m,K,M,\beta}, \quad \forall m \in \mathscr{M},$$

$$\text{(H6)} \qquad \varphi_m \in \mathscr{F}^3_{m,K,M,\gamma}, \quad \forall m \in \mathscr{M},$$

$$\text{(H7)} \qquad \varphi_m \in \mathscr{F}^4_{m,K,M,\delta}, \quad \forall m \in \mathscr{M}.$$

We observe that if (H2) holds, then it follows from (H4) and (H5) that $\forall m \in \mathscr{M}, L_m + \varphi_m \in \mathscr{F}^1_{m,K,M}, \mathscr{F}^2_{m,K,M,\beta}$, respectively with K replaced by $K + \lambda$.

The definition of $\mathscr{F}^1_{m,K,M}$ explicitly formulates what we previously referred to as the local monotonicity condition: One of the two vectors u, v is restricted to a suitable ball $B_{1,m}(M)$ containing the solution of the p.d.e. Let us also note that the lack of explicit dependence on t is purely for the sake of simplicity.

(H8) For each $m \in \mathscr{M}$, (2.1) has a unique solution $\omega_m : [0,T] \to \cap_{j=1}^4 B_{j,m}\big(\dfrac{M}{2}\big)$.

$$\text{(H9)} \qquad \max_{0 \le t \le T} \Big\| \frac{d^j}{dt^j} \omega_m(t) \Big\|_{H_m} \le c_j, \quad j = 0, \dots, J,$$

for a sufficiently large J and constants c_j independent of m.

To simplify the notation, we shall suppress subscripts m and H_m whenever possible. Let us also mention that in case the problem is a stiff nonlinear system of o.d.e.'s, not associated with any semidiscretization, we think of it as posed on $\mathbb{R}^m$ for a fixed m. In such a case $\varepsilon_m = 0$.

Remark 2.1. One could argue that hypotheses (H2) and (H3) are global in nature; however, many classes of important p.d.e.'s e.g. parabolic and hyperbolic, as well as specific equations such as the Korteweg–de Vries equation, the Nonlinear Schrödinger equation and the Navier–Stokes equations of fluid mechanics satisfy them. This is in sharp contrast with hypotheses (H4)–(H7) which can only be used in a local setting in order to treat the above-mentioned equations. $\square$

2.2. **The Implicit Runge–Kutta methods.** For $q \geq 1$ integer, a $q-$stage IRK method is given as a set of constants $A = (a_{ij}) \in \mathbb{R}^{q\times q}, b = (b_1, \dots, b_q)^T \in \mathbb{R}^q, \tau = (\tau_1, \dots, \tau_q)^T \in \mathbb{R}^q$. We shall assume that these methods satisfy certain *stability* and *consistency* conditions. Indeed, we require the *algebraic stability condition* cf. [6]

$$\text{(S)} \qquad \begin{cases} b_i \geq 0, \quad i = 1, \dots, q, \\ \text{the } q \times q \text{ array with entries } m_{ij} = a_{ij}b_i + a_{ji}b_j - b_ib_j \\ \text{is positive semidefinite.} \end{cases}$$

The consistency conditions are given by the *simplifying assumptions* [6]

$$\text{(B)} \qquad \sum_{j=1}^{q} b_j \tau_j^\ell = \frac{1}{\ell+1}, \quad \ell = 0, \dots, \nu - 1,$$

$$\text{(C)} \qquad \sum_{j=1}^{q} a_{ij} \tau_j^\ell = \frac{\tau_i^{\ell+1}}{\ell+1}, \quad i = 1, \dots, q, \ \ell = 0, \dots, p - 1,$$

$$\text{(D)} \qquad \sum_{i=1}^{q} a_{ij} \tau_i^\ell b_i = \frac{b_j}{\ell+1}(1 - \tau_j^{\ell+1}), \quad j = 1, \dots, q, \ \ell = 0, \dots, \rho - 1,$$

for some integers $\nu, p, \rho \geq 1$. We assume that

$$\nu \leq \rho + p + 1, \tag{2.3a}$$

$$\nu \leq 2p + 2. \tag{2.3b}$$

The existence of the numerical approximations is obtained by assuming the following positivity property

$$\text{(P)} \qquad \begin{cases} A \text{ is invertible and there exists a positive diagonal matrix } D \text{ such that} \\ x^T C x > 0, \ \forall x \in \mathbb{R}^q, x \neq 0, \ \text{where } C = DA^{-1}D^{-1}. \end{cases}$$

Two classes of IRK methods satisfying the hypotheses above are the Gauss–Legendre methods for which $\nu = 2q, p = q, \rho = q$ and the Radau IIA methods for which $\nu = 2q - 1, p = q, \rho = q - 1$, [6]. We also mention two diagonally implicit (DIRK) methods of orders 3 and 4, respectively. (The fourth-order method does not satisfy (2.3a). However, cf. [12], [11], our theory holds for this method as well.)

3. The base scheme

As noted earlier, the techniques employed in this section are well known. The purpose of the detailed treatment of the base scheme is to provide a benchmark (in terms of the spatial and temporal orders of accuracy of its global error) against which we measure the accuracy of the linearized schemes that are introduced in the subsequent sections.

We begin with a preliminary result which shows that in view of (H1) it is possible to disregard the terms $\varepsilon(t)$ while constructing the temporal discretizations. Denoting $L + \varphi$ by f, we have

Lemma 3.1. *Let $\omega : [0,T] \to H_m$ be the solution of* (2.1) *and let $v : [0,T] \to H_m$ be a solution of the initial value problem*

$$\left\{ \begin{aligned} &\frac{dv}{dt} = f(v), \quad 0 \le t \le T, \\ &v(0) = \omega^0. \end{aligned} \right. \tag{3.1}$$

Then, there exists a constant c, *independent of* m *such that*

$$\max_{0\le t\le T} \|(\omega - v)(t)\| \le cm^{-s}. \tag{3.2}$$

Proof. From (2.1) and (3.1), we get,

$$\frac{d}{dt}(\omega - v) = f(\omega) - f(v) + \varepsilon(t).$$

Taking the inner product with $\omega - v$, from (H2) and (H4) we get

$$\frac{d}{dt}\|\omega - v\|^2 \le 2(\lambda + K)\|\omega - v\|^2 + 2\|\varepsilon(t)\|\,\|\omega - v\|, \quad 0 \le t \le T.$$

Using (H1), we easily get (3.2). □

Let N be a positive integer and let $k = \frac{T}{N}$ represent the temporal step size. We introduce the map $\mathscr{R}(k) = \mathscr{R} : H_m \to H_m$ as follows: For $v \in H_m$, let the *intermediate values* $v^i \in H_m, 1 \le i \le q$, be given by

$$v^i = v + k\sum_{j=1}^{q} a_{ij} f(v^j), \quad i = 1, \dots, q. \tag{3.3}$$

We then set

$$\mathscr{R}v = v + k\sum_{i=1}^{q} b_i f(v^i). \tag{3.4}$$

Note that the existence of $\mathscr{R}v$ depends solely on the existence of the intermediate values $\{v^i\}_{i=1}^q$ satisfying (3.3). Furthermore, since A is invertible in view of (P), (3.4) may be written as

$$\mathscr{R}v = (1 - b^T A^{-1}\boldsymbol{e})v + b^T A^{-1}(v^1, \dots, v^q)^T, \tag{3.4$'$}$$

where $\boldsymbol{e} = (1, \dots, 1)^T \in \mathbb{R}^q$.

We shall next consider the question of existence of the intermediate values. Using a well-known version of Brouwer's fixed point theorem, we shall prove that if k is sufficiently small, then for each $v \in H_m$, there exists al least one solution set $\{\{v^i\}_{i=1}^q, \mathscr{R}v\}$ to (3.3), (3.4). For simplicity of notation, we shall represent this set simply by $\mathscr{R}v$. Note however that, for nonlinear f, the map $\mathscr{R}$ cannot be expected to be single-valued in general.

Lemma 3.2. *Let $(H, (\cdot,\cdot)_H)$ be a finite dimensional Hilbert space and denote by $\|\cdot\|_H$ the associated norm. Suppose that $g : H \to H$ is continuous and that there exists $\alpha > 0$ such that $(g(x), x)_H \ge 0$ for all $x \in H$ with $\|x\|_H = \alpha$. Then, there exists $x^* \in H$ such that $g(x^*) = 0$ and $\|x^*\|_H \le \alpha$.* □

Proposition 3.1. *Assume that* (H2), (H3) *and* (P) *hold. Then there exists* $k_0 = k_0(A, b, \lambda, \eta) > 0$ *such that for each* $0 < k \le k_0$, *and each* $v \in H_m$, *there exists a solution* $\{\{v^i\}_{i=1}^q, \mathscr{R}v\}$ *to* (3.3), (3.4). *Furthermore, all such solutions* $\{\{v^i\}_{i=1}^q\}$ *satisfy*

$$\max_{1\le i\le q} \|v^i\| \le c\|v\|, \tag{3.5}$$

for some constant $c = c(A, \lambda, \eta) > 0$. *If* (S) *is also assumed to hold, then all such solutions* $\mathscr{R}v$ *satisfy the estimate*

$$\|\mathscr{R}v\| \le (1 + ck)\|v\|, \tag{3.6}$$

for some constant $c = c(A, b, \lambda, \eta) > 0$.

Proof. We first establish (3.5) and (3.6). From (3.3), we obtain

$$\sum_{j=1}^q c_{ij} d_j v^j = \sum_{j=1}^q c_{ij} d_j v + k d_i f(v^i), \quad i = 1, \dots, q,$$

where C, D are as in (P). Taking the inner product of the $i-$th equation with $d_i v^i$, summing over i, from (P), (H2) and (H3) it follows that for some constants c_1, c_2 depending only on A,

$$c_1 \sum_{i=1}^q \|v^i\|^2 \le c_2 \|v\| \Big(\sum_{i=1}^q \|v^i\|^2 \Big)^{1/2} + k(\lambda + \eta) \big(\max_i d_i^2 \big) \sum_{i=1}^q \|v^i\|^2.$$

Choosing $k_0 = \frac{c_1}{2(\lambda+\eta)(\max_i d_i^2)}$, we obtain (3.5) for any $0 < k \le k_0$.

Now from (3.3) and (3.4) it follows that

$$\begin{aligned} \|\mathscr{R}v\|^2 &= \|v\|^2 + 2k \sum_{i=1}^q b_i (f(v^i), v) + k^2 \sum_{i,j=1}^q b_i b_j (f(v^i), f(v^j)) \\ &= \|v\|^2 + 2k \sum_{i=1}^q b_i (f(v^i), v^i) - k^2 \sum_{i,j=1}^q m_{ij} (f(v^i), f(v^j)). \end{aligned}$$

Using (H2), (H3) and (S),

$$\|\mathscr{R}v\|^2 \le \|v\|^2 + 2k(\lambda + \eta) \big(\max_{1\le i\le q} b_i \big) \sum_{i=1}^q \|v^i\|^2. \tag{3.7}$$

Using (3.5) in (3.7), we obtain (3.6).

Concerning the question of existence, we first note that if $v = 0$, then $\mathscr{R}v = 0$ is a solution, in view of the fact that $f(0) = 0$. Hence, let $v \ne 0$ and define the map $G = (g_1, \dots, g_q)^T : (H_m)^q \to (H_m)^q$ by

$$g_i(V) = \sum_{j=1}^q c_{ij} d_i d_j (v^i - v) - k d_i^2 f(v^i), \quad i = 1, \dots, q,$$

for $V = (v^1, \dots, v^q)^T, v^i \in H_m, i = 1, \dots, q$. Let $((\cdot, \cdot))$ denote the usual (product) inner product on $(H_m)^q$, and $|\!|\!| \cdot |\!|\!|$ the associated norm.

Then we have

$$((G(V),V)) = \sum_{i,j=1}^{q} c_{ij} d_i d_j \{(v^i, v^j) - (v, v^i)\} - k \sum_{i=1}^{q} d_i^2 (f(v^i), v^i).$$

We see immediately that for $0 < k \le k_0$

$$\begin{aligned}((G(V),V)) &\ge c_1 \|V\|^2 - c_2 \|v\| \, \|V\| - k(\lambda+\eta)\big(\max_i d_i^2\big) \|V\|^2 \\ &\ge \|V\| \Big\{ \frac{c_1}{2} \|V\| - c_2 \|v\| \Big\}.\end{aligned}$$

Hence $((G(V),V)) \ge 0$ for all $V \in (H_m)^q$ satisfying $\|V\| = \frac{2c_2}{c_1}\|v\|$. Using Lemma 3.2, we infer that there exists $V^* \in (H_m)^q$ such that $G(V^*) = 0$. Obviously $(v^1, \ldots, v^q)^T = V^*$ is a solution of (3.3). □

We shall often use steps similar to those leading to the estimate (3.5). In such occurrences, these shall be referred to as *diagonalization arguments.*

We next consider the following stability result

Proposition 3.2. *Assume that* (H2), (H4), (S) *and* (P) *hold. Then, there exists* $k_0 = k_0(A, b, \lambda, K) > 0$ *such that if* $\{v, \{v^i\}_{i=1}^q, \mathscr{R}v\}$ *and* $\{w, \{w^i\}_{i=1}^q, \mathscr{R}w\}$ *satisfy* (3.3) *with* $0 < k \le k_0$ *and* $\{v^i\}_{i=1}^q \subset B_1(M)$, (*or* $\{w^i\}_{i=1}^q \subset B_1(M)$), *then*

$$\max_{1\le i\le q} \|v^i - w^i\| \le c\|v - w\|, \tag{3.8}$$

$$\|\mathscr{R}v - \mathscr{R}w\| \le (1 + ck)\|v - w\|, \tag{3.9}$$

for some constant $c = c(A, b, \lambda, K)$.

Proof. Applying a diagonalization procedure to the system

$$v^i - w^i = v - w + k \sum_{j=1}^{q} a_{ij} \big(f(v^j) - f(w^j)\big), \quad i = 1, \ldots, q,$$

we obtain (3.8) from (H2), (H4) and (P). Furthermore, using (S), we obtain

$$\begin{aligned}\|\mathscr{R}v - \mathscr{R}w\|^2 &\le \|v - w\|^2 + 2k \sum_{i=1}^{q} b_i \big(f(v^i) - f(w^i), v^i - w^i\big) \\ &\quad - k^2 \sum_{i,j=1}^{q} m_{ij} \big(f(v^i) - f(w^i), f(v^j) - f(w^j)\big) \qquad (3.10) \\ &\le \|v - w\|^2 + 2k\big(\max_{1\le i\le q} b_i\big)(\lambda + K) \sum_{i=1}^{q} \|v^i - w^i\|^2.\end{aligned}$$

(3.9) now follows from (3.8) and (3.10). □

Note that, as a result of the above, there exists at most one set $\{v^i\}_{i=1}^q$ in $B_1(M)$ satisfying (3.3).

We now focus attention on the local truncation errors. Letting $t^n = nk$ and $t^{n,i} = t^n + k\tau_i, i = 1, \ldots, q, n = 0, \ldots, N - 1$, we have:

Proposition 3.3. *Assume that hypotheses* (H1), (H2), (H3), (H4), (H8), (H9), (B), (C) *and* (P) *hold. Then there exists $k_0 > 0$ such that for all $0 < k \le k_0$ and for $n = 0, \dots, N-1$, there exist $\omega^{n,i}, \omega^{n+1} = \mathscr{R}\omega(t^n)$ satisfying*

$$\omega^{n,i} = \omega(t^n) + k \sum_{j=1}^{q} a_{ij} f(\omega^{n,j}), \quad i = 1, \dots, q, \tag{3.11}$$

$$\omega^{n+1} = \mathscr{R}\omega(t^n) = \omega(t^n) + k \sum_{i=1}^{q} b_i f(\omega^{n,i}). \tag{3.12}$$

Furthermore,

$$\max_{1 \le i \le q} \|\omega(t^{n,i}) - \omega^{n,i}\| \le ck(k^p + m^{-s}), \tag{3.13}$$

$$\|\omega^{n+1} - \omega(t^{n+1})\| \le ck(k^{\min\{p,\nu\}} + m^{-s}), \tag{3.14}$$

for some constant c independent of k and m.

Proof. The existence of $\{\omega^{n,i}\}_{i=1}^q$ and hence of ω^{n+1} follows from Proposition 3.1. Now let $\{\rho^{n,i}\}_{i=1}^q$ and ρ^{n+1} in H_m be given by

$$\rho^{n,i} = \omega(t^{n,i}) - \omega(t^n) - k \sum_{j=1}^{q} a_{ij} f(\omega(t^{n,j})), \tag{3.15}$$

$$\rho^{n+1} = \omega(t^{n+1}) - \omega(t^n) - k \sum_{i=1}^{q} b_i f(\omega(t^{n,i})). \tag{3.16}$$

From (2.1), Taylor's theorem, (H1) and (H9),

$$\begin{aligned} \rho^{n,i} &= \omega(t^{n,i}) - \omega(t^n) - k \sum_{j=1}^{q} a_{ij} \big[\frac{d\omega}{dt}(t^{n,j}) - \varepsilon(t^{n,j})\big] \\ &= \sum_{\ell=1}^{p} k^\ell \frac{\tau_i^\ell}{\ell!} \frac{d^\ell \omega}{dt^\ell}(t^n) - k \sum_{j=1}^{q} a_{ij} \sum_{\ell=1}^{p} k^{\ell-1} \frac{\tau_j^{\ell-1}}{(\ell-1)!} \frac{d^\ell \omega}{dt^\ell}(t^n) + O(k^{p+1} + km^{-s}). \end{aligned}$$

Using (C), it follows easily that

$$\max_{1 \le i \le q} \|\rho^{n,i}\| \le ck(k^p + m^{-s}). \tag{3.17}$$

Now it follows from (3.15) and (3.11) that

$$\omega(t^{n,i}) - \omega^{n,i} = k \sum_{j=1}^{q} a_{ij} \big[f(\omega(t^{n,j})) - f(\omega^{n,j})\big] + \rho^{n,i}. \tag{3.18}$$

In view of (P), (H2), (H4), (H8) and (3.17), a diagonalization argument gives (3.13) for k sufficiently small.

Proceeding as is the derivation of (3.17) but using (B) instead of (C), we obtain

$$\|\rho^{n+1}\| \le ck(k^\nu + m^{-s}). \tag{3.19}$$

Moreover, it follows from (3.12), (3.16) and (3.18) that

$$\omega(t^{n+1}) - \omega^{n+1} = k\sum_{i=1}^{q} b_i\big[f(\omega(t^{n,i})) - f(\omega^{n,i})\big] + \rho^{n+1}$$
$$= \sum_{i,j=1}^{q} b_i(A^{-1})_{ij}\big[\omega(t^{n,j}) - \omega^{n,j} - \rho^{n,j}\big] + \rho^{n+1}.$$

(3.14) now follows from (3.13), (3.17) and (3.19). □

In case $\varepsilon_m = 0$, i.e., when we have no semidiscretization of a p.d.e. in mind, the results of Propositions 3.3 hold without any spatial contribution m^{-s} in the bounds (3.13) or (3.14). The same holds for the rest of the analogous estimates in Sections 3–6.

We are now ready to state and prove the main result of this section.

Theorem 3.1. *Assume that the hypotheses of Propositions 3.1, 3.2 and 3.3 hold. Assume additionally that*

(i) $s_1 \le s$.

Then there exist $k_0, m_0, c_0 > 0$ *such that for all* $0 < k \le k_0$ *and all* $m \ge m_0$ *satisfying*

(ii) $k^{p+1}m^{s_1} \le c_0$,

there exists a sequence $V^0, \{\{V^{n,i}\}_{i=1}^q, V^{n+1}\}_{n=0}^{N-1} \subset H_m$ *given by*

$$\begin{cases} V^0 = \omega^0, \\ V^{n,i} = V^n + k\sum_{j=1}^{q} a_{ij}f(V^{n,j}), \quad i = 1,\dots,q, \\ V^{n+1} = \mathscr{R}V^n = V^n + k\sum_{i=1}^{q} b_i f(V^{n,i}). \end{cases} \tag{3.20}$$

In addition, the following estimate holds

$$\max_{0\le n\le N} \|\omega(t^n) - V^n\| \le ck\{k^{\min\{p,\nu\}} + m^{-s}\}, \tag{3.21}$$

for some constant c independent of k and m. We shall call (3.20) *the "base scheme".*

Proof. Applying Proposition 3.1 repeatedly, we can establish the existence of a sequence $\{\{V^{n,i}\}_{i=1}^q, V^{n+1}\}_{n=0}^{N-1}$ satisfying (3.20). Now using (H8), (2.2) and (3.13),

$$\begin{aligned} |||\omega^{n,i}|||_1 &\le |||\omega^{n,i} - \omega(t^{n,i})|||_1 + |||\omega(t^{n,i})|||_1 \\ &\le cm^{s_1}k\{k^p + m^{-s}\} + \frac{M}{2}, \quad i = 1,\dots,q. \end{aligned} \tag{3.22}$$

Hence, in view of (i) and (ii), for k sufficiently small, it follows that

$$\omega^{n,i} \in B_1\Big(3\frac{M}{4}\Big), \quad i = 1,\dots,q,\ n = 0,\dots,N-1. \tag{3.23}$$

Applying Proposition 3.2, we see that

$$\|\omega^{n+1} - V^{n+1}\| \le (1+ck)\|\omega(t^n) - V^n\|.$$

From (3.14) and the triangle inequality, it follows that for $n = 0, \dots, N-1$,

$$\|\omega(t^{n+1}) - V^{n+1}\| \le (1+ck)\|\omega(t^n) - V^n\| + ck\{k^{\min\{p,\nu\}} + m^{-s}\}.$$

(3.21) now follows easily from recursion. □

Remarks

1) It is obvious that Theorem 3.1 remains valid for any choice of V^0 in H_m that satisfies

$$\|V^0 - \omega^0\| \le cm^{-s}, \tag{3.24}$$

where c is independent of m. Consequently, we shall refer to (3.20) with V^0 satisfying (3.24) as the "base scheme" as well.

2) (i) and (ii) form a set of convenient sufficient conditions that guarantee that $\omega^{n,i} \in B_1(M)$ for all n, i. In special cases, (3.23) may be proved in a more direct manner, cf. e.g. [3].

3) If $s_j \le s$ and $k^{p+1} m^{s_j} \le c_0$ for some $j, 1 \le j \le 4, c_0$ sufficiently small, then

$$\omega^{n,i} \in B_j\Big(3\frac{M}{4}\Big), \quad i = 1, \dots, q,\ 0 \le n \le N-1. \tag{3.25}$$

In general $1 \le p \le q$ whilst ν may be as large as $2q$, as in the case of Gauss–Legendre methods. For some specific problems, using (D), (2.3a), (2.3b) and specialized techniques, one may obtain an improved rate of convergence estimate for the base scheme. See for instance [12], [11]. In order to accomodate such special cases, we shall make the assumption

$$\|\omega(t^n) - V^n\| \le c\{k^\sigma + m^{-s}\}, \quad n = 0, \dots, N, \tag{H10}$$

for some integer σ, with $p \le \sigma \le \nu$ and for some constant c independent of k and m.

Finally, let us remark that with slightly more stringent conditions than (i), (ii), one may prove uniqueness of the $V^{n,i}$ as well. For example, consider the following

Corollary 3.1. *Assume that* (H10) *holds and that in addition to the assumptions of Theorem 3.1 we have*

(i) $s_1 < s$,

(ii) $k^\sigma m^{s_1} \le c_0,\ c_0$ *sufficiently small.*

Then, for a given choice of V^0 *satisfying* (3.24)*, there exists a unique solution* $\{\{V^{n,i}\}_{i=1}^q, V^{n+1}\}_{n=0}^{N-1}$ *to* (3.20)*.*

Proof. In view of Proposition 3.2, it suffices to show that

$$\max_{1\le i\le q} |\!|\!| V^{n,i} |\!|\!|_1 \le M, \quad n = 0, \dots, N-1. \tag{3.26}$$

To obtain this, from (3.8) and (H10) it follows that

$$\max_{1\le i\le q} \|V^{n,i} - \omega^{n,i}\| \le c\|V^n - \omega(t^n)\| \le c\{k^\sigma + m^{-s}\}.$$

Hence, using (i), (ii), (2.2) and (3.23), we obtain

$$\max_{1\le i\le q} |\!|\!| V^{n,i} |\!|\!|_1 \le cm^{s_1}\{k^\sigma + m^{-s}\} + \max_{1\le i\le q} |\!|\!| \omega^{n,i} |\!|\!|_1 \le M,$$

which is the desired result. □

4. Newton's method

To begin, let us recall that Newton's method for approximating a root of a smooth function $g : X \to X$, where X is a normed linear space, is given by

$$Dg(x_\ell)(x_{\ell+1} - x_\ell) = -g(x_\ell), \quad \ell = 0, 1, \ldots, \quad x_0 \text{ given.}$$

In our particular context, given approximations $U^j \in H_m, U^j \approx u(t^j), j = 0, \ldots, n$, Newton's iterative procedure for approximating the intermediate values $\{U^{n,i}\}$ takes the form

$$U^{n,i}_{\ell+1} - k \sum_{j=1}^{q} a_{ij} Df(U^{n,j}_\ell)(U^{n,j}_{\ell+1} - U^{n,j}_\ell) = U^n + k \sum_{j=1}^{q} a_{ij} f(U^{n,j}_\ell), \tag{4.1}$$
$$i = 1, \ldots, q, \ \ell = 0, \ldots, \ell_n - 1.$$

The *starting values* $U^{n,i}_0$ are assumed given, and $\ell_n \geq 1$ is the number of iterations to be performed at step n. We then define U^{n+1} by

$$U^{n+1} = (1 - b^T A^{-1} \boldsymbol{e}) U^n + b^T A^{-1} \begin{pmatrix} U^{n,1}_{\ell_n} \\ \vdots \\ U^{n,q}_{\ell_n} \end{pmatrix}. \tag{4.2}$$

Starting values $U^{n,i}_0$ may be generated by a variety of techniques. For example, one could use the collocation polynomial from the previous step as advocated in [9]. In this paper, we generate them simply by extrapolation from previously computed values $U^n, U^{n-1}, \ldots$ according to

$$U^{n,i}_0 = \sum_{j=0}^{p_n} \mu^{p_n}_{ij} U^{n-j}, \quad i = 1, \ldots, q, \ n = 0, \ldots, N-1, \tag{4.3}$$

where $p_n \leq n$ is a nonnegative integer and where the extrapolation coefficients are generated as follows: For integer ℓ such that $0 \leq \ell \leq n$, let $\{L^{\ell,n}_i\}_{i=0}^{\ell}$ be the (Lagrange) polynomials of degree ℓ that satisfy $L^{\ell,n}_i(t^{n-j}) = \delta_{ij}, 0 \leq i, j \leq \ell$. Then set

$$\mu^\ell_{ij} = L^{\ell,n}_j(t^n + k\tau_i) = \prod_{\substack{r=0 \\ r \neq j}}^{\ell} \frac{\tau_i + r}{r - j}, \quad 1 \leq i \leq q, \ 0 \leq j \leq \ell. \tag{4.4}$$

Using Taylor's theorem, it can be shown that for any smooth function u,

$$\sum_{j=0}^{\ell} \mu^\ell_{ij} u(t^{r-j}) = u(t^r + k\tau_i) + O(k^{\ell+1}), \quad 1 \leq i \leq q, \ r \geq \ell. \tag{4.5}$$

In view of the fact that the accuracy of the extrapolated values is limited by the number of available past data, as well as by $p + 1$ and σ we shall take

$$p_n = \min\{n, p, \sigma - 1\}. \tag{4.6}$$

Theorem 4.1. *Assume that* (H10) *and the hypotheses of Theorem 3.1 are satisfied and that we are given initial data* $U^0,\dots,U^{\bar p}, \bar p=\min\{p,\sigma-1\}$, *satisfying*

$$\|U^j-\omega(t^j)\|\le c\{k^\sigma+m^{-s}\},\quad 0\le j\le \bar p, \tag{4.7}$$

for some constant c *independent of* k *and* m.

Assume in addition that

(i) (H5) *holds,*

(ii) (H7) *holds and* $\delta<s$,

(iii) $s_1,s_2,s_3<s$,

(iv) $\ell_n\ge \log_2(\sigma-\bar p+1),\quad \bar p\le n\le N-1$.

Then, there exist $k_0,m_0,c_0>0$ *such that for all* $0<k\le k_0$ *and for all* $m\ge m_0$ *satisfying*

(v) $kKm^\beta\le c_0$,

(vi) $k^{\bar p+1}m^{s_j}\le c_0,\quad$ *for* $j=1,2,4$,

(vii) $k^{\bar p+1}m^\delta\le c_0$,

there exists a unique sequence $\{U^n\}_{n=0}^N$ *which for* $\bar p+1\le n\le N$ *is generated by* (4.1), (4.2) *and* (4.3) *with* $p_n=\bar p$. *Furthermore,*

$$\max_{0\le n\le N}\|U^n-\omega(t^n)\|\le c\{k^\sigma+m^{-s}\}, \tag{4.8}$$

for some constant c *independent of* k *and* m.

Moreover, if $p\le\sigma\le 2p$, *then the conclusion of the theorem holds with* $\ell_n=1$ *provided*

(viii) $k^{2\bar p-\sigma+2}Km^\delta$ *is sufficiently small.*

Proof. It follows from (4.7) and (H10) that

$$\|U^n-V^n\|\le c\{k^\sigma+m^{-s}\},\quad 0\le n\le\bar p,$$

where V^n is defined by (3.20). We shall prove inductively that there hold:

$$\|U^n-V^n\|\le \tilde c_n\{k^\sigma+m^{-s}\},\quad \bar p\le n\le N, \tag{I_i}$$

$$\tilde c_n=\{1+\tilde c k\}\tilde c_{n-1}+\tilde c k,\quad \bar p+1\le n\le N, \tag{I_{ii}}$$

where the nonnegative constant $\tilde c$ depends only on the IRK method and the constant c in (3.9). An important consequence of (I_{ii}) is that

$$\tilde c_n\le c^*:=(\tilde c_{\bar p}+1)\mathrm{e}^{\tilde cT},\quad \bar p\le n\le N.$$

Now assume that (I_i), (I_{ii}) hold up to $n, \bar p\le n\le N-1$. To extend these to $n+1$, we shall prove inductively that

$$U_\ell^{n,i}\in B_j(M),\quad \ell\ge 0,\ j=1,2,4,\ i=1,\dots,q, \tag{II_i}$$

$$\max_{1\le i\le q}\|\tilde U^{n,i}-U_\ell^{n,i}\|\le (ckKm^\delta)^{2^\ell-1}\max_{1\le i\le q}\|\tilde U^{n,i}-U_\ell^{n,i}\|^{2^\ell},\quad \ell\ge 0, \tag{II_{ii}}$$

where $\{\tilde U^{n,i}\}_{i=1}^q$ are (exact) solutions of (3.3) with $v=U^n, \tilde U^{n+1}=\mathscr{R}U^n$ and where c depends only on the IRK method. Note first that from (I_i) and (H10) it follows that

$$\|U^j-\omega(t^j)\|\le\|U^j-V^j\|+\|V^j-\omega(t^j)\| \tag{4.9}$$

$$\leq (c^* + c)\{k^\sigma + m^{-s}\}, \quad 0 \leq j \leq n.$$

We next verify (II_i) for $\ell = 0$. (Obviously (II_{ii}) holds for $\ell = 0$.) Indeed, from (4.5), (4.9), (3.13),

$$\begin{aligned}
\|\tilde{U}^{n,i} - U_0^{n,i}\| &\leq \|\tilde{U}^{n,i} - \omega^{n,i}\| + \|\omega^{n,i} - \omega(t^{n,i})\| \\
&\leq \Big\| \sum_{j=0}^{\bar{p}} \mu_{ij}^{\bar{p}} \big[\omega(t^{n-j}) - U^{n-j}\big] \Big\| + \Big\| \omega(t^{n,i}) - \sum_{j=0}^{\bar{p}} \mu_{ij}^{\bar{p}} \omega(t^{n-j}) \Big\| \\
&\leq \|\tilde{U}^{n,i} - \omega^{n,i}\| + cc^*\{k^{\bar{p}+1} + m^{-s}\}, \quad i = 1, \ldots, q.
\end{aligned} \tag{4.10}$$

Now from (3.23), (3.8) and (4.9),

$$\|\tilde{U}^{n,i} - \omega^{n,i}\| \leq c\|U^n - \omega(t^n)\| \leq cc^*\{k^\sigma + m^{-s}\}, \quad i = 1, \ldots, q. \tag{4.11}$$

Hence, in view of (iii), (iv) and choosing k small and m large, we obtain

$$|||\tilde{U}^{n,i} - \omega^{n,i}|||_j \leq cc^* m^{s_j}\{k^\sigma + m^{-s}\} \leq \frac{M}{8}, \quad j = 1, 2, 4. \tag{4.12}$$

Thus, from (3.25) and (4.12) it follows that

$$\tilde{U}^{n,i} \in B_j\Big(7\frac{M}{8}\Big), \quad j = 1, 2, 4. \tag{4.13}$$

Now from (4.10) and (4.11), it follows that

$$\|\tilde{U}^{n,i} - U_0^{n,i}\| \leq c\{k^{\bar{p}+1} + m^{-s}\}, \tag{4.14}$$

with $c = Cc^*$ where C does not depend on m, k, n and the induction indices. Choosing k small and m large, $|||\tilde{U}^{n,i} - U_0^{n,i}|||_j \leq \frac{M}{8}$. This, together with (4.13) give the desired result.

Now assume that (II_i) and (II_{ii}) hold up to some $\ell \geq 0$. To show that $\{U_{\ell+1}^{n,i}\}_{i=1}^q$ exist (uniquely) satisfying (4.1), we consider the associated homogeneous system

$$y^i - k \sum_{j=1}^q a_{ij} Df(U_\ell^{n,j}) y^j = 0, \quad i = 1, \ldots, q.$$

Using a diagonalization procedure, it follows from (II_i), (H2) and (H5) that

$$\big(c_1 - c_2 k(\lambda + Km^\beta)\big) \sum_{i=1}^q \|y^i\|^2 \leq 0,$$

for some constants c_1, c_2 depending only on A. Hence, taking kKm^β sufficiently small, according to (v), forces $y^i = 0, i = 1, \ldots, q$.

We shall next prove the estimate

$$\max_{1\leq i\leq q} \|\tilde{U}^{n,i} - U_{\ell+1}^{n,i}\| \leq ckKm^\delta \max_{1\leq i\leq q} \|\tilde{U}^{n,i} - U_\ell^{n,i}\|^2, \tag{4.15}$$

for some $c = c(A, q)$. Indeed, for $i = 1, \dots, q$,

$$
\begin{aligned}
\tilde{U}^{n,i} - U^{n,i}_{\ell+1} &= k \sum_{j=1}^{q} a_{ij} \big[f(\tilde{U}^{n,j}) - f(U^{n,j}_\ell) - Df(U^{n,j}_\ell)(U^{n,j}_{\ell+1} - U^{n,j}_\ell) \big] \\
&= k \sum_{j=1}^{q} a_{ij} \Big[Df(U^{n,j}_\ell)(\tilde{U}^{n,j} - U^{n,j}_{\ell+1}) \\
&\quad + \int_0^1 (1-t) D^2\varphi\big(t\tilde{U}^{n,j} + (1-t)U^{n,j}_\ell\big)[\tilde{U}^{n,j} - U^{n,j}_\ell]^2 \, dt \Big],
\end{aligned} \tag{4.16}
$$

where $D^2\varphi(u)[v]^2 = D^2\varphi(u)[v,v]$. We need to estimate the argument of $D^2\varphi$. From (II_i), (4.13) and for $0 \le t \le 1$, we have

$$\max_{1 \le i \le q} \|t\tilde{U}^{n,i} + (1-t)U^{n,i}_\ell\|_j \le M, \quad j = 1, 2, 4.$$

Thus, applying the diagonalization procedure to (4.16), from (H2), (H5) and (H7) we obtain

$$
\begin{aligned}
c_1 \sum_{i=1}^{q} \|\tilde{U}^{n,i} - U^{n,i}_{\ell+1}\|^2 &\le c_2 k(\lambda + Km^\beta) \sum_{i=1}^{q} \|\tilde{U}^{n,i} - U^{n,i}_{\ell+1}\|^2 \\
&\quad + c_3 k K m^\delta \sum_{i=1}^{q} \|\tilde{U}^{n,i} - U^{n,i}_\ell\|^2 \, \|\tilde{U}^{n,i} - U^{n,i}_{\ell+1}\|,
\end{aligned}
$$

from which (4.15) follows if kKm^β is sufficiently small, with $c = \frac{2qc_3}{c_1}$. This in turn implies that (II_{ii}) holds for $\ell + 1$ as well. We next show that $U^{n,i}_{\ell+1} \in B_j(M)$.

From (II_{ii}) and (4.14),

$$\max_{1 \le i \le q} \|\tilde{U}^{n,i} - U^{n,i}_{\ell+1}\| \le k^{2^{\ell+1}-1} c\big(cKm^\delta\{k^{\bar{p}+1} + m^{-s}\}\big)^{2^{\ell+1}-1}\{k^{\bar{p}+1} + m^{-s}\},$$

for some $c = Cc^*$, with C as above. We choose k and m so that in view of (ii) and (vii) we have $c\big(cKm^\delta\{k^{\bar{p}+1} + m^{-s}\}\big)^{2^{\ell+1}-1} \le 1$. We obtain

$$\max_{1 \le i \le q} \|\tilde{U}^{n,i} - U^{n,i}_{\ell+1}\| \le k^{2^{\ell+1}-1}\{k^{\bar{p}+1} + m^{-s}\}. \tag{4.17}$$

As done before, choosing k small and m large, forces (II_i) to be satisfied for $\ell + 1$. This completes the secondary induction argument (II) and we return to the primary argument (I). Now if $\ell_n \ge \log_2(\sigma - \bar{p} + 1)$, it follows from (4.2) and (4.17) that

$$\|\tilde{U}^{n+1} - U^{n+1}\| \le c(A, b) \max_{1 \le i \le q} \|\tilde{U}^{n,i} - U^{n,i}_{\ell_n}\| \le ck\{k^\sigma + m^{-s}\}. \tag{4.18}$$

Using the triangle inequality, (4.18), (3.9) and (I_i), we obtain

$$
\begin{aligned}
\|U^{n+1} - V^{n+1}\| &\le \|U^{n+1} - \tilde{U}^{n+1}\| + \|\tilde{U}^{n+1} - V^{n+1}\| \\
&\le \big[(1 + ck)\tilde{c}_n + ck\big]\{k^\sigma + m^{-s}\}.
\end{aligned} \tag{4.19}
$$

This establishes both (I_i) and (I_{ii}) and defines $\tilde{c}_n$. (4.8) now follows from (I_i) and (H10). Finally, if (viii) holds, then, from (4.14) and (4.15) we obtain

$$\begin{aligned}\max_{1\le i\le q}\|\tilde{U}^{n,i}-U_1^{n,i}\| &\le ckKm^\delta(c^*)^2\{k^{\bar{p}+1}+m^{-s}\}^2\\ &\le \big((c^*)^2k^{2\bar{p}-\sigma+2}Km^\delta\big)ck\{k^\sigma+m^{-s}\}\\ &\le ck\{k^\sigma+m^{-s}\}.\end{aligned}$$

Hence, we may establish (4.19) and thus (I_i), (I_{ii}) for this case as well. The proof of the theorem is now complete. □

We now consider briefly the practically important issue of generating initial data $U^0,\dots,U^{\bar{p}}$ satisfying (4.7). Indeed, this can be done by a variety of techniques including the use of explicit Runge–Kutta methods or Taylor expansions. The iterative scheme (4.1) can be used as well with the added benefit of the guidance offered by the theoretical framework of Theorem 4.1 In this respect, the relevant considerations are the following

(a) Take $U^0=\omega^0$ (or $(\omega^0+O(m^{-s}))$.
(b) Generate $U_0^{n,i}$ by

$$U_0^{n,i}=\sum_{j=0}^{n}\mu_{ij}^nU^{n-j}=\tilde{U}^{n,i}+O(k^{n+1}+m^{-s}),\quad i=1,\dots,q,\ n=0,\dots,\bar{p}-1.$$

(c) Increase ℓ_n to compensate for the reduced accuracy of the initial approximations $U_0^{n,i}$ and compute U^{n+1} by (4.2), $0\le n\le\bar{p}-1$.
(d) The desired estimates will hold if the two conditions $k^{\bar{p}+1}m^{s_j}\le c_0$ and $k^{\bar{p}+1}m^\delta\le c_0$ are replaced with $k^{n+1}m^{s_j}\le c_0$ and $k^{n+1}m^\delta\le c_0$, respectively. If these conditions become stringent for $n=0$, we recommend the use of more accurate formulas based on Taylor's Theorem such as

$$\begin{aligned}U^{0,i}&=\omega^0+k\tau_i\omega_t^0=\omega^0+k\tau_if(\omega^0)\\ &=\tilde{U}^{0,i}+O(k^2+m^{-s}).\end{aligned}$$

5. Efficient implementation of Newton's method

Newton's scheme, as described in Theorem 4.1 and specifically in its implementation (4.1), requires forming the operator $\mathscr{J}:(H_m)^q\to(H_m)^q$

$$\mathscr{J}=\begin{pmatrix}I-ka_{11}Df(U_\ell^{n,1}) & -ka_{12}Df(U_\ell^{n,2}) & \dots & -ka_{1q}Df(U_\ell^{n,q})\\ -ka_{21}Df(U_\ell^{n,1}) & I-ka_{22}Df(U_\ell^{n,2}) & \dots & -ka_{2q}Df(U_\ell^{n,q})\\ \vdots & \vdots & \ddots & \vdots\\ -ka_{q1}Df(U_\ell^{n,1}) & -ka_{q2}Df(U_\ell^{n,2}) & \dots & I-ka_{qq}Df(U_\ell^{n,q})\end{pmatrix}$$

as well as solving the associated linear system at each new ℓ and n. In practice, this translates into a $q\dim H_m\times q\dim H_m$ system. Obviously, this could be prove to be prohibitively costly when $\dim H_m$ is very large. One possibility that immediately comes to mind is to evaluate $\mathscr{J}$ at $U_0^{n,j}$ and use it according to the iterative procedure

$$U_{\ell+1}^{n,i} - k\sum_{j=1}^{q} a_{ij} Df(U_0^{n,j})(U_{\ell+1}^{n,j} - U_\ell^{n,j}) = U^n + k\sum_{j=1}^{q} a_{ij} f(U_\ell^{n,j}),\ \ \ell = 0,\dots,.$$

The usefulness of this particular approach is limited because we saw in Theorem 4.1 that, under rather general conditions, a single Newton iteration is sufficient to preserve the convergence rate of the base scheme. On the other hand, we may use the same operator over a number of time steps.

It is clear that a great number of strategies are possible for efficient implementation of (4.1). We shall concentrate on evaluating the operators $\mathscr{J}$ at some U_*^n independent of the stage number j. to this end, let U_*^n denote an appropriately chosen element of H_m. Let $U_\ell^{n,i}$ satisfy

$$U_{\ell+1}^{n,i} - k\sum_{j=1}^{q} a_{ij} Df(U_*^n)(U_{\ell+1}^{n,j} - U_\ell^{n,j}) = U^n + k\sum_{j=1}^{q} a_{ij} f(U_\ell^{n,j}), \tag{5.1}$$
$$i = 1,\dots,q,\ \ \ell = 0,\dots,\ell_n - 1.$$

This scheme is known as the "modified Newton method". Now assume that A has distinct eigenvalues $\lambda_1,\dots,\lambda_q$. This is indeed the case for the Gauss–Legendre and the Radau IIA methods, cf. [6]. The decomposition $A = S^{-1}\Lambda S$ naturally induces a decomposition of the system $\mathscr{J} z = b$ whereby q systems $\big(I - k\lambda_i Df(U_*^n)\big)\tilde z_i = \tilde b_i, i = 1,\dots,q$, are to be solved instead. These q systems are independent of each other and can be solved simultaneously on a computer with at least q independent processors. This strategy has been successfully implemented in some specific settings in [10].

Concerning the modified Newton method, we have

Theorem 5.1. *Assume that* (H10) *and the hypotheses of Theorem 3.1 are satisfied and that we are given initial data* $U^0,\dots,U^{\bar p}, \bar p = \min\{p, \sigma - 1\}$, *satisfying*

$$\|U^j - \omega(t^j)\| \le c\{k^\sigma + m^{-s}\}, \quad 0 \le j \le \bar p, \tag{5.2}$$

for some constant c *independent of* k *and* m.

Assume in addition that

(i) (H5) *holds,*

(ii) (H7) *holds and* $\delta < s$,

(iii) $s_1, s_2, s_3 < s$,

(iv) $\ell_n \ge \sigma - \bar p, \quad \bar p \le n \le N - 1.$

Then, there exist $k_0, m_0, c_0 > 0$ *such that for all* $0 < k \le k_0$ *and for all* $m \ge m_0$ *satisfying*

(v) $kKm^\beta \le c_0$,

(vi) $kKm^\delta \le c_0$,

(vii) $k^{\bar p+1} m^{s_j} \le c_0, \quad$ *for* $j = 1, 2, 4$,

(viii) $k^{\bar p+1} m^\delta \le c_0$,

there exists a unique sequence $\{U^n\}_{n=0}^N$, *which for* $\bar p+1 \le n \le N$ *is generated by* (5.1), (4.2) *and* (4.3) *with* $p_n = \bar p$, *and*

$$U_*^n = U^n. \tag{5.3}$$

Furthermore,

$$\max_{0\le n\le N} \|U^n - \omega(t^n)\| \le c\{k^\sigma + m^{-s}\}, \tag{5.4}$$

for some constant c *independent of* k *and* m.

Proof. We shall omit details that would otherwise be repetitions of similar ones exhibited in the proof of Theorem 4.1. Again, we shall use the primary induction hypotheses

$$\|U^n - V^n\| \le \tilde c_n\{k^\sigma + m^{-s}\}, \quad 0 \le n \le N, \tag{I_i}$$

$$\tilde c_n = \{1 + \tilde c k\}\tilde c_{n-1} + \tilde c k, \quad 1 \le n \le N, \tag{I_{ii}}$$

where the nonnegative constant $\tilde c$ depends only on the IRK method and the constant c in (3.9) and (H10). Also, let c^* be as in (4.9).

Assume that (I_i), (I_{ii}) hold up to $n, \bar p \le n \le N-1$. To extend these to $n+1$, we shall prove inductively that

$$U_\ell^{n,i} \in B_j(M), \quad \ell \ge 0, \ j = 1,2,4, \tag{II_i}$$

$$\max_{1\le i\le q} \|\tilde U^{n,i} - U_\ell^{n,i}\| \le k^\ell \max_{1\le i\le q} \|\tilde U^{n,i} - U_0^{n,i}\|^\ell, \quad \ell \ge 0, \tag{II_{ii}}$$

where $\{\tilde U^{n,i}\}_{i=1}^q$ are the (exact) solutions of (3.3) with $v = U^n$. Using arguments similar to those used in the proof of Theorem 4.1, we may prove that, under the stated conditions,

$$U^n, U_0^{n,i} \in B_j(M), \quad j = 1,2,4, \ i = 1,\dots,q. \tag{5.5}$$

Obviously (II_{ii}) holds for $\ell = 0$. Now assume that both (II_i) and (II_{ii}) hold up to some $\ell \ge 0$. We have

$$\begin{aligned}
\tilde U^{n,i} - U_{\ell+1}^{n,i} &= k\sum_{j=1}^q a_{ij}\big[f(\tilde U^{n,j}) - f(U_\ell^{n,j}) - Df(U^n)(U_{\ell+1}^{n,j} - U_\ell^{n,j})\big]\\
&= k\sum_{j=1}^q a_{ij}\Big[Df(U^n)(\tilde U^{n,j} - U_{\ell+1}^{n,j})\\
&+ \big[Df(U_\ell^{n,j}) - Df(U^n)\big](\tilde U^{n,j} - U_\ell^{n,j})\\
&+ \int_0^1 (1-t)D^2 f\big(t\tilde U^{n,j} + (1-t)U_\ell^{n,j}\big)[\tilde U^{n,j} - U_\ell^{n,j}]^2\,dt\Big]\\
&= k\sum_{j=1}^q a_{ij}\Big[Df(U^n)(\tilde U^{n,j} - U_{\ell+1}^{n,j})\\
&+ \int_0^1 (1-t)D^2\varphi\big(t\tilde U_\ell^{n,j} + (1-t)U^n\big)[U_\ell^{n,j} - U^n, \tilde U^{n,j} - U_\ell^{n,j}]\,dt
\end{aligned}$$

$$+ \int_0^1 (1-t) D^2\varphi\big(t\tilde{U}^{n,j} + (1-t)U_\ell^{n,j}\big)[\tilde{U}^{n,j} - U_\ell^{n,j}]^2\, dt\Big].$$

As before, we can show that $\tilde{U}^{n,i} \in B_j(M),\ j = 1,2,4$. Using a diagonalization argument, it follows from (5.5), (II_{ii}), (H2), (H5) and (H7) that

$$c_1 \max_{1\le i\le q} \|\tilde{U}^{n,i} - U_{\ell+1}^{n,i}\| \le c_2 k \max_{1\le i\le q} \big\{(\lambda + Km^\beta)\|\tilde{U}^{n,i} - U_{\ell+1}^{n,i}\| + Km^\delta \|\tilde{U}^{n,i} - U_\ell^{n,i}\|^2 \\ + Km^\delta \|U_\ell^{n,i} - U^n\|\, \|\tilde{U}^{n,i} - U_\ell^{n,i}\|\big\},$$

for some constants c_1, c_2 depending only on the IRK method. Choosing k so that $c_2 k(\lambda + Km^\beta) \le \frac{c_1}{2}$, we obtain

$$\tag{5.6} \max_{1\le i\le q} \|\tilde{U}^{n,i} - U_{\ell+1}^{n,i}\| \le c_3 k K m^\delta \max_{1\le i\le q} \big\{\|\tilde{U}^{n,i} - U_\ell^{n,i}\|^2 \\ + \|U_\ell^{n,i} - U^n\|\, \|\tilde{U}^{n,i} - U_\ell^{n,i}\|\big\}.$$

We can show that for some constant $c_4 = c_4(c^*)$,

$$\|U_\ell^{n,i} - U^n\| \le c_4\{k + m^{-s}\}.$$

Similarly, from (II_{ii}) and (4.14),

$$\max_{1\le i\le q} \|\tilde{U}^{n,i} - U_\ell^{n,i}\| \le c_5\{k^{\bar{p}+1} + m^{-s}\},$$

for some $c_5 = c_5(c^*)$. Choosing k and m so that

$$\tag{5.7} \max\big\{c_3 c_4 K m^\delta (k + m^{-s}), c_3 c_5 K m^\delta (k^{\bar{p}+1} + m^{-s})\big\} \le \frac{1}{2},$$

we obtain (II_{ii}) for $\ell + 1$. Hence, we can now show that $U_{\ell+1}^{n,i} \in B_j(M)$, under the stated conditions.

In view of the fact that $\ell_n \ge \sigma - \bar{p}$, and proceeding exactly as we did in the proof of Theorem 4.1, we can close the primary induction argument, proving the theorem. □

Remark 5.1. Theorem 5.1 requires in particular that km^δ be sufficiently small. This condition may be weakened somewhat by modifying the proof as follows: We choose k and m so that instead of (5.7) we have

$$\tag{5.8} \max\big\{c_3 c_4 K m^\delta (k^{1+\vartheta} + m^{-s}), c_3 c_5 K m^\delta (k^{\bar{p}+1} + m^{-s})\big\} \le \frac{1}{2},$$

with $0 \le \vartheta < 1$, and require $k^{1+\vartheta} m^\delta$ to be small. As a consequence, (II_{ii}) must be modified to

$$\max_{1\le i\le q} \|\tilde{U}^{n,i} - U_\ell^{n,i}\| \le k^{\ell(1-\vartheta)} \max_{1\le i\le q} \|\tilde{U}^{n,i} - U_0^{n,i}\|^\ell, \quad \ell \ge 0.$$

As a result, an increased number of iterations must be performed.

6. A simpler iterative scheme

We shall next consider an iterative scheme where $\mathscr{J}$ is constant and which is sometimes called an "explicit–implicit" type method. This extremely efficient option can be applied however, only when the constant γ in (H6) is zero

$$U^{n,i}_{\ell+1} - k\sum_{j=1}^{q} a_{ij} L U^{n,j}_{\ell+1} = U^n + k\sum_{j=1}^{q} a_{ij}\varphi(U^{n,j}_{\ell}), \quad i = 1,\dots,q,\ \ \ell = 0,\dots,\ell_n - 1. \tag{6.1}$$

From the error equation

$$\begin{aligned}\tilde U^{n,i} - U^{n,i}_{\ell+1} - k\sum_{j=1}^{q} a_{ij} L(\tilde U^{n,j} - U^{n,j}_{\ell+1}) &= k\sum_{j=1}^{q} a_{ij}\big[\varphi(\tilde U^{n,j}) - \varphi(U^{n,j}_{\ell})\big]\\ &= k\sum_{j=1}^{q} a_{ij}\int_0^1 D\varphi(t\tilde U^{n,j} + (1-t)U^{n,j}_{\ell})(\tilde U^{n,j} - U^{n,j}_{\ell})\,dt,\end{aligned}$$

we obtain in view of (H2) and (H6) with $\gamma = 0$,

$$\max_{1\le i\le q}\|\tilde U^{n,i} - U^{n,i}_{\ell+1}\| \le ckK \max_{1\le i\le q}\|\tilde U^{n,i} - U^{n,i}_{\ell}\|.$$

Operating within the framework of an induction argument, we obtain

$$\begin{aligned}\max_{1\le i\le q}\|\tilde U^{n,i} - U^{n,i}_{\ell}\| &\le (cK)^{\ell_n}(kc^*)k^{\ell_n - 1}\{k^{\bar p+1} + m^{-s}\}\\ &\le (cK)^{\ell_n} k\{k^{\sigma} + m^{-s}\},\end{aligned}$$

for $kc^* \le 1$ and $\ell_n \ge \sigma - \bar p + 1$. We have,

Theorem 6.1. *Assume that* (H10) *and the hypotheses of Theorem 3.1 are satisfied and that we are given initial data* $U^0,\dots,U^{\bar p}, \bar p = \min\{p, \sigma - 1\}$*, satisfying*

$$\|U^j - \omega(t^j)\| \le c\{k^{\sigma} + m^{-s}\}, \quad 0 \le j \le \bar p,$$

for some constant c independent of k and m.

Assume in addition that

(i) (H6) *holds and* $\gamma = 0$,

(ii) $s_1, s_2, s_3, s_4 < s$,

(iii) $\ell_n \ge \sigma - \bar p + 1, \quad \bar p \le n \le N - 1.$

Then, there exist $k_0, m_0, c_0 > 0$ such that for all $0 < k \le k_0$, and for all $m \ge m_0$ satisfying

(iv) $k^{\bar p+1} m^{s_j} \le c_0$ *for* $j = 1, 2, 3, 4$,

there exists a unique sequence $\{U^n\}_{n=0}^{N}$ which for $\bar p + 1 \le n \le N$ is generated by (6.1), (4.2) *and* (4.3) *with $p_n = \bar p$. Furthermore,*

$$\max_{0\le n\le N}\|U^n - \omega(t^n)\| \le c\{k^{\sigma} + m^{-s}\},$$

for some constant c independent of k and m. □

7. Examples

Let Ω be an open, bounded, connected subset of $\mathbb{R}^d$. For integer $\mu \geq 0$ and $p \in [0, \infty]$, let $W^{\mu,p} = W^{\mu,p}(\Omega)$ denote the usual Sobolev spaces of complex-valued functions defined on Ω and having generalized derivatives up to order μ in $L^p(\Omega)$. The norm on $W^{\mu,p}$ will be denoted by $\|\cdot\|_{\mu,p}$. In particular, $L^p = W^{0,p}(\Omega)$ and for $p = 2$ we let $H^\mu = W^{\mu,2}$. We let $\|\cdot\| = \|\cdot\|_{0,2}$ and $\|\cdot\|_\mu = \|\cdot\|_{\mu,2}$. In some specific instances, as in the case of the KdV equation below, we shall restrict attention to the real-valued functions.

7.1. **The Korteweg–de Vries equation.** We consider the problem of approximating 1−periodic solutions of the KdV equation

$$\begin{cases} u_t + uu_x + u_{xxx} = 0, & 0 \leq x \leq 1, \quad 0 \leq t \leq T, \\ u(x, 0) = u^0(x), & 0 \leq x \leq 1, \end{cases} \tag{7.1.1}$$

where u^0 is a sufficiently smooth 1−periodic function, i.e. $u^0 \in H^\mu_{\text{per}} = W^{\mu,2}_{\text{per}}$ for μ sufficiently large, where for $\mu \geq 1, 1 \leq p \leq \infty$,

$$W^{\mu,p}_{\text{per}} = \{v \in W^{\mu,p}(0,1) : v^{(j)}(0) = v^{(j)}(1),\ 0 \leq j \leq \mu - 1\}.$$

For the existence, uniqueness and regularity of solutions of (7.1.1) we refer to [5]. Specifically, it is known that if $u^0 \in H^\mu_{\text{per}}, \mu \geq 3$, then there exists a solution $u : [0, T] \to H^\mu_{\text{per}}$, $\forall T > 0$. Moreover, for $j \geq 0$ such that $\mu - 3j \geq 0$,

$$\sup_{0 \leq t \leq T} \left\|\frac{d^j u}{dt^j}\right\|_{\mu - 3j} \leq c(\|u^0\|_\mu). \tag{7.1.2}$$

There is a large body of work devoted to the numerical approximation of solutions of the KdV equation, including finite difference, finite element as well as spectral methods. Herein, we operate within the framework already established in [3], [8] and [12]. In particular, the analysis of convergence of the base scheme is drawn from [12].

For integer $r \geq 3$, let $S^r_h \subset H^{r-1}_{\text{per}} \cap W^{2,\infty}_{\text{per}}$ denote the space of 1−periodic splines of degree $\leq r - 1$, defined on a uniform partition $x_j = jh, j = 0, \ldots, m$, of $[0, 1]$, with $h = \frac{1}{m}$. It is known that $\dim S^r_h = m$. We set $H_m = S^r_h$ and equip it with the L^2 inner product

$$(v, w)_m = (v, w) = \int_0^1 v(x) w(x)\, dx, \quad \forall v, w \in S^r_h.$$

The spaces $\{S^r_h\}_{h>0}$ possesses the following approximation property: For each $v \in H^r_{\text{per}}$, there exists $\chi \in S^r_h$ such that

$$\sum_{j=0}^{\mu-1} h^j \|v - \chi\|_j \leq ch^\mu \|v\|_\mu, \quad 1 \leq \mu \leq r, \tag{7.1.3}$$

for some constant c independent of h and v. If in addition $v \in W^{2,\infty}_{\text{per}}$, then

$$\sum_{j=0}^{1} h^j \|v - \chi\|_{j,\infty} \leq ch^2 \|v\|_{2,\infty}. \tag{7.1.4}$$

Moreover, the spaces S_h^r possess the following *inverse* properties

$$\|\chi\|_\beta \le ch^{-(\beta-\alpha)}\|\chi\|_\alpha, \quad 0 \le \alpha \le \beta \le r-1, \tag{7.1.5}$$

$$\|\chi\|_{\alpha,\infty} \le ch^{-(\alpha+\frac{1}{2})}\|\chi\|, \quad 0 \le \alpha \le r-1. \tag{7.1.6}$$

As basis for S_h^r, we use a set of *modified* basis functions $\tilde\varphi_1, \dots, \tilde\varphi_m$ associated with the nodes $x_1, \dots, x_m$ (cf. [14]). For $v \in H^1_{\text{per}}$, we define the *quasi-interpolant* $\tilde v$ by

$$\tilde v(x) = \sum_{j=1}^m v(x_j)\tilde\varphi_j(x).$$

It is known (cf. [14]) that the quasi-interpolant enjoys the following optimal approximation property: For $v \in H^r_{\text{per}}$,

$$\|v - \tilde v\| \le ch^r\|v\|_r. \tag{7.1.7}$$

For $r \ge 3$, let $u : [0,T] \to H^r_{\text{per}}$ be the solution of (7.1.1) and let $\omega = \omega_h : [0,T] \to S_h^r$ denote the quasi-interpolant $\omega(x,t) = \sum_{j=1}^m u(x_j,t)\tilde\varphi_j(x)$. It is shown in [8] that

$$(\omega_t + \omega\omega_x, \chi) - (\omega_{xx}, \chi_x) = (\varepsilon(t), \chi) \quad 0 \le t \le T, \ \forall x \in S_h^r, \tag{7.1.8}$$

where $\varepsilon : [0,T] \to S_h^r$ is a (small) smooth function (truncation error),

Define the operators $L_h, \varphi_h : S_h^r \to S_h^r$ by

$$(L_h v, \chi) = (v_{xx}, \chi_x) \quad \forall \chi \in S_h^r,$$
$$(\varphi_h(v), \chi) = -(vv_x, \chi) \quad \forall \chi \in S_h^r,$$

respectively. Note that $\varphi_h(v) = -P_0(vv_x)$ where P_0 denotes the L^2-orthogonal projection operator onto S_h^r. We may rewrite (7.1.8) as

$$\omega_t = L_h\omega + \varphi_h(\omega) + \varepsilon(t), \quad 0 \le t \le T. \tag{7.1.9}$$

Having cast our problem in the form of (2.1), we next undertake the systematic verification of the hypotheses (H1)–(H10).

With $s = r$, (H1) is proved in [8] (inequality (1.33)). It easily follows from periodicity that (H2) and (H3) hold with $\lambda = \eta = 0$.

We set all four norms $\|\!|\cdot|\!\|_i$ equal to $\|\cdot\|_{1,\infty}$. It then follows from (7.1.6) that (2.2) holds with $s_i = \frac{3}{2}, i = 1,2,3,4$. We also let

$$M = 2\sup_{0\le t\le T}\big[c\|u(t)\|_r + c\|u(t)\|_{2,\infty} + \|u(t)\|_{1,\infty}\big], \tag{7.1.10}$$

where u is the solution of (7.1.1) and c is a constant depending on the constants in (7.1.3), (7.1.4), (7.1.6) and (7.1.7). Also, in verifying hypotheses (H4)–(H7), we shall use different constants K_i and then set $K = \max\{K_1, K_2, K_3, K_4\}$.

Integrating by parts and using periodicity, we obtain

$$\begin{aligned}(\varphi_h(v) - \varphi_h(w), v - w) &= -\frac{1}{2}(v_x, [v-w]^2)\\ &\le \frac{1}{2}\|v\|_{1,\infty}\,\|v - w\|^2 \quad \forall v, w \in S_h^r.\end{aligned}$$

Hence we see that (H4) holds with $K_1 = \frac{M}{2}$.

Using $Dg(x)y = \lim_{\varepsilon\to 0}[g(x+\varepsilon y)-g(x)]$, we see that $D\varphi_h(v)w = -P_0[(vw)_x]$, $\forall v, w \in S_h^r$. Hence,

$$(D\varphi_h(v)w, w) = -((vw)_x, w) = -\frac{1}{2}(v_x, w^2)$$
$$\leq \frac{1}{2}\|v\|_{1,\infty}\,\|w\|^2.$$

So we see that (H5) holds with $K_2 = \frac{M}{2}$ and $\beta = 0$. Now, using (7.1.4),

$$\begin{aligned}\|D\varphi_h(v)w\| &\leq \|(vw)_x\|\\ &\leq \big(\|v\|_{1,\infty} + ch^{-1}\|v\|_{0,\infty}\big)\|w\|\\ &\leq ch^{-1}\|v\|_{1,\infty}\,\|w\|.\end{aligned}$$

Hence, we see that (H6) holds with $K_3 = cM$ and $\gamma = 1$.

Further, $D^2\varphi_h(z)[v,w] = -P_0[(vw)_x]$ for $z, v, w \in S_h^r$. Hence, we easily obtain

$$\|D^2\varphi_h(z)[v,w]\| \leq ch^{-\frac{3}{2}}\|v\|\,\|w\|,$$

where c depends on the constants in (7.1.5) and (7.1.6). Thus, (H7) holds with $K_4 = c$ and $\delta = \frac{3}{2}$.

Now for $0 \leq t \leq T$, choosing $\chi \in S_h^r$ suitably and using (7.1.3), (7.1.4), (7.1.6) and (7.1.7), we obtain

$$\begin{aligned}\|\omega - u\|_{1,\infty} &\leq \|\omega-\chi\|_{1,\infty} + \|\chi - u\|_{1,\infty}\\ &\leq ch^{-\frac{3}{2}}\|\omega-\chi\| + ch\|u\|_{2,\infty}\\ &\leq ch^{-\frac{3}{2}}\big\{\|\omega-u\| + \|u-\chi\|\big\} + ch\|u\|_{2,\infty}\\ &\leq ch^{r-\frac{3}{2}}\|u\|_r + ch\|u\|_{2,\infty}.\end{aligned}\tag{7.1.11}$$

It then follows from the triangle inequality that

$$\sup_{0\leq t\leq T}\|\omega\|_{1,\infty} \leq \frac{M}{2}.$$

Hence, (H8) is satisfied in view of (7.1.10). Indeed, this motivates our choice of M. (H9) is inequality (1.35) in [8].

As for (H10), it is proved in [12] that the (temporal) rate of convergence of the base scheme is the *classical* rate $\sigma = \nu$. The results of Sections 4 and 5 apply, yielding approximations U^n satisfying $\max_{0\leq n\leq N}\|U^n - \omega(t^n)\| \leq c(k^\sigma + h^r)$. Hence, from (7.1.7) and the triangle inequality it follows that

$$\max_{0\leq n\leq N}\|u(t^n) - U^n)\| \leq c\{k^\sigma + h^r\},$$

where u is the solution of (7.1.1).

Let us note that the above results require certain relations between k and h to hold. Specifically, Theorem 4.1 requires

$$kh^{-\frac{3}{2(\bar p+1)}} \leq c_0',\tag{7.1.12}$$

for sufficiently small c_0'. This is a mild condition except for the case $\bar{p} = 0$ corresponding e.g. to the Backward Euler method. Also, (7.1.12) guarantees that taking $\ell_n = 1$ in Newton's method will suffice.

On the other hand, condition (vi) of Theorem 5.1 translates into the requirement $kh^{-3/2}$ be sufficiently small. We may weaken this restriction say to kh^{-1} small by taking $\vartheta = \frac{1}{2}$ in Remark 5.1. This will however come at the expense of doubling the number of iterations.

7.2. **The nonlinear Schrödinger equation.** We consider the problem of approximating the complex-valued solution u of the following initial and boundary value problem for the cubic Schrödinger equation:

$$\begin{cases} u_t = \mathrm{i}\Delta u + \mathrm{i}|u|^2 u, & \text{in } \bar{\Omega} \times [0,T], \\ u = 0, & \text{on } \partial\Omega \times [0,T], \\ u(x,0) = u^0(x), & \text{in } \bar{\Omega}, \end{cases} \tag{7.2.1}$$

where Ω is an open, bounded, connected subset of $\mathbb{R}^d$ and u^0 is a given complex-valued function defined on $\bar{\Omega}$. We assume that (7.2.1) possesses a unique solution u which is sufficiently smooth up to $\partial\Omega$.

We shall operate within the framework established in [11]. In particular, we shall use the space $C(\bar{\Omega})$ of continuous, complex-valued functions defined on $\bar{\Omega}$, and let H_0^1 denote the subspace of H^1 consisting of those functions that vanish on $\partial\Omega$ in the sense of trace.

For integer $r \geq 2$ and $0 < h < 1$, $Z_h^r \subset H^1 \cap C^0(\bar{\Omega})$ will represent an approximating finite-dimensional space of functions. Such spaces typically consist of piecewise polynomial functions of degree $\leq r-1$ defined on a suitable partition of Ω. Note that the elements of Z_h^r are complex-valued. In particular, we assume that $Z_h^r = S_h^r + \mathrm{i}S_h^r$ where S_h^r is an approximating space of real-valued functions. Indeed, the properties of Z_h^r listed below are all derived from corresponding properties of S_h^r.

We assume that these spaces possess good approximation properties; indeed that there exists a constant c independent of h such that for each $v \in H^r \cap H_0^1$, there exists $\chi \in Z_h^r$ such that

$$\|v - \chi\| \leq ch^r \|v\|_r, \tag{7.2.2}$$

and if in addition $v \in W^{2,\infty}(\Omega)$, then

$$\|v - \chi\|_{L^\infty} \leq ch^2 \|v\|_{2,\infty}. \tag{7.2.3}$$

We shall assume that the elements of Z_h^r satisfy the following *inverse* inequalities

$$\|\chi\|_{0,\infty} \leq ch^{-d/2} \|\chi\|, \tag{7.2.4}$$

$$\|\chi\|_1 \leq ch^{-1} \|\chi\|. \tag{7.2.5}$$

Let $V = Z_h^r + (H^2 \cap H_0^1)$. We assume the existence of a family of sesquilinear forms $B_h^r : V \times V \to \mathbb{C}$ with the following properties:

$$B_h^r(v,v) \ \text{ is real for } v \in V, \tag{7.2.6}$$

$$B_h^r(v,v) \geq c\|v\|_1^2 \ \text{ for } c > 0, \forall v \in Z_h^r, \tag{7.2.7}$$

$$B_h^r(v,\chi) = -(\Delta v, \chi) \quad \forall \chi \in Z_h^r, \quad v \in H^2 \cap H_0^1. \tag{7.2.8}$$

With B_h^r we associate an *elliptic projection* operator $P_E : H^2 \cap H_0^1 \to Z_h^r$ by

$$B_h^r(P_E v, \chi) = B_h^r(v,\chi) = -(\Delta v, \chi) \quad \forall \chi \in Z_h^r. \tag{7.2.9}$$

We assume that for some constant c independent of h

$$\|P_E v - v\| \leq c h^r \|v\|_r \qquad \forall v \in H^r \cap H_0^1. \tag{7.2.10}$$

The most well-known family of such sesquilinear forms is provided by the so-called standard Galerkin method. In this case $Z_h^r \subset H_0^1$ and

$$B_h^r(v,w) = \int_\Omega \nabla v \cdot \nabla \bar{w}\, dx.$$

Let $u_h : [0,T] \to Z_h^r$ denote the elliptic projection $P_E u$ of the solution of (7.2.1). Then

$$(u_{ht}, \chi) = -\mathrm{i} B_h^r(u_h,\chi) + \mathrm{i}(|u_h|^2 u_h + \psi(t), \chi), \tag{7.2.11}$$

where $\psi = P_0[u_{ht} - u_t - \mathrm{i}(|u_h|^2 u_h - |u|^2 u)]$ and P_0 denotes the L^2-orthogonal projection operator onto Z_h^r. Then ψ satisfies

$$\sup_{0\leq t\leq T} \|\frac{d^j\psi}{dt^j}\| \leq c_j h^r, \quad j = 0,1,\ldots. \tag{7.2.12}$$

To prove this one just needs to note that

$$\sup_{0\leq t\leq T} \|\frac{d^j u_h}{dt^j}\|_{0,\infty} \leq c_j, \quad j = 0,1,\ldots,$$

for c_j independent of h under the hypothesis that $r > \frac{d}{2}$. Set

$$H_m = S_h^r \times S_h^r, \quad m = h^{-d}, \quad s = \frac{r}{d}.$$

We equip H_m with the inner product

$$(v,w) = (v,w)_{H_m} = \int_\Omega (v_1 w_1 + v_2 w_2)\, dx, \quad v = (v_1,v_2)^T, w = (w_1,w_2)^T \in H_m,$$

and associated norm $\|v\| = \|v\|_{H_m} = (v,v)^{1/2}$. We define the operator $\Delta_h : S_h^r \to S_h^r$ by

$$(\Delta_h v, \chi) = -B_h^r(v,\chi), \quad \forall \chi \in S_h^r,$$

and thence the operator $L : H_m \to H_m$ by

$$L = \begin{pmatrix} 0 & -\Delta_h \\ \Delta_h & 0 \end{pmatrix}.$$

Now consider the function $g : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$g(x,y) = \big(g_1(x,y), g_2(x,y)\big)^T = \big(-(x^2+y^2)y, (x^2+y^2)x\big)^T.$$

g naturally induces a map $\varphi(v_1,v_2) = \big(\varphi_1(v_1,v_2), \varphi_2(v_1,v_2)\big)^T : H_m \to H_m$ where

$$\begin{aligned}(\varphi_1(v_1,v_2),\chi) &= (g_1(v_1,v_2),\chi), \quad \forall \chi \in S_h^r,\\ (\varphi_2(v_1,v_2),\chi) &= (g_2(v_1,v_2),\chi), \quad \forall \chi \in S_h^r.\end{aligned}$$

With the maps $\omega, \varepsilon : [0,T] \to H_m$ given by $\omega = (\operatorname{Re} u_h, \operatorname{Im} u_h)^T, \varepsilon = (\operatorname{Re}\psi, \operatorname{Im}\psi)^T$, we see that (7.2.11) can be written in the equivalent form

$$\omega_t = L\omega + \varphi(\omega) + \varepsilon(t), \quad 0 \le t \le T, \tag{7.2.13}$$

which is the required form (2.1).

However, it turns out that φ does not satisfy hypothesis (H4). In order to overcome this difficulty, we introduce a map $\tilde{\varphi} : H_m \to H_m$ as follows: Let $z \in C_0^\infty(\mathbb{R})$ be a *cutoff* function

$$z(\xi) = \begin{cases} 1 & |\xi| \le M, \\ 0 & |\xi| \ge 2M. \end{cases}$$

We let $\tilde{g} : \mathbb{R}^2 \to \mathbb{R}^2$ be given by

$$\begin{aligned}\tilde{g}(x,y) &= \big(\tilde{g}_1(x,y), \tilde{g}_2(x,y)\big)^T \\ &= \big(-z(\xi)(x^2+y^2)y, z(\xi)(x^2+y^2)x\big)^T, \quad \xi = (x^2+y^2)^{1/2}.\end{aligned}$$

Now let $\tilde{\varphi}$ be the map naturally induced by $\tilde{g}$

$$\tilde{\varphi}(v) = z(\xi)\big(-P_0(v_1^2+v_2^2)v_2, P_0(v_1^2+v_2^2)v_1\big)^T,$$

$v = (v_1,v_2)^T \in H_m, \xi = (v_1^2+v_2^2)^{1/2}$.

We shall show below that ω also satisfies the equation

$$\omega_t = L\omega + \tilde{\varphi}(\omega) + \varepsilon(t), \quad 0 \le t \le T. \tag{7.2.14}$$

Now (H1) follows from (7.2.12) and the fact that $m^{-s} = h^r$. Also, it is easily seen that $(Lv,v) = (\tilde{\varphi}(v),v) = 0$, $\forall v \in H_m$. Thus (H2) and (H3) hold with $\lambda = \eta = 0$, respectively.

Also, setting

$$\|v\|_i = \|v\|_{0,\infty} = \max\{\|v_1\|_{0,\infty}, \|v_2\|_{0,\infty}\}, \quad i = 1,2,3,4, \tag{7.2.15}$$

for $v = (v_1,v_2)^T \in H_m$, we see that (2.2) holds with $s_i = \frac{1}{2}, i = 1,2,3,4$.

Now set

$$M = 2 \sup_{0\le t\le T} \big[c\|u(t)\|_r + c\|u(t)\|_{2,\infty} + \|u(t)\|_{0,\infty}\big],$$

where u is the solution of (7.2.1) and c is a constant depending on the constants in (7.2.2), (7.2.3) and (7.2.4).

It is clear that $\tilde{g}$ and its derivatives of arbitrary order are bounded on $\mathbb{R}^2$. Hence it follows at once that $\tilde{\varphi}$ satisfies hypothesis (H4) without the stipulation that the argument of $\tilde{\varphi}$ belong to $B_1(M)$.

Also, for $v, w \in H_m$ with $\xi = (w_1^2 + w_2^2)^{1/2}$,

$$D\tilde{\varphi}(w)v = \begin{pmatrix} -P_0\{[z'(\xi)\xi + 2z(\xi)]w_1w_2v_1 + [z'(\xi)\xi w_1w_2 + z(\xi)(3w_2^2 + w_1^2)]v_2\} \\ P_0\{[z'(\xi)\xi w_1^2 + z(\xi)(3w_1^2 + w_2^2)]v_1 + [z'(\xi)\xi + 2z(\xi)]w_1w_2v_2\} \end{pmatrix}.$$

It is easy to see that (H5) and (H6) hold with $\beta = 0$ and $\gamma = 0$, respectively, again without restriction on the argument of $\tilde{\varphi}$.

Let $z, v, w \in H_m$ with $|||z|||_4 \le M$. Then,

$$D^2\tilde{\varphi}(z)[v, w] = D^2\varphi(z)[v, w] = 2\begin{pmatrix} -P_0\{z_1w_2v_1 + z_2w_1v_1 + 3z_2w_2v_2 + z_1w_1v_2\} \\ P_0\{3z_1w_1v_1 + z_2w_2v_1 + z_1w_2v_2 + z_2w_1v_2\} \end{pmatrix}.$$

From (7.2.4)

$$\|z_iv_jw_\ell\| \le M\|v_j\|_{0,\infty}\,\|w_\ell\| \le ch^{-d/2}\|v_j\|\,\|w_\ell\|;$$

so (H7) holds with $\delta = \frac{1}{2}$.

To ascertain (H8), proceeding as we did in the case of the KdV equation, we obtain from (7.2.2), (7.2.4) and (7.2.10),

$$\|u_h - u\| \le ch^2\|u\|_{2,\infty} + ch^{r-d/2}\|u\|_r.$$

Since $\|\omega\|_{0,\infty} \le \|u_h\|_{0,\infty}$, for $r \ge \frac{d}{2}$ we obtain

$$\sup_{0\le t\le T}\|\omega(t)\|_{0,\infty} \le \sup_{0\le t\le T}\left[c\|u(t)\|_{2,\infty} + c\|u(t)\|_r + \|u(t)\|_{0,\infty}\right] \le \frac{M}{2}.$$

In view of (7.2.15), this not only establishes (H8), but also shows that (7.2.14) is satisfied. Also, since the operators $\frac{d}{dt}$ and P_E commute, we may easily verify (H9) using (7.2.10).

To obtain the results of Sections 3, 4, 5 and 6, we argue as follows: In the case of Theorem 3.1, given any V^0 satisfying (3.24), we obtain the existence of a unique sequence $\{\{V^{n,i}\}_{i=1}^q, V^{n+1}\}_{n=0}^{N-1}$ satisfying (3.20) with $f = L + \tilde{\varphi}$, with $\{V^n\}_{n=0}^N$ satisfying (3.21). In view of (3.26), (7.2.15) and the definition of $\tilde{\varphi}$, it follows that $V^0, \{\{V^{n,i}\}_{i=1}^q, V^{n+1}\}_{n=0}^{N-1}$ is the solution of the base scheme. Furthermore, it is proved in [11] that the following improved estimate holds

$$\max_{0\le n\le N}\|\omega(t^n) - V^n\| \le c\{k^\sigma + h^r\},$$

where the integer σ is given by

$$\sigma = \begin{cases} \nu & \text{if } \Omega \text{ is polyhedral or } d = 1, \\ \min\{p+3, \nu\} & \text{otherwise.} \end{cases}$$

Let us note here that these results require the condition $r > \frac{d}{2}$ and $k^\sigma h^{-d/2} \le c_0$.

A similar reasoning can be applied to the results of Sections 4, 5 and 6. Indeed, all of these apply with $f = L + \tilde{\varphi}$. Recall that a cornerstone of the proofs was the fact that $U_\ell^{n,i} \in B_1(M)$, and in addition $U^n = U^* \in B_1(M)$ in the case of Theorem 5.1. Since $\varphi(v) = \tilde{\varphi}(v), \forall v \in B_1(M)$, the conclusions of Theorems 4.1, 5.1 and 6.1 remain in force for $f = L + \varphi$ as well. Furthermore, the iterative procedures (4.1), (5.1) and (6.1) involve linear systems that are invertible under their respective prevailing conditions. Hence, the schemes outlined have unique solutions, which may be calculated by using either φ or $\tilde{\varphi}$. Obviously, it would be more convenient to use φ, in which case, $D\varphi$ would be given by

$$D\varphi(w)v = 2\begin{pmatrix} -P_0[2w_1w_2v_1 + (3w_2^2 + w_1^2)v_2] \\ P_0[(3w_1^2 + w_2^2)v_1 + 2w_1w_2v_2] \end{pmatrix}.$$

Finally, using (7.2.10) and the triangle inequality, we obtain convergence of the numerical approximations U^n to $u(t^n)$ at the rate $O(k^\sigma + h^r)$.

Of course the conditions of Theorems 4.1, 5.1 and 6.1 hold, under the guise of specific constraints on k, h, r, d. In particular, the conditions $s_j < s$ translate into $r > \frac{d}{2}$, which was a basic assumption for the convergence of the base scheme. In addition, we also require that $k^{\bar{p}+1}h^{-d/2} \le c_0$. This is slightly more restrictive than the condition $k^\sigma h^{-d/2} \le c_0$. For $d \le 3$ and $\bar{p} \ge 1$ a mild condition of the type $k = o(h^{3/4})$ must be satisfied. Also, Newton's method will require only one iteration under the condition that k be taken sufficiently small. On the other hand, condition (v) in Theorem 5.1 is equivalent to $kh^{-d/2}$ being sufficiently small, which could be restrictive for $d = 3$. Hence, the Explicit–Implicit iteration could provide a better alternative.

References

1. G. D. Akrivis, V. A. Dougalis and O. A. Karakashian, *On fully discrete Galerkin methods of second-order temporal accuracy for the nonlinear Schrödinger equation*, Numer. Math. **59** (1991) 31–53.
2. R. Alexander, *The modified Newton method in the solution of stiff ordinary differential equations*, Math. Comp. **57** (1991) 673–701.
3. G. A. Baker, V. A. Dougalis and O. A. Karakashian, *Convergence of Galerkin approximations for the Korteweg–de Vries equation*, Math. Comp. **40** (1983) 419–433.
4. J. L. Bona, V. A. Dougalis, O. A. Karakashian and W. R. McKinney, *Conservative high-order numerical methods for the generalized Korteweg–de Vries equation*, Phil. Trans. Roy. Soc. London Ser. A **351** (1995) 107–164.
5. J. L. Bona and R. Smith, *The initial value problem for the Korteweg–de Vries equation*, Phil. Trans. Roy. Soc. London Ser. A **298** (1975) 555–604.
6. J. C. Butcher, *The numerical analysis of ordinary differential equations; Runge-Kutta methods and general linear methods*, John Wiley, Chichester, 1987.
7. M. Crouzeix, W. H. Hundsdorfer and M. N. Spijker, *On the existence of solutions to the algebraic equations in implicit Runge–Kutta methods*, BIT **23** (1983) 84–91.
8. V. A. Dougalis and O. A. Karakashian, *On some high order accurate fully discrete Galerkin methods for the Korteweg–de Vries equation*, Math. Comp. **45** (1985) 329–345.
9. E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II: Stiff and Differential–Algebraic Problems*, Springer–Verlag, Berlin Heidelberg, Springer Series in Computational Mathematics v. 14, 1991.
10. O. Karakashian and W. Rust, *On the parallel implementation of implicit Runge–Kutta methods*, SIAM J. Sci. Sta. Comput. **9** (1988) 1085–1090.
11. O. Karakashian, G. D. Akrivis and V. A. Dougalis, *On optimal-order error estimates for the nonlinear Schrödinger Equation*, SIAM J. Numer. Anal. **30** (1993) 377–400.
12. O. Karakashian and W. McKinney, *On optimal high order in time approximations for the Korteweg–de Vries equation*, Math. Comp. **55** (1990) 473–496.
13. J. M. Sanz-Serna and D. F. Griffiths, *A new class of results for the algebraic equations of implicit Runge–Kutta processes*, IMA J. Numer. Anal. **11** (1991) 449–455.

14. V. Thomée and B. Wendroff, *Convergence estimates for Galerkin methods for variable coefficient initial value problems*, SIAM J. Numer. Anal. **11** (1974) 1059–1068.

Computer Science Department, University of Ioannina, 451 10 Ioannina, Greece

Mathematics Department, University of Athens, 15784 Zographou, Greece and IACM, FORTH

Mathematics Department, University of Tennessee, Knoxville TN 37996-1300