ELSEVIER

Applied Numerical Mathematics 17 (1995) 97–118

APPLIED NUMERICAL MATHEMATICS

Convergence of Runge–Kutta methods for differential-algebraic systems of index 3

Laurent Jay *

Department of Computer Science, University of Minnesota, 4-192 EE/CS Building, 200 Union Street, Minneapolis, MN 55455-0159, USA

Abstract

This article deals with convergence results of stiffly accurate implicit Runge–Kutta methods when applied to differential-algebraic equations of index 3 in Hessenberg form. Under certain hypotheses global superconvergence is shown, proving a result conjectured by Hairer, Lubich and Roche. Numerical examples are provided which illustrate the theoretical results.

Keywords: Differential-algebraic equations; Index 3; Runge–Kutta methods

1. Introduction

This article presents optimal convergence results for stiffly accurate implicit Runge–Kutta (RK) methods when applied to semi-explicit index-3 differential-algebraic equations (DAEs) in Hessenberg form. Index-3 problems frequently arise in the modelling of constrained mechanical systems (for details see [2, Section 6.2], [7, pp. 6–7], and [9, pp. 483–486 and 539–540]). For solving such problems an index reduction is usually possible by differentiating the constraints, although some difficulties may occur (see [2, Sections 2.5.3 and 5.4.1] for a detailed discussion). However, for multibody systems containing very stiff springs, i.e., whose Hooke's constant $1/\varepsilon^2$ is very large, the numerical solution behaves like that for the limit problem ($\varepsilon \to 0$) which is of index 3 (see [7, pp. 10–12] and [13]). In this situation an index reduction is not applicable and the convergence behaviour for the index-3 case must be studied. This remark remains valid for the equations of motion of very stiff mechanical systems in which a large potential forces the motion to be close to a manifold (see [13]).

Convergence results have been obtained for BDF methods (see [2, Section 3.2.4] and [3]) and for implicit RK methods applied to solvable linear constant coefficients systems of arbitrary

* E-mail: na.ljay@na-net.ornl.gov.

SSDI 0168-9274(95)00013-5

index (see [4]). Non-optimal orders of convergence of RK methods for semi-explicit index-3 DAEs in Hessenberg form have been demonstrated in [7, Section 6] and sharper estimates have been numerically observed and hypothesized (see [7, pp. 18–19 and 86]). The main result of this article (Theorem 6.1 below) is a proof of the conjecture of Hairer, Lubich and Roche [7, p. 86], giving sharp convergence bounds for stiffly accurate RK methods, such as the Lobatto IIIC and Radau IIA schemes. It generalizes a well-known result of Butcher [5] for ordinary differential equations to semi-explicit index-3 DAEs in Hessenberg form. This result has an application in the convergence analysis of these methods when applied to stiff mechanical systems (see [13, Theorem 3.1]). Furthermore it extends the results of [11] for collocation methods to general implicit RK methods, but with completely different techniques. New insight is provided for the structure of the global error.

In this paper we consider the following system of semi-explicit index-3 DAEs given in an autonomous and Hessenberg form

$$y' = f(y, z), \tag{1.1a}$$

$$z' = k(y, z, u), \tag{1.1b}$$

$$0 = g(y), \tag{1.1c}$$

where the initial values (y_0, z_0, u_0) at x_0 are assumed to be *consistent*, i.e., they satisfy

$$0 = g(y), \tag{1.1c}$$

$$0 = (g_y f)(y, z), \tag{1.1d}$$

$$0 = (g_{yy}(f, f) + g_y f_y f + g_y f_z k)(y, z, u). \tag{1.1e}$$

We suppose f, g, and k to be sufficiently differentiable, and that

$$(g_y f_z k_u)(y, z, u) \quad \text{is invertible} \tag{1.2}$$

in a vicinity of the exact solution (*index*-3 assumption). In a neighbourhood of any fixed values satisfying (1.1e) and (1.2), (1.1e) defines an implicit function for u, i.e., $u = G(y, z)$.

The application of Runge–Kutta methods to (1.1a)–(1.1c) is presented in Section 2. Existence and uniqueness of the RK solution, and influence of perturbations are studied in Section 3. New improved estimates for the perturbed solution are given. They are essential to obtain sharp convergence results. Section 4 deals with the calculation of expressions encountered in the preceding section and involving the RK coefficients. Optimal local error estimates for the y-component and a certain projection of the z-component are then given in Section 5. For the z-component this requires the use of the new technique of DA3-series. With the help of the results contained in the previous sections, a global convergence theorem is presented in Section 6 proving the conjecture of Hairer, Lubich and Roche [7, p. 86]. Finally, Section 7 includes some numerical experiments illustrating the theoretical results.

2. Runge–Kutta methods for semi-explicit index-3 DAEs

Definition 2.1. One step of an *s-stage Runge–Kutta* (RK) method applied to (1.1a)–(1.1c) reads (see [2, p. 75] and [7, p. 71])

$$y_1 = y_0 + h\sum_{i=1}^{s} b_i Y_i', \qquad z_1 = z_0 + h\sum_{i=1}^{s} b_i Z_i', \qquad u_1 = u_0 + h\sum_{i=1}^{s} b_i U_i', \tag{2.1a}$$

where

$$Y_i' = f(Y_i, Z_i), \qquad Z_i' = k(Y_i, Z_i, U_i), \qquad 0 = g(Y_i), \tag{2.1b}$$

and the *internal stages* are given by

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij} Y_j', \qquad Z_i = z_0 + h\sum_{j=1}^{s} a_{ij} Z_j', \qquad U_i = u_0 + h\sum_{j=1}^{s} a_{ij} U_j'. \tag{2.1c}$$

We define the s-dimensional vectors $\mathbb{1} := (1, \ldots, 1)^T$ and $e_s := (0, \ldots, 0, 1)^T$. We denote $A := (a_{ij})_{i,j=1}^{s}$ the *RK matrix*, $b := (b_1, \ldots, b_s)^T$ the *weight vector*, $c := (c_1, \ldots, c_s)^T := A\mathbb{1}_s$ the *node vector*, and $C := \mathrm{diag}(c_1, \ldots, c_s)$. Let $B(p)$, $C(q)$, and $D(r)$ be the following *simplifying assumptions*, written in matrix notation, which are related to the construction of such methods (see [6, Section 34], [7, pp. 15–16], and [9, Section IV.5])

$$\begin{aligned}
B(p)&: \quad b^T C^{k-1}\mathbb{1} = \frac{1}{k}, && k = 1, \ldots, p;\\
C(q)&: \quad AC^{k-1}\mathbb{1} = \frac{1}{k}C^k\mathbb{1}, && k = 1, \ldots, q;\\
D(r)&: \quad b^T C^{k-1}A = \frac{1}{k}\left(b^T - b^T C^k\right), && k = 1, \ldots, r.
\end{aligned}$$

We are interested in RK methods satisfying the hypotheses

(I): the RK matrix A is invertible;

(S): the method is *stiffly accurate*, i.e., $a_{si} = b_i$ for $i = 1, \ldots, s$.

Remarks. The following results can be easily proved.

(1) (S) together with $B(1)$ leads to $c_s = 1$. Moreover, if $C(q)$ and $D(r)$ are satisfied then $B(p)$ holds with $p \geqslant \max(q, r+1)$.

(2) (S) implies that $y_1 = Y_s$, $g(y_1) = g(Y_s) = 0$, $z_1 = Z_s$, and $u_1 = U_s$ in (2.1).

(3) (I) and (S) imply that $R(\infty) = 0$ where R is the stability function of the RK method (see [9, Proposition IV.3.8]).

3. Existence, uniqueness, and influence of perturbations

This section is devoted to the analysis of the solution of the nonlinear system (2.1) with (y_0, z_0, u_0) replaced by approximate h-dependent starting values $(\eta, \zeta, \nu) = (\eta(h), \zeta(h), \nu(h))$. An important result is given by Theorem 3.4 which will be useful in Section 6 to the study of the error propagation. We first investigate the existence and uniqueness of the RK solution.

Theorem 3.1 ([7, Theorem 6.1], [12, Theorem IV.3.1]). *Let us suppose that*

$$g(\eta) = O(h^\tau), \quad \tau \geqslant 3, \tag{3.1a}$$

$$(g_y f)(\eta, \zeta) = O(h^\kappa), \quad \kappa \geqslant 2, \tag{3.1b}$$

$$(g_{yy}(f, f) + g_y f_y f + g_y f_z k)(\eta, \zeta, \nu) = O(h), \tag{3.1c}$$

$$(g_y f_z k_u)(y, z, u) \text{ is invertible in a neighbourhood of } (\eta, \zeta, \nu), \tag{3.1d}$$

and that $C(2)$ *and* (I) *are fulfilled. Then for* $h \leqslant h_0$ *there exists a locally unique solution to*

$$Y_i = \eta + h \sum_{j=1}^{s} a_{ij} f(Y_j, Z_j), \tag{3.2a}$$

$$Z_i = \zeta + h \sum_{j=1}^{s} a_{ij} k(Y_j, Z_j, U_j), \tag{3.2b}$$

$$0 = g(Y_i) \tag{3.2c}$$

for $i = 1, \ldots, s$, *which satisfies*

$$Y_i - \eta = O(h), \qquad Z_i - \zeta = O(h), \qquad U_i - \nu = O(h). \tag{3.3}$$

Remarks.

(1) If the function k of (1.1) is linear in u then the assumptions $C(2)$ and (3.1c) can be omitted. In this situation, $\tau \geqslant 2$ and $\kappa \geqslant 1$ are sufficient. However, if $C(2)$ is not satisfied, $\tau = 2$ or $\kappa = 1$, we only have the estimate $U_i - \nu = O(1)$ (for more details see Lemma 3.3 and [7, p. 74]).

(2) The value of ν in (3.1c) only prescribes the solution of (3.2) to be close to the manifold defined by (1.1e). However, (Y_i, Z_i, U_i) are clearly independent of ν.

(3) If the function k of (1.1) is not linear in u then $C(2)$ and (I) show the necessity of having $s \geqslant 2$.

The next result, a more complete and precise formulation of [7, Theorem 6.2], is concerned with the influence of perturbations to (3.2).

Theorem 3.2. *Let* (Y_i, Z_i, U_i) *be given by* (3.2) *and let us consider perturbed values* $(\widehat{Y}_i, \widehat{Z}_i, \widehat{U}_i)$ *satisfying*

$$\widehat{Y}_i = \widehat{\eta} + h \sum_{j=1}^{s} a_{ij} f(\widehat{Y}_j, \widehat{Z}_j) + h\delta_i, \tag{3.4a}$$

$$\widehat{Z}_i = \widehat{\zeta} + h \sum_{j=1}^{s} a_{ij} k(\widehat{Y}_j, \widehat{Z}_j, \widehat{U}_j) + h\mu_i, \tag{3.4b}$$

$$0 = g(\widehat{Y}_i) + \theta_i, \tag{3.4c}$$

for $i = 1, \ldots, s$. *In addition to the assumptions of Theorem* 3.1 *let us suppose that*

$$\begin{aligned} &\Delta\eta = O(h^3), \qquad \Delta\zeta = O(h^2), \qquad \widehat{U}_i - \nu = O(h), \\ &\delta_i = O(h^2), \qquad \mu_i = O(h), \qquad \theta_i = O(h^3). \end{aligned} \tag{3.5}$$

Then we have for $h \leqslant h_0$ the estimates

$$\Delta Y_i = P_y \Delta\eta + hc_i f_z P_z \Delta\zeta + \mathrm{O}\Big(h\|\Delta\eta\| + h^2\|\Delta\zeta\| + \frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|\Big), \tag{3.6a}$$

$$\Delta Z_i = -\frac{1}{h}\sigma_i \cdot SQ_y\Delta\eta + P_z\Delta\zeta + \mathrm{O}\Big(\|\Delta\eta\| + h\|\Delta\zeta\| + \frac{1}{h^3}\|Q_y\Delta\eta\|^2 + \frac{1}{h}\|Q_z\Delta\zeta\|^2 + \|\delta\| + h\|\mu\| + \frac{1}{h}\|\theta\|\Big), \tag{3.6b}$$

$$P_{z,i}\Delta Z_i = P_z\Delta\zeta + \mathrm{O}\Big(\|Q_y\Delta\eta\| + h\|P_y\Delta\eta\| + h\|\Delta\zeta\| + \frac{1}{h^3}\|Q_y\Delta\eta\|^2 + \frac{1}{h}\|Q_z\Delta\zeta\|^2 + h\|\delta\| + h\|\mu\| + \|\theta\|\Big), \tag{3.6c}$$

$$\Delta U_i = \mathrm{O}\Big(\frac{1}{h^2}\|Q_y\Delta\eta\| + \frac{1}{h}\|P_y\Delta\eta\| + \frac{1}{h}\|Q_z\Delta\zeta\| + \|P_z\Delta\zeta\| + \frac{1}{h}\|\delta\| + \|\mu\| + \frac{1}{h^2}\|\theta\|\Big) \tag{3.6d}$$

where $\sigma_i = e_i^{\mathrm{T}} A^{-1}\mathbb{1}$ with $e_i = (0,\ldots,0,1,0,\ldots,0)^{\mathrm{T}}$ (the s-dimensional vector with all components equal to 0 excepted the ith which is equal to 1), $\delta = (\delta_1,\ldots,\delta_s)^{\mathrm{T}}$, $\|\delta\| = \max_i \|\delta_i\|$, and similarly for μ and θ. P_y, Q_y, P_z, and Q_z are projectors *defined under the condition* (1.2) *by*

$$\begin{aligned} &S := k_u(g_y f_z k_u)^{-1} g_y, \\ &Q_y := f_z S, \qquad P_y := I - Q_y, \qquad Q_z := Sf_z, \qquad P_z := I - Q_z. \end{aligned} \tag{3.7}$$

Remarks.

(1) We have used the notation $\Delta\eta = \widehat{\eta} - \eta$, $\Delta\zeta = \widehat{\zeta} - \zeta$, $Y = (Y_1,\ldots,Y_s)^{\mathrm{T}}$, $\Delta Y = \widehat{Y} - Y$, $\|\Delta Y\| = \max_i \|\Delta Y_i\|$, and similarly for the z- and u-components.

(2) The missing arguments for f_z, S, P_y, Q_z, etc., are (η,ζ,ν) or $(\eta,\zeta,G(\eta,\zeta))$ with G as described in the Introduction. Those of $P_{z,i}$ are $(Y_i, Z_i, G(Y_i, Z_i))$ or (Y_i, Z_i, U_i).

(3) The conditions (3.5) ensure that all O-terms in the proof below are small.

(4) If $g(\widehat{\eta}) = 0 = g(\eta)$ then $Q_y\Delta\eta = \mathrm{O}(\|\Delta\eta\|^2)$. Consequently, this term may be neglected and the hypothesis $\Delta\eta = \mathrm{O}(h^3)$ can be relaxed to $\mathrm{O}(h^2)$. If we have $(g_y f)(\widehat{\eta},\widehat{\zeta}) = 0 = (g_y f)(\eta,\zeta)$ then similarly $\Delta\zeta = \mathrm{O}(h)$ suffices.

(5) If the function k of (1.1) is linear in u then the terms $\|Q_y\Delta\eta\|^2$ and $\|Q_z\Delta\zeta\|^2$ in (3.6a)–(3.6c) are multiplied by one additional factor h. In this case $\Delta\eta = \mathrm{O}(h^2)$, $\Delta\zeta = \mathrm{O}(h)$, $\widehat{U}_i - \nu = \mathrm{O}(1)$, $\delta_i = \mathrm{O}(h)$, $\mu_i = \mathrm{O}(1)$, and $\theta_i = \mathrm{O}(h^2)$ are sufficient, but then we only have the estimate $\Delta U_i = \mathrm{O}(1)$.

(6) The constants implied by the O-terms in (3.6) depend on bounds for certain derivatives of f, g, and k, but not on the constants entering in the O-terms in (3.1a), (3.1b), and (3.5), if h is sufficiently small.

(7) It can be observed that the terms $\|\theta\|/h^2$, $\|\theta\|/h$, $\|\delta\|/h$, $\|\delta\|$, and $\|\mu\|$ are not present in (3.6a) and (3.6c).

Proof of Theorem 3.2. Subtracting (3.2) from (3.4) we obtain by linearization

$$\Delta Y_i = \Delta\eta + h\sum_{j=1}^{s} a_{ij} f_y(Y_j, Z_j)\Delta Y_j + h\sum_{j=1}^{s} a_{ij} f_z(Y_j, Z_j)\Delta Z_j + h\delta_i + \mathrm{O}(h\|\Delta Y\|^2 + h\|\Delta Z\|^2), \tag{3.8a}$$

$$\Delta Z_i = \Delta\zeta + h\sum_{j=1}^{s} a_{ij} k_y(Y_j, Z_j, U_j)\Delta Y_j + h\sum_{j=1}^{s} a_{ij} k_z(Y_j, Z_j, U_j)\Delta Z_j + h\sum_{j=1}^{s} a_{ij} k_u(Y_j, Z_j, U_j)\Delta U_j + h\mu_i + \mathrm{O}(h\|\Delta Y\|^2 + h\|\Delta Z\|^2 + h\|\Delta U\|^2), \tag{3.8b}$$

$$0 = g_y(Y_i)\Delta Y_i + \theta_i + \mathrm{O}(\|\Delta Y_i\|^2), \tag{3.8c}$$

which can be rewritten, using tensor notation,

$$\Delta Y = \mathbb{1}\otimes\Delta\eta + h(A\otimes I)\{f_y\}\Delta Y + (A\otimes I)\{f_z\}h\Delta Z + h\delta + \mathrm{O}(h\|\Delta Y\|^2 + h\|\Delta Z\|^2), \tag{3.9a}$$

$$h\Delta Z = \mathbb{1}\otimes h\Delta\zeta + h^2(A\otimes I)\{k_y\}\Delta Y + h(A\otimes I)\{k_z\}h\Delta Z + (A\otimes I)\{k_u\}h^2\Delta U + h^2\mu + \mathrm{O}(h^2\|\Delta Y\|^2 + h^2\|\Delta Z\|^2 + h^2\|\Delta U\|^2), \tag{3.9b}$$

$$0 = \{g_y\}\Delta Y + \theta + \mathrm{O}(\|\Delta Y\|^2), \tag{3.9c}$$

where

$$\{g_y\} := \mathrm{blockdiag}\,(g_y(Y_1), \ldots, g_y(Y_s)), \tag{3.10a}$$

$$\{f_z\} := \mathrm{blockdiag}\,(f_z(Y_1, Z_1), \ldots, f_z(Y_s, Z_s)), \tag{3.10b}$$

$$\{k_u\} := \mathrm{blockdiag}\,(k_u(Y_1, Z_1, U_1), \ldots, k_u(Y_s, Z_s, U_s)), \tag{3.10c}$$

and similarly for $\{f_y\}$, $\{k_y\}$, and $\{k_z\}$. Inserting (3.9a) into (3.9c) and (3.9b) into the resulting formula leads to

$$\begin{aligned} -\{g_y\}(A\otimes I)\{f_z\}(A\otimes I)\{k_u\}h^2\Delta U &= \{g_y\}\Big[\mathbb{1}\otimes\Delta\eta + (A\otimes I)\{f_z\}(\mathbb{1}\otimes h\Delta\zeta) + h(A\otimes I)\{f_y\}\Delta Y \\ &\quad + h^2(A\otimes I)\{f_z\}(A\otimes I)\{k_y\}\Delta Y + h(A\otimes I)\{f_z\}(A\otimes I)\{k_z\}h\Delta Z \\ &\quad + (A\otimes I)\{f_z\}h^2\mu + h\delta\Big] + \theta + \mathrm{O}(\|\Delta Y\|^2 + h\|\Delta Z\|^2 + h^2\|\Delta U\|^2). \end{aligned} \tag{3.11}$$

In accordance with (3.3) we have

$$g_y(Y_i)\, a_{ij} f_z(Y_j, Z_j)\, a_{jk} k_u(Y_k, Z_k, U_k) = a_{ij}a_{jk}(g_y f_z k_u)(\eta, \zeta, \nu) + \mathrm{O}(h), \tag{3.12}$$

thus, the left matrix of (3.11) can be written as

$$\{g_y\}(A\otimes I)\{f_z\}(A\otimes I)\{k_u\} = A^2\otimes(g_y f_z k_u)(\eta,\zeta,\nu) + O(h) \tag{3.13}$$

and is invertible by (I) and (3.1d) for h sufficiently small. Putting

$$\begin{aligned} &G_y := \{g_y\}, && F_z := (A\otimes I)\{f_z\}(A\otimes I)^{-1},\\ &K_u := (A\otimes I)^2\{k_u\}(A\otimes I)^{-2}, && S_A := K_u(G_y F_z K_u)^{-1} G_y,\\ &Q_{y,A} := F_z S_A, && P_{y,A} := I - Q_{y,A},\\ &Q_{z,A} := S_A F_z, && P_{z,A} := I - Q_{z,A}, \end{aligned} \tag{3.14}$$

we remark for example that $G_y = G_y Q_{y,A}$, $S_A = S_A Q_{y,A}$, $G_y F_z = G_y F_z Q_{z,A}$, and $F_z P_{z,A} = P_{y,A} F_z$. Hence, from (3.11) and (3.9) we obtain

$$\begin{aligned} h^2\Delta U = &-(A\otimes I)^{-2}(G_y F_z K_u)^{-1} G_y\Big[\mathbb{1}\otimes\Delta\eta + F_z(A\mathbb{1}\otimes h\Delta\zeta) + h(A\otimes I)\{f_y\}\Delta Y\\ &+ h^2 F_z(A\otimes I)^2\{k_y\}\Delta Y + hF_z(A\otimes I)^2\{k_z\}h\Delta Z\Big]\\ &+ O(\|\Delta Y\|^2 + h\|\Delta Z\|^2 + h^2\|\Delta U\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|), \end{aligned} \tag{3.15a}$$

$$\begin{aligned} h\Delta Z = &(A\otimes I)^{-1}\Big[-S_A(\mathbb{1}\otimes\Delta\eta) + P_{z,A}(A\mathbb{1}\otimes h\Delta\zeta) - hS_A(A\otimes I)\{f_y\}\Delta Y\\ &+ h^2 P_{z,A}(A\otimes I)^2\{k_y\}\Delta Y + hP_{z,A}(A\otimes I)^2\{k_z\}h\Delta Z\Big]\\ &+ O(\|\Delta Y\|^2 + h\|\Delta Z\|^2 + h^2\|\Delta U\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|), \end{aligned} \tag{3.15b}$$

$$\begin{aligned} \Delta Y = &P_{y,A}(\mathbb{1}\otimes\Delta\eta) + F_z P_{z,A}(A\mathbb{1}\otimes h\Delta\zeta) + hP_{y,A}(A\otimes I)\{f_y\}\Delta Y\\ &+ h^2 P_{y,A}F_z(A\otimes I)^2\{k_y\}\Delta Y + hP_{y,A}F_z(A\otimes I)^2\{k_z\}h\Delta Z\\ &+ O(\|\Delta Y\|^2 + h\|\Delta Z\|^2 + h^2\|\Delta U\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|). \end{aligned} \tag{3.15c}$$

The O-terms in (3.15a)–(3.15c) lead to the estimate

$$O\Big(\frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|P_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|P_z\Delta\zeta\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|\Big).$$

If the function k of (1.1) is linear in u, they can be replaced by

$$O(\|\Delta Y\|^2 + h\|\Delta Z\|^2 + h^2\|\Delta Y\|\cdot\|\Delta U\| + h^2\|\Delta Z\|\cdot\|\Delta U\| + h\|\delta\| + h^2\|\mu\| + \|\theta\|)$$

yielding

$$O\Big(\frac{1}{h}\|Q_y\Delta\eta\|^2/h + \|P_y\Delta\eta\|^2 + h\|Q_z\Delta\zeta\|^2 + h\|P_z\Delta\zeta\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|\Big).$$

The term $P_y\Delta\eta$ entering in $P_{z,i}\Delta Z_i$ leads to the estimate $O(h\|P_y\Delta\eta\|)$, because of $(I\otimes P_{z,i})S_A(\mathbb{1}\otimes P_y) = O(h^2)$ which is a consequence of $P_z k_u \equiv 0$ and $g_y P_y \equiv 0$. The estimate (3.6c) for the perturbations δ and θ simply follows from (3.6a, b, d), (3.8b), and $P_z k_u \equiv 0$. □

Our next aim in Theorem 3.4 is to show that the estimates (3.6a) and (3.6c) can be improved for $i = s$. The following lemma will be useful in the proof of this theorem.

Lemma 3.3. *Besides the hypotheses of Theorem* 3.1, *let us suppose further that* $C(q)$ *holds. Then the solution* (Y_i, Z_i, U_i) *of* (3.2) *satisfies*

$$Y_i = \widetilde{\eta} + \sum_{m=1}^{\lambda} \frac{c_i^m h^m}{m!} D_m Y(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu}) + O(h^{\lambda+1}), \tag{3.16a}$$

$$Z_i = \widetilde{\zeta} + \sum_{n=1}^{\gamma} \frac{c_i^n h^n}{n!} D_n Z(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu}) + O(h^{\gamma+1}), \tag{3.16b}$$

$$U_i = \widetilde{\nu} + \sum_{p=1}^{\mu} \frac{c_i^p h^p}{p!} D_p U(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu}) + O(h^{\mu+1}), \tag{3.16c}$$

where $\lambda = \min(\tau, \kappa+1, q)$, $\gamma = \min(\tau-2, \kappa, q-1)$, $\mu = \min(\tau-3, \kappa-2, q-2)$, *and* $D_m Y$, $D_n Z$, $D_p U$ *are functions composed with derivatives of* f, g, *and* k. $(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu})$ *are consistent values close to* (η, ζ, ν) *but constructed independently of* ν. *They are uniquely determined by* (1.1c)–(1.1e), $P_y(\eta, \zeta, \nu^*)(\widetilde{\eta} - \eta) = 0$, *and* $P_z(\eta, \zeta, \nu^*)(\widetilde{\zeta} - \zeta) = 0$ *with* $\nu^* := G(\eta, \zeta)$.

Proof. We find $Q_y(\eta, \zeta, \nu^*)(\widetilde{\eta} - \eta) = O(h^\tau)$ and $Q_z(\eta, \zeta, \nu^*)(\widetilde{\zeta} - \zeta) = O(h^{\min(\tau,\kappa)})$. We define $(y(x), z(x), u(x))$ the solution of (1.1) with initial values $y(x_0) = \widetilde{\eta}$, $z(x_0) = \widetilde{\zeta}$, and $u(x_0) = \widetilde{\nu}$. The exact solution at $x_0 + c_i h$ satisfies (3.4) with $\theta_i = 0$ and

$$\delta_i = \frac{h^q}{q!} y^{(q+1)}(x_0) \left(\frac{c_i^{q+1}}{q+1} - \sum_{j=1}^{s} a_{ij} c_j^q \right) + O(h^{q+1}) = O(h^q), \tag{3.17a}$$

$$\mu_i = \frac{h^q}{q!} z^{(q+1)}(x_0) \left(\frac{c_i^{q+1}}{q+1} - \sum_{j=1}^{s} a_{ij} c_j^q \right) + O(h^{q+1}) = O(h^q). \tag{3.17b}$$

The difference from the numerical solution (3.2) can thus be estimated with Theorem 3.2, yielding

$$\|Y_i - y(x_0 + c_i h)\| = O(h^{\min(\tau+1, \kappa+2, q+1)}), \tag{3.18a}$$

$$\|Z_i - z(x_0 + c_i h)\| = O(h^{\min(\tau-1, \kappa+1, q)}), \tag{3.18b}$$

$$\|U_i - u(x_0 + c_i h)\| = O(h^{\min(\tau-2, \kappa-1, q-1)}). \quad \square \tag{3.18c}$$

Here is the main result of this section. New improved estimates for the perturbed solution are given. They will be essential in the proof of Theorem 6.1.

Theorem 3.4. *In addition to the assumptions of Theorem* 3.2 *and Lemma* 3.3, *including those of Theorem* 3.1, *let us suppose that* $B(1)$, $D(r)$, *and* (S) *hold. Then we have*

$$\Delta Y_s = P_y \Delta\eta + h f_z P_z \Delta\zeta + O\Big(h\|\Delta\eta\| + h^2 \|P_z \Delta\zeta\| + h^{m+2} \|Q_z \Delta\zeta\| + \frac{1}{h^2} \|Q_y \Delta\eta\|^2 + \|Q_z \Delta\zeta\|^2 + h\|\delta\| + h^2 \|\mu\| + \|\theta\| \Big), \tag{3.19a}$$

$$hP_{z,s}\Delta Z_s = hP_z\Delta\zeta + \mathrm{O}\Big(h\|Q_y\Delta\eta\| + h^2\|P_y\Delta\eta\| + h^2\|P_z\Delta\zeta\| + h^{n+2}\|Q_z\Delta\zeta\| + \frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h^2\|\delta\| + h^2\|\mu\| + h\|\theta\|\Big), \tag{3.19b}$$

$$hQ_{z,s}\Delta Z_s = -\sigma\cdot SQ_y\Delta\eta + \mathrm{O}\Big(h\|\Delta\eta\| + h^2\|\Delta\zeta\| + \frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|\Big), \tag{3.19c}$$

where $\sigma = b^{\mathrm{T}}A^{-2}\mathbb{1}$, $m = \min(\tau-3, \kappa-2, q-2, \max(r-1,0))$, $n = \min(\tau-3, \kappa-2, q-2, r)$. P_y, Q_y, P_z, Q_z, S, *and* f_z *are evaluated at* (η, ζ, ν^*) *with* ν^* *defined as in Lemma* 3.3. *The arguments of* $P_{z,s}$ *and* $Q_{z,s}$ *are* $(Y_s, Z_s, G(Y_s, Z_s))$ *or* (Y_s, Z_s, U_s).

Remarks.

(1) If the function k of (1.1) is linear in u, then $m = \min(\tau-2, \kappa-1, q-1, \max(r-1,0))$, $n = \min(\tau-2, \kappa-1, q-1, r)$. The fifth remark of Theorem 3.2 also holds here.

(2) The important results consist in the splitting of $\Delta\zeta$ according to the projections P_z and Q_z, and in the h-exponents in front of $\|Q_z\Delta\zeta\|$ in (3.19a) and (3.19b).

(3) It must be stressed that m and n satisfy $0 \leqslant m \leqslant n \leqslant m+1$.

(4) In the proof the missing arguments for f_z, P_y, P_z, etc., are $(\tilde\eta, \tilde\zeta, \tilde\nu)$ defined in Lemma 3.3. At the end of the proof a final estimate shows that they can be replaced by (η, ζ, ν^*).

(5) The sixth remark of Theorem 3.2 is also valid here for almost all constants entering in the O-expressions of (3.19). The exceptions are the constants implied by the O-terms $\mathrm{O}(h^{m+2}\|Q_z\Delta\zeta\|)$ and $\mathrm{O}(h^{n+2}\|Q_z\Delta\zeta\|)$ in (3.19a) and (3.19b) which depend on those of (3.1a) and (3.1b) if m or $n \geqslant 1$. Nevertheless, this will not affect the proof of Theorem 6.1 (see Section 6) where Theorem 3.4 will be applied.

Proof of Theorem 3.4. We resume the proof of Theorem 3.2 with the help of Lemma 3.3, using the same notations and definitions. Because of

$$\begin{aligned} \Delta Y_s &= (e_s^{\mathrm{T}}\otimes I)\Delta Y, \\ P_{z,s}\Delta Z_s &= (e_s^{\mathrm{T}}\otimes P_{z,s})\Delta Z, \\ Q_{z,s}\Delta Z_s &= (e_s^{\mathrm{T}}\otimes Q_{z,s})\Delta Z, \end{aligned} \tag{3.20}$$

Eq. (3.19) is a simple consequence of (3.6a)–(3.6c) with the exception of the h-exponents in front of $\|Q_z\Delta\zeta\|$ in (3.19a) and (3.19b) which remain to be shown. They will be computed with similar techniques used in the proof of [10, Theorem 4.4].

The formulas (3.15b) and (3.15c) can be rewritten

$$\begin{pmatrix} I - hS - h^2T \end{pmatrix}\begin{pmatrix} \Delta Y \\ h\Delta Z \end{pmatrix} = V + \mathrm{O}\Big(\frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|P_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|P_z\Delta\zeta\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|\Big), \tag{3.21}$$

where the matrices S, T, and the vector V are given by

$$S = \begin{pmatrix} P_{y,A}(A\otimes I)\{f_y\} & P_{y,A}F_z(A\otimes I)^2\{k_z\} \\ -(A\otimes I)^{-1}S_A(A\otimes I)\{f_y\} & (A\otimes I)^{-1}P_{z,A}(A\otimes I)^2\{k_z\} \end{pmatrix}, \tag{3.22a}$$

$$T = \begin{pmatrix} P_{y,A}F_z(A\otimes I)^2\{k_y\} & O \\ (A\otimes I)^{-1}P_{z,A}(A\otimes I)^2\{k_y\} & O \end{pmatrix}, \tag{3.22b}$$

$$V = \begin{pmatrix} P_{y,A}(\mathbb{1}\otimes\Delta\eta) + F_zP_{z,A}(A\mathbb{1}\otimes h\Delta\zeta) \\ (A\otimes I)^{-1}[-S_A(\mathbb{1}\otimes\Delta\eta) + P_{z,A}(A\mathbb{1}\otimes h\Delta\zeta)] \end{pmatrix}. \tag{3.22c}$$

We put $K_{u,0} := I\otimes k_u(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu})$ corresponding to K_u with (Y_i, Z_i, U_i) replaced by $(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu})$. By the use of

$$\Delta\zeta = P_z\Delta\zeta + Q_z\Delta\zeta, \qquad Q_y = SQ_y, \qquad Q_z^2 = Q_z, \qquad P_{z,A}K_u = 0,$$

V can be estimated by

$$V = W + \begin{pmatrix} \mathbb{1}\otimes P_y\Delta\eta + A\mathbb{1}\otimes hf_zP_z\Delta\zeta + O(h\|\Delta\eta\| + h^2\|P_z\Delta\zeta\|) \\ -A^{-1}\mathbb{1}\otimes Q_y\Delta\eta + \mathbb{1}\otimes hP_z\Delta\zeta + O(h\|\Delta\eta\| + h^2\|P_z\Delta\zeta\|) \end{pmatrix}, \tag{3.23a}$$

where we have isolated in W the terms including $Q_z\Delta\zeta$

$$W = \begin{pmatrix} -F_zP_{z,A}(K_u - K_{u,0})(A\mathbb{1}\otimes h(g_yf_zk_u)^{-1}g_yf_zQ_z\Delta\zeta) \\ -(A\otimes I)^{-1}P_{z,A}(K_u - K_{u,0})(A\mathbb{1}\otimes h(g_yf_zk_u)^{-1}g_yf_zQ_z\Delta\zeta) \end{pmatrix}. \tag{3.23b}$$

Computing the inverse of the left matrix in (3.21) by means of the series of Von Neumann we arrive at

$$\begin{aligned} \begin{pmatrix} \Delta Y \\ h\Delta Z \end{pmatrix} &= \sum_{\rho=0}^{n} (hS + h^2T)^\rho W \\ &\quad + \begin{pmatrix} \mathbb{1}\otimes P_y\Delta\eta + A\mathbb{1}\otimes hf_zP_z\Delta\zeta + O(h\|\Delta\eta\| + h^2\|P_z\Delta\zeta\| + h^{n+2}\|Q_z\Delta\zeta\|) \\ -A^{-1}\mathbb{1}\otimes Q_y\Delta\eta + \mathbb{1}\otimes hP_z\Delta\zeta + O(h\|\Delta\eta\| + h^2\|P_z\Delta\zeta\| + h^{n+2}\|Q_z\Delta\zeta\|) \end{pmatrix} \\ &\quad + O\Big(\frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|P_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|P_z\Delta\zeta\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|\Big). \end{aligned} \tag{3.24}$$

Our next aim is to develop at the point $(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu})$ the expression involving S, T, and W in (3.24) into h-powers. By the use of Lemma 3.3, the expressions G_y, F_z, K_u, $\{f_y\}$, $\{k_y\}$, and $\{k_z\}$ can be expanded, e.g.,

$$\begin{aligned} F_z &= \sum_{k=0}^{\gamma} h^kAC^kA^{-1}\otimes B_k + O(h^{\gamma+1}) \\ &= I\otimes f_z + hACA^{-1}\otimes\big(f_{yz}(f,\cdot) + f_{zz}(k,\cdot)\big) + \cdots, \end{aligned} \tag{3.25}$$

where the B_k are functions compound with derivatives of f, g, and k and evaluated at $(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu})$. In order to develop $(G_yF_zK_u)^{-1}$ we first consider

$$G_y F_z K_u = I \otimes (g_y f_z k_u) + \sum_{0<i+j+k\leqslant\omega} h^{i+j+k}(C^i A C^j A^{-1} A^2 C^k A^{-2}) \otimes D_{ijk} + \mathrm{O}(h^{\omega+1}) \tag{3.26}$$

where $\omega = \mu$ ($\omega = \gamma$ if k is linear in u because $k_u(y,z,u)$ is independent of u) and the D_{ijk} are other expressions like the B_k evaluated at $(\tilde\eta, \tilde\zeta, \tilde\nu)$. By using once again the series of Von Neumann we obtain

$$\begin{aligned}(G_y F_z K_u)^{-1} &= I \otimes (g_y f_z k_u)^{-1} \\ &\quad + \sum_{0<|\alpha|+|\beta|+|\delta|\leqslant\omega} h^{|\alpha|+|\beta|+|\delta|} \left(\prod_{i=1}^{\omega} C^{\alpha_i} A C^{\beta_i} A^{-1} A^2 C^{\delta_i} A^{-2} \right) \otimes E_{\alpha\beta\delta} + \mathrm{O}(h^{\omega+1}),\end{aligned} \tag{3.27}$$

where the $E_{\alpha\beta\delta}$ are other expressions like the B_k evaluated at $(\tilde\eta, \tilde\zeta, \tilde\nu)$, $\alpha = (\alpha_1, \ldots, \alpha_\omega)$, $\beta = (\beta_1, \ldots, \beta_\omega)$, and $\delta = (\delta_1, \ldots, \delta_\omega)$ are multi-indices in $\mathbb{N}^\omega$. The norm of a multi-index $\kappa = (\kappa_1, \ldots, \kappa_\omega) \in \mathbb{N}^\omega$ is defined by $|\kappa| := \sum_{i=1}^{\omega} \kappa_i$. Hence, we are now able to develop the sum containing $Q_z \Delta\zeta$ in (3.24).

In order to show (3.19a), we carefully examine only two representative terms, as the others can be treated similarly. We first consider

$$H := -(e_s^{\mathrm{T}} \otimes I) F_z P_{z,A} (K_u - K_{u,0}) (A\mathbb{1} \otimes h(g_y f_z k_u)^{-1} g_y f_z Q_z \Delta\zeta), \tag{3.28}$$

which is simply a constituent of W and it appears in ΔY_s when $\rho = 0$ in the sum of (3.24). By expanding H into h-powers we arrive at

$$H = h \sum_{\substack{1\leqslant|\sigma|\leqslant m \\ \sigma:=(\alpha,\beta,\delta,\nu)}} h^{|\sigma|} C_\sigma \cdot K_\sigma Q_z \Delta\zeta + \mathrm{O}(h^{m+2} \|Q_z \Delta\zeta\|), \tag{3.29}$$

where α, β, δ, and ν are multi-indices, the K_σ are of the same type as the D_{ijk}, and the coefficients C_σ are given by

$$C_\sigma = e_s^{\mathrm{T}} A C^{\nu_1} A^{-1} A^2 C^{\nu_2} A^{-2} \left(\prod_{i=1}^{\omega} C^{\alpha_i} A C^{\beta_i} A^{-1} A^2 C^{\delta_i} A^{-2} \right) C^{\nu_3} A C^{\nu_4} A^{-1} \underbrace{A^2 C^{\nu_5} A^{-2} A\mathbb{1}}_{A \cdot A C^{\nu_5} A^{-1} \mathbb{1}} \tag{3.30}$$

with $\nu_5 > 0$. If $D(r)$ is satisfied with $r \geqslant 1$, then in accordance with Theorem 4.1 below, these coefficients vanish for $|\sigma| + 1 = |\alpha| + |\beta| + |\delta| + |\nu| + 1 \leqslant r$. For $r = 0$, we have $H = \mathrm{O}(h^2 \|Q_z \Delta\zeta\|)$ as a consequence of $K_u - K_{u,0} = \mathrm{O}(h)$. We thus get $H = \mathrm{O}(h^{m+2} \|Q_z \Delta\zeta\|)$.

Assuming that $m \geqslant 3$, we now consider a second expression involving $Q_z \Delta\zeta$ which enters in the computation of ΔY_s, coming from $h^2 S^2 W$ in (3.24),

$$\begin{aligned}J := &-h^2 (e_s^{\mathrm{T}} \otimes I) F_z P_{z,A} (A \otimes I)^2 \{k_z\} (A \otimes I)^{-1} P_{z,A} (A \otimes I)^2 \{k_z\} \\ &\times (A \otimes I)^{-1} P_{z,A} (K_u - K_{u,0}) (A\mathbb{1} \otimes h(g_y f_z k_u)^{-1} g_y f_z Q_z \Delta\zeta).\end{aligned} \tag{3.31}$$

As seen above we get

$$J = h^3 \sum_{\substack{1\leqslant|\sigma|\leqslant m-2 \\ \sigma:=(\alpha,\beta,\delta,\varepsilon,\theta,\kappa,\lambda,\varsigma,\upsilon,\nu)}} h^{|\sigma|} D_\sigma \cdot L_\sigma Q_z \Delta\zeta + \mathrm{O}(h^{m+2} \|Q_z \Delta\zeta\|), \tag{3.32}$$

where the L_σ are other expressions like the D_{ijk} and evaluated at $(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu})$. The coefficients D_σ are given by

$$\begin{aligned} D_\sigma = {} & e_s^{\mathrm{T}} A C^{\nu_1} A^{-1} \\ & \times A^2 C^{\nu_2} A^{-2} \left(\prod_{i=1}^{\omega} C^{\alpha_i} A C^{\beta_i} A^{-1} A^2 C^{\delta_i} A^{-2} \right) C^{\nu_3} A C^{\nu_4} A^{-1} \underbrace{A^2 C^{\nu_5} A^{-1}}_{A \cdot A C^{\nu_5} A^{-1}} \\ & \times A^2 C^{\nu_6} A^{-2} \left(\prod_{i=1}^{\omega} C^{\varepsilon_i} A C^{\theta_i} A^{-1} A^2 C^{\kappa_i} A^{-2} \right) C^{\nu_7} A C^{\nu_8} A^{-1} \underbrace{A^2 C^{\nu_9} A^{-1}}_{A \cdot A C^{\nu_9} A^{-1}} \\ & \times A^2 C^{\nu_{10}} A^{-2} \left(\prod_{i=1}^{\omega} C^{\lambda_i} A C^{\varsigma_i} A^{-1} A^2 C^{\upsilon_i} A^{-2} \right) C^{\nu_{11}} A C^{\nu_{12}} A^{-1} \underbrace{A^2 C^{\nu_{13}} A^{-2} A \mathbb{1}}_{A \cdot A C^{\nu_{13}} A^{-1} \mathbb{1}} \end{aligned} \tag{3.33}$$

with $\nu_{13} > 0$, and according to Theorem 4.1 they vanish if $|\sigma| + 3 \leqslant r$. Therefore, J can be estimated by $\mathrm{O}(h^{m+2} \| Q_z \Delta\zeta \|)$ too. All other remaining terms can be treated in a similar way, so that (3.19a) results.

The h-exponent in front of $Q_z \Delta\zeta$ in (3.19b) remains to be proved. We use the same techniques as above to estimate ΔY_s. The main difference is that we expand some specific terms in the matrix products involving S and T in (3.24) into h-powers not at the point $(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu})$ but at $(Y_s, Z_s, G(Y_s, Z_s))$. Such an expansion concerns only the first factors of these matrix products which are equal to K_u. From Lemma 3.3 we easily get

$$Y_i = Y_s - \sum_{m=1}^{\lambda} \frac{(1 - c_i^m) h^m}{m!} D_m Y(Y_s, Z_s, G(Y_s, Z_s)) + \mathrm{O}(h^{\lambda+1}), \tag{3.34a}$$

$$Z_i = Z_s - \sum_{n=1}^{\gamma} \frac{(1 - c_i^n) h^n}{n!} D_n Z(Y_s, Z_s, G(Y_s, Z_s)) + \mathrm{O}(h^{\gamma+1}), \tag{3.34b}$$

$$U_i = G(Y_s, Z_s) - \sum_{p=1}^{\mu} \frac{(1 - c_i^p) h^p}{p!} D_p U(Y_s, Z_s, G(Y_s, Z_s)) + \mathrm{O}(h^{\mu+1}). \tag{3.34c}$$

We rewrite an expression entering in the vector W (see (3.23b)) in the form

$$P_{z,A}(K_u - K_{u,0})(A \otimes I) = P_{z,A}(A \otimes I)(A \otimes I)^{-1}(K_u - K_{u,0})(A \otimes I). \tag{3.35}$$

The above expression $(A \otimes I)^{-1}(K_u - K_{u,0})(A \otimes I)$ can be expanded leading to

$$(A \otimes I)^{-1}(K_u - K_{u,0})(A \otimes I) = \sum_{k=1}^{\omega} h^k A C^k A^{-1} \otimes K_{k,s}, \tag{3.36}$$

where the $K_{k,s}$ are other expressions like the B_k. To end the proof, the arguments are similar to those used when estimating ΔY_s. The only problem could arise from coefficients of the form $e_s^{\mathrm{T}} A^{-1}(I - C^k) \cdots \mathbb{1}_s$ or $e_s^{\mathrm{T}} A^{-1} C^k \cdots \mathbb{1}_s$ with $k \geqslant 1$. But these coefficients do not appear, because of the premultiplication with $P_{z,s}$ and the fact that $P_z(Y_s, Z_s, G(Y_s, Z_s)) K_{u,s} \equiv 0$ with $K_{u,s} :=$

$K_u(Y_s, Z_s, G(Y_s, Z_s))$. Without such a premultiplication these coefficients appear in ΔZ_s coming from, e.g., the expansion of

$$(e_s^T \otimes I)(A^{-1} \otimes I)(I \otimes K_{u,s})(G_y F_z K_u)^{-1} \cdots (\mathbb{1} \otimes h(g_y f_z k_u)^{-1} g_y f_z Q_z \Delta\zeta). \quad \square \tag{3.37}$$

4. Properties of the Runge–Kutta coefficients

The following theorem deals with the valuation of expressions encountered in the demonstration of Theorem 3.4.

Theorem 4.1. *Let us assume that the conditions $B(1)$, $D(r)$, (I), and (S) hold. For a fixed $\rho \in \mathbb{N}\backslash\{0\}$, let us consider a multi-index $\nu = (\nu_1, \ldots, \nu_\rho)$ satisfying $\nu_i \geqslant 1$ and let $\alpha \geqslant 0$. If $|\nu| \leqslant r$, then we have*

$$e_s^T C^\alpha \left(\prod_{i=1}^{\rho} M_i \right) \mathbb{1} = e_s^T C^\alpha M_1 \cdots M_\rho \mathbb{1} = 0, \tag{4.1}$$

where the matrices M_i are of the form

$$A^{\nu_i}, \qquad A^{\sigma_i} C^{\nu_i} A^{-\sigma_i}, \qquad A^{\sigma_i}(I - C^{\nu_i}) A^{-\sigma_i - 1} \tag{4.2}$$

with $\sigma_i \in \{0, 1, 2\}$ and it is supposed that $M_\rho = AC^{\nu_\rho} A^{-1}$.

Remark. In the proof we adopt the convention that if a null factor multiplies a term of the form $b^T C^{-m}$ with $m \geqslant 1$, then this expression has to be omitted. For example (4.3) with $k = 0$ reads $b^T A^{-1} = e_s^T - 0 \cdot b^T C^{-1} = e_s^T$.

Proof of Theorem 4.1. In this proof k and l denote two non-negative integers. Assuming that $k \leqslant r$, the multiplication of (S) and $D(r)$ with A^{-1} leads to

$$b^T C^k A^{-1} = e_s^T - k b^T C^{k-1}. \tag{4.3}$$

(S), $D(r)$, and (4.3) together imply that

$$b^T C^k A C^l A^{-2} = e_s^T + \frac{l(l-1)}{k+1} b^T C^{l-2} - \frac{(k+l+1)(k+l)}{k+1} b^T C^{k+l-1} \tag{4.4}$$

provided $k + l \leqslant r - 1$. Similarly, if $k + l \leqslant r$, we get

$$b^T C^k (I - C^l) A^{-2} = l e_s^T + k(k-1) b^T C^{k-2} - (k+l)(k+l-1) b^T C^{k+l-2} \tag{4.5}$$

and for $k + l \leqslant r - 1$ we have

$$\begin{aligned} & b^T C^k A (I - C^l) A^{-3} \\ & \quad = 2l e_s^T + \frac{l(l-1)(l-2)}{k+1} b^T C^{l-3} \\ & \qquad + k(k-1) b^T C^{k-2} - \frac{(k+l+1)(k+l)(k+l-1)}{k+1} b^T C^{k+l-2}. \end{aligned} \tag{4.6}$$

A repeated application of (S), $D(r)$, (4.3)–(4.6) to (4.1) shows that this expression is a linear combination of terms $b^{\mathrm{T}}C^kA^{-1}\mathbb{1}$ with $1 \leqslant k \leqslant r$. They all vanish because of

$$b^{\mathrm{T}}C^kA^{-1}\mathbb{1} = e_s^{\mathrm{T}}\mathbb{1} - kb^{\mathrm{T}}C^{k-1}\mathbb{1} = 1 - k\frac{1}{k} = 0, \tag{4.7}$$

which is a consequence of (4.3) and $B(r)$ (the first remark of Section 2 applies). □

5. The local error

We consider one step of a RK method (2.1) with consistent initial values (y_0, z_0, u_0) at x_0 and we want to give estimates for the *local error*

$$\begin{aligned} \delta y_h(x_0) &= y_1 - y(x_0 + h), \\ \delta z_h(x_0) &= z_1 - z(x_0 + h), \\ \delta u_h(x_0) &= u_1 - u(x_0 + h). \end{aligned} \tag{5.1}$$

Optimal local error estimates for the y-component and a certain projection of the z-component are given in the following theorem. For the z-component this requires the use of the new technique of DA3-series. Because of the considerable mathematical material related to DA3-series, we only give a sketch of the proof and we refer the reader to [12] for more details.

Theorem 5.1.

(a) *Let us suppose that the RK method satisfies* (I), $B(p)$, *and* $C(q)$ *with* $q \geqslant 2$. *Then we have*

$$\delta y_h(x_0) = \mathrm{O}(h^{\min(p,q)+1}), \qquad P_y(x_0 + h)\delta y_h(x_0) = \mathrm{O}(h^{\min(p,q+1)+1}), \tag{5.2a}$$

$$\delta z_h(x_0) = \mathrm{O}(h^q), \qquad P_z(x_0 + h)\delta z_h(x_0) = \mathrm{O}(h^{\min(p,q)+1}), \tag{5.2b}$$

$$\delta u_h(x_0) = \mathrm{O}(h^{\min(p,q-2)+1}) \tag{5.2c}$$

where $P_y(x)$ *and* $P_z(x)$ *are the projectors* (3.7) *evaluated at* $(y(x), z(x), u(x))$, *the exact solution of* (1.1) *at* x.

(b) *Moreover, if in addition* $D(r)$ *and* (S) *hold, then we obtain*

$$\delta y_h(x_0) = \mathrm{O}(h^{k+1}), \quad \textit{with } k = \min(p, 2q - 1, q + r), \tag{5.3}$$

$$P_z(x_0 + h)\delta z_h(x_0) = \mathrm{O}(h^{\ell+1}), \quad \textit{with } \ell = \min(p, 2q - 2, q + r). \tag{5.4}$$

Remarks.

(1) If the function k of (1.1) is linear in u, then, in (5.3) and (5.4) we have

$$k = \min(p, 2q, q + r), \tag{5.3′}$$

$$\ell = \min(p, 2q - 1, q + r), \tag{5.4′}$$

and the condition $q \geqslant 2$ can be omitted.

(2) Concerning the demonstration of part (b), a detailed proof is given in [12]. In this article the general "rooted-tree type" theory about the Taylor expansion of the exact and of the numerical solutions is not presented. We do not give the complete definitions and the obvious results, which are necessary here, of the extension to index-3 problems of the theory of Butcher (see [6, Section 30] and [8, Section II.2]) for first-order ordinary differential equations. For systems of index 1 and 2 this has been realized (see [15], [7, Section 5], and [9, Section VI.8]). The entire derivation can be obtained in total analogy by following, for example, the last quoted reference which is devoted to index-2 problems. The set of resulting trees is denoted by DAT3. The labelling in LDAT3, the corresponding set of monotonically labelled trees, must be defined conscientiously. The concept of B-series (see [8, Section II.12]) can also be extended, leading to the so-called "DA3-series", and the composition law of DA3-series can be formulated.

Proof of Theorem 5.1 (*Outline*). Part (a) is already known (see [7, Lemma 6.3]).

Part (b) remains to be demonstrated. The local error of the y-component can be found by repeated application of simplifying assumptions to the order conditions, such as shown in the proof of [7, Theorem 5.9] (summarized in [9, Theorem VI.8.10]).

However, the estimate (5.4) is much more difficult to prove. The first idea is to develop the local error of the z-component, not at x_0, but at $x_0 + h$. Using the shortened notation $\Psi(x) := (y(x), z(x), u(x))$ for the exact solution of (1.1) at x, we get the $\mathrm{DA3}_z$-series

$$\begin{aligned} z_1 - z(x_0+h) &= \mathrm{DA3}_z(\boldsymbol{a}, \Psi(x_0+h)) - z(x_0+h) \\ &= \sum_{v \in \mathrm{DAT3}_z} \alpha(v) \frac{h^{\rho(v)}}{\rho(v)!} \boldsymbol{a}(v) F(v)\big(\Psi(x_0+h)\big), \end{aligned} \tag{5.5}$$

where for a tree w, $F(w)$ is the corresponding elementary differential, $\alpha(w)$ indicates the number of possible monotonic labellings of w, and $\rho(w)$ denotes its order. The expression of $\boldsymbol{a}(v)$ depending on the coefficients of the method, can be found by the use of the DA3-series composition law, leading to

$$\boldsymbol{a}(v) = \frac{(-1)^{\rho(v)}}{\alpha(v)} \sum_{\text{labellings of } v} \left(\sum_{j=0}^{\rho(v)} (-1)^j \binom{\rho(v)}{j} \big(\gamma(s_j(v))\Phi(s_j(v)) - 1\big) \right), \tag{5.6}$$

where the first summation is over all the different $\alpha(v)$ labellings of v and $s_i(v)$ is the "subtree" formed by the i first indices of a particular labelling of v. For a tree w, it can be noticed that $\Phi(w)$, which is function of the RK coefficients, and the rational number $\gamma(w)$, are in fact independent of the labelling.

The main idea is to show that for all trees $v \in \mathrm{DAT3}_z$ which are not of the form $[u]_z$ with $u \in \mathrm{DAT3}_u$, i.e., for which $F(v) \not\equiv k_u F(u)$, we have $\boldsymbol{a}(v) = 0$ if $\rho(v) \leqslant \ell$ (with ℓ given by (5.4)). This will give the desired result since the remaining trees are of higher order or satisfy $v = [u]_z$ with $u \in \mathrm{DAT3}_u$ implying that $P_z F(v) \equiv P_z k_u F(u) \equiv 0$.

Now we consider a tree $v \in \mathrm{DAT3}_z$, $v \neq [u]_z$ with $u \in \mathrm{DAT3}_u$, satisfying $\rho(v) \leqslant \ell$. We can suppose that $\rho(v) \geqslant \min(p, q) + 1$, because we already know by (5.2b) that $\boldsymbol{a}(v) = 0$ for all trees of order $\rho(v) \leqslant \min(p, q)$. We first recursively simplify the terms $\gamma(s_i(v))\Phi(s_i(v)) - 1$ in (5.6)

with the help of $C(q)$. The order conditions reduced then by $D(r)$ to those of the bushy trees can also be eliminated. But a linear combination of various terms of the form

$$\sum_{l=0}^{\rho(v)-m} \mu_v(v_l)(-1)^{\rho(v_l)} \binom{\rho(v)}{\rho(v_l)} \Big(\gamma(v_l)\Phi(v_l) - 1\Big) \tag{5.7}$$

remains, where $m = \rho(v_0) = \rho(u) + 1$, $\Phi(v_l) = \sum_{i=1}^{s} b_i c_i^l \Phi_i(u)$ with $u \in \mathrm{DAT3}_u$ satisfying $\rho(u) \geqslant q - 1$, $\rho(v_l) = m + l$, and $\gamma(v_l) = \gamma(v_0)(m + l)/m$. In general several trees u exist which are not reducible by $C(q)$. The coefficients $\mu_v(v_l)$ count the number of times that its multiplicand in (5.7) appears in the sum (5.6) after reduction by $C(q)$. For u fixed we have the relations

$$\mu_v(v_l) = \binom{m+l-1}{l} \cdot \mu_v(v_0)$$

which are related to the number of labellings of v. Defining

$$\theta_l := (-1)^l \binom{\rho(v)-m}{l} \cdot \mu_v(v_0),$$

the sum (5.7) can be rewritten

$$\begin{aligned} &\sum_{l=1}^{\rho(v)-m} \Big[\mu_v(v_l)(-1)^{m+l} \binom{\rho(v)}{m+l} \Big(\gamma(v_l)\Phi(v_l) - 1\Big) - \theta_l(-1)^m \binom{\rho(v)}{m} \Big(\gamma(v_0)\Phi(v_0) - 1\Big)\Big] \\ &= \mu_v(v_0) \sum_{l=1}^{\rho(v)-m} (-1)^{m+l} \frac{\rho(v)!}{l!\,(\rho(v)-m-l)!\,m!} \Big(\gamma(v_0)\big(\Phi(v_l) - \Phi(v_0)\big) + \frac{l}{m+l}\Big), \end{aligned} \tag{5.8}$$

and the last expression in parentheses can be proved to vanish, first with the help of $D(r)$, then as in the proof related to the local error of the y-component. □

6. Convergence results

We present here the main result of this article and we partly follow [7, pp. 78–82]. Theorem 6.1 proves the conjecture, based on numerical experiments, stated in [7, p. 86].

Theorem 6.1. *Let us consider the differential-algebraic system* (1.1a)–(1.1c) *of index* 3 *with consistent initial values* (y_0, z_0, u_0) *at* x_0 *and the RK method* (2.1). *Let us assume that the RK coefficients satisfy* $B(p)$, $C(q)$ *with* $q \geqslant 2$, $D(r)$, (I), *and* (S). *Then for* $x_n - x_0 = nh \leqslant$ Const, *the global error satisfies*

$$\begin{aligned} &y_n - y(x_n) = \mathrm{O}(h^{\min(p,2q-2,q+r)}), && z_n - z(x_n) = \mathrm{O}(h^q), \\ &P_z(x_n)\big(z_n - z(x_n)\big) = \mathrm{O}(h^{\min(p,2q-2,q+r)}), && u_n - u(x_n) = \mathrm{O}(h^{q-1}). \end{aligned} \tag{6.1}$$

Remarks.

(1) If in addition the function k of (1.1) is linear in u, we get

$$y_n - y(x_n) = O(h^{\min(p,2q-1,q+r)}),$$
$$P_z(x_n)\big(z_n - z(x_n)\big) = O(h^{\min(p,2q-1,q+r)}), \tag{6.1'}$$

and the assumption $C(2)$ can be omitted. In the proof in this case we have $\gamma = \min(q - 1, \max(r - 1, 0))$, $\delta = \min(q - 1, r)$, k and ℓ are given by (5.3') and (5.4'), and the terms $\|(Q_z)_n \Delta z_n\|^2$ in (6.2b)–(6.2d) have to be replaced by $h\|(Q_z)_n \Delta z_n\|^2$. In the situation of $q = 1$, the proof given in [7, Theorem 6.4] requires slight modifications.

(2) The theorem remains valid in the case of variable stepsizes with $h = \max_i h_i$.

Proof of Theorem 6.1. In a first step, we can show that global convergence of order q for the (y, z)-components occurs (see [7, Theorem 6.4]).

In the second step, we use once again the techniques of the previous step. We denote two neighbouring RK solutions by $\{\tilde{y}_n, \tilde{z}_n\}$, $\{\hat{y}_n, \hat{z}_n\}$ and their difference by $\Delta y_n = \tilde{y}_n - \hat{y}_n$, $\Delta z_n = \tilde{z}_n - \hat{z}_n$. We assume that $\Delta y_n = O(h^{q+1})$ and $\Delta z_n = O(h^{q+1})$ (see [7, Formula (6.28)]). Because of $g(\tilde{y}_n) = 0 = g(\hat{y}_n)$ the fourth remark of Theorem 3.2 holds, implying that

$$(Q_y)_n \Delta y_n = O(\|\Delta y_n\|^2) = O(h^{q+1}\|(P_y)_n \Delta y_n\|). \tag{6.2a}$$

By the use of the results of the first step, Theorem 3.4 can be applied with $\delta = 0$, $\mu = 0$, and $\theta = 0$, yielding

$$\begin{aligned}&(P_y)_{n+1}\Delta y_{n+1}\\ &\quad = (P_y)_n \Delta y_n + h(f_z)_n (P_z)_n \Delta z_n\\ &\qquad + O\left(h\|(P_y)_n \Delta y_n\| + h^2\|(P_z)_n \Delta z_n\| + h^{\gamma+2}\|(Q_z)_n \Delta z_n\| + \|(Q_z)_n \Delta z_n\|^2\right),\end{aligned} \tag{6.2b}$$

$$\begin{aligned}&h(P_z)_{n+1}\Delta z_{n+1}\\ &\quad = h(P_z)_n \Delta z_n\\ &\qquad + O\left(h^2\|(P_y)_n \Delta y_n\| + h^2\|(P_z)_n \Delta z_n\| + h^{\delta+2}\|(Q_z)_n \Delta z_n\| + \|(Q_z)_n \Delta z_n\|^2\right),\end{aligned} \tag{6.2c}$$

$$h(Q_z)_{n+1}\Delta z_{n+1} = O\left(h\|(P_y)_n \Delta y_n\| + h^2\|\Delta z_n\| + \|(Q_z)_n \Delta z_n\|^2\right), \tag{6.2d}$$

where $\gamma = \min(q - 2, \max(r - 1, 0))$, $\delta = \min(q - 2, r)$ and $(P_y)_n$, $(Q_y)_n$, $(P_z)_n$, $(Q_z)_n$, $(f_z)_n$ are evaluated at $(\hat{y}_n, \hat{z}_n, \hat{u}_n^*)$. We define $\hat{u}_n^* := G(\hat{y}_n, \hat{z}_n)$ with G as described in the Introduction. This choice of $\hat{u}_n^*$ does not influence the values $(\hat{y}_n, \hat{z}_n)$ (see the second remark of Theorem 3.1 and the second remark of Section 2) and simplifies the proof. The estimates (6.2) lead to (by induction or similarly to the proof of [14, Theorem 3.3])

$$\|\Delta y_n\| \leqslant C\left(\|(P_y)_0 \Delta y_0\| + \|(P_z)_0 \Delta z_0\| + h^{\delta+1}\|(Q_z)_0 \Delta z_0\|\right), \tag{6.3a}$$

$$h\|(P_z)_n \Delta z_n\| \leqslant C\left(h\|(P_y)_0 \Delta y_0\| + h\|(P_z)_0 \Delta z_0\| + h^{\delta+2}\|(Q_z)_0 \Delta z_0\|\right), \tag{6.3b}$$

$$h\|(Q_z)_n \Delta z_n\| \leqslant C\left(h\|(P_y)_0 \Delta y_0\| + h\|(P_z)_0 \Delta z_0\| + h^2\|(Q_z)_0 \Delta z_0\|\right). \tag{6.3c}$$

Hence it follows from standard techniques (see [7, Fig. 4.1, p. 36] or [8, Fig. II.3.2, p. 160]) that

$$\begin{aligned} &y_n - y(x_n) = \mathrm{O}(h^{\min(k,\ell,q+\delta)}), \\ &P_z(x_n)\big(z_n - z(x_n)\big) = \mathrm{O}(h^{\min(k,\ell,q+\delta)}), \end{aligned} \tag{6.4}$$

where k and ℓ are given in (5.3) and (5.4). □

The estimates (6.1) and (6.1′) give us more insight into the structure of the global error for the z-component. If the numerical solution is projected onto the manifold $(g_y f)(y,z) = 0$, then the accuracy of the z-component can be improved. This can be done similarly as for index-2 problems (see [9, Section VI.7, p. 513] and [1, Section 3]), but here theoretically merely at the end of the integration process. We thus obtain:

Corollary 6.2. *Under the assumptions of Theorem* 6.1, *let* $\widehat{z}_n, \mu_n$ *be the solution of*

$$\widehat{z}_n = z_n + k_u(y_n, z_n, u_n)\mu_n, \qquad 0 = (g_y f)(y_n, \widehat{z}_n), \tag{6.5}$$

where (y_n, z_n, u_n) *is the numerical solution* (2.1) *after n steps. Then we get*

$$\widehat{z}_n - z(x_n) = \begin{cases} \mathrm{O}(h^{\min(p,2q-1,q+r)}), & \text{if } k \text{ is linear in } u, \\ \mathrm{O}(h^{\min(p,2q-2,q+r)}), & \text{else.} \end{cases} \tag{6.6}$$

Moreover, if we define $\widehat{u}_n$ *as the solution together with* y_n *and* $\widehat{z}_n$ *of* (1.1e), *we have*

$$\widehat{u}_n - u(x_n) = \begin{cases} \mathrm{O}(h^{\min(p,2q-1,q+r)}), & \text{if } k \text{ is linear in } u, \\ \mathrm{O}(h^{\min(p,2q-2,q+r)}), & \text{else.} \end{cases} \tag{6.7}$$

The problem (1.1)–(1.2) is not ill-posed if all constraints (1.1c)–(1.1e) are taken into account and not only (1.1c) since it is of index 1. It is in fact preferable to effect the projection (6.5) after every step, because the numerical solution is stabilized as concerns the influence of perturbations (see formulas (3.6c) and (3.19b)). Nevertheless, a very accurate approximation of the u-component of the solution may be unnecessary (see the second remark of Theorem 3.1), therefore the projection onto the manifold (1.1e) can be omitted. This remark is important if one wants to avoid the computation of extra derivatives such as g_{yy}. For stiffly accurate RK methods a fairly good choice is often given by $u_1 := U_s$.

Corollary 6.3. *For the s-stage* ($s \geqslant 3$) *Lobatto IIIC method applied to the index-3 system* (1.1a)–(1.1c), *the global error satisfies*

$$\begin{aligned} &y_n - y(x_n) = \mathrm{O}(h^{2s-4}), && P_z(x_n)\big(z_n - z(x_n)\big) = \mathrm{O}(h^{2s-4}), \\ &z_n - z(x_n) = \mathrm{O}(h^{s-1}), && \widehat{z}_n - z(x_n) = \mathrm{O}(h^{2s-4}), \\ &u_n - u(x_n) = \mathrm{O}(h^{s-2}), && \widehat{u}_n - u(x_n) = \mathrm{O}(h^{2s-4}). \end{aligned} \tag{6.8}$$

Moreover, if k is linear in u we have ($s \geqslant 2$)

$$\begin{aligned} &y_n - y(x_n) = \mathrm{O}(h^{2s-3}), && P_z(x_n)\big(z_n - z(x_n)\big) = \mathrm{O}(h^{2s-3}), \\ &\widehat{z}_n - z(x_n) = \mathrm{O}(h^{2s-3}), && \widehat{u}_n - u(x_n) = \mathrm{O}(h^{2s-3}). \end{aligned} \tag{6.8′}$$

Proof. The proof is obtained by putting $p = 2s - 2$, $q = s - 1$, $r = s - 1$ in (6.1), (6.1′), (6.6), and (6.7). □

The next result provides an alternative proof of [11, Corollary 2.3] demonstrated with completely different techniques.

Corollary 6.4. *For the s-stage* ($s \geqslant 2$) *Radau IIA method applied to the index-3 system* (1.1a)–(1.1c), *the global error satisfies*

$$\begin{aligned} &y_n - y(x_n) = O(h^{2s-2}), && P_z(x_n)\big(z_n - z(x_n)\big) = O(h^{2s-2}),\\ &z_n - z(x_n) = O(h^s), && \hat{z}_n - z(x_n) = O(h^{2s-2}),\\ &u_n - u(x_n) = O(h^{s-1}), && \hat{u}_n - u(x_n) = O(h^{2s-2}). \end{aligned} \tag{6.9}$$

Moreover, if k is linear in u we have ($s \geqslant 1$)

$$\begin{aligned} &y_n - y(x_n) = O(h^{2s-1}), && P_z(x_n)\big(z_n - z(x_n)\big) = O(h^{2s-1}),\\ &\hat{z}_n - z(x_n) = O(h^{2s-1}), && \hat{u}_n - u(x_n) = O(h^{2s-1}). \end{aligned} \tag{6.9′}$$

Proof. The proof is obtained by putting $p = 2s - 1$, $q = s$, $r = s - 1$ in (6.1), (6.1′), (6.6), and (6.7). □

Remark. For a constant-stepsize application of the implicit Euler method ($s = 1$) with k linear in u, it can be shown that $u_n - u(x_n) = O(h)$ for $n \geqslant 2$ (see [3] and [7, p. 90]).

An application of Theorem 6.1 concerns the convergence analysis of Runge–Kutta methods when applied to stiff mechanical systems (see [13, Theorem 3.1]).

7. Numerical experiments

The global convergence results of Section 6 have been confirmed by numerical tests.

Example 7.1. Consider the following index-3 problem

$$\begin{aligned} &y_1' = 2y_1y_2z_1z_2, && y_2' = -y_1y_2z_2^2,\\ &z_1' = (y_1y_2 + z_1z_2)u, && z_2' = -y_1y_2^2z_2^2u,\\ &0 = y_1y_2^2 - 1, \end{aligned} \tag{7.1}$$

which is of the form (1.1a)–(1.1c) with k linear in u. For the consistent initial values $y_0 = (1,1)^T$, $z_0 = (1,1)^T$, and $u_0 = 1$ the exact solution is given by

$$y_1(x) = z_1(x) = e^{2x}, \qquad y_2(x) = z_2(x) = e^{-x}, \qquad u(x) = e^x. \tag{7.2}$$

In Fig. 1 the global errors at $x_{end} = 0.1$ for the Lobatto IIIC methods ($s = 2, 3, 4, 5, 6$) applied to (7.1) are plotted as functions of h (the stepsizes have been chosen alternatively as $\frac{1}{3}h$ and $\frac{2}{3}h$). Since

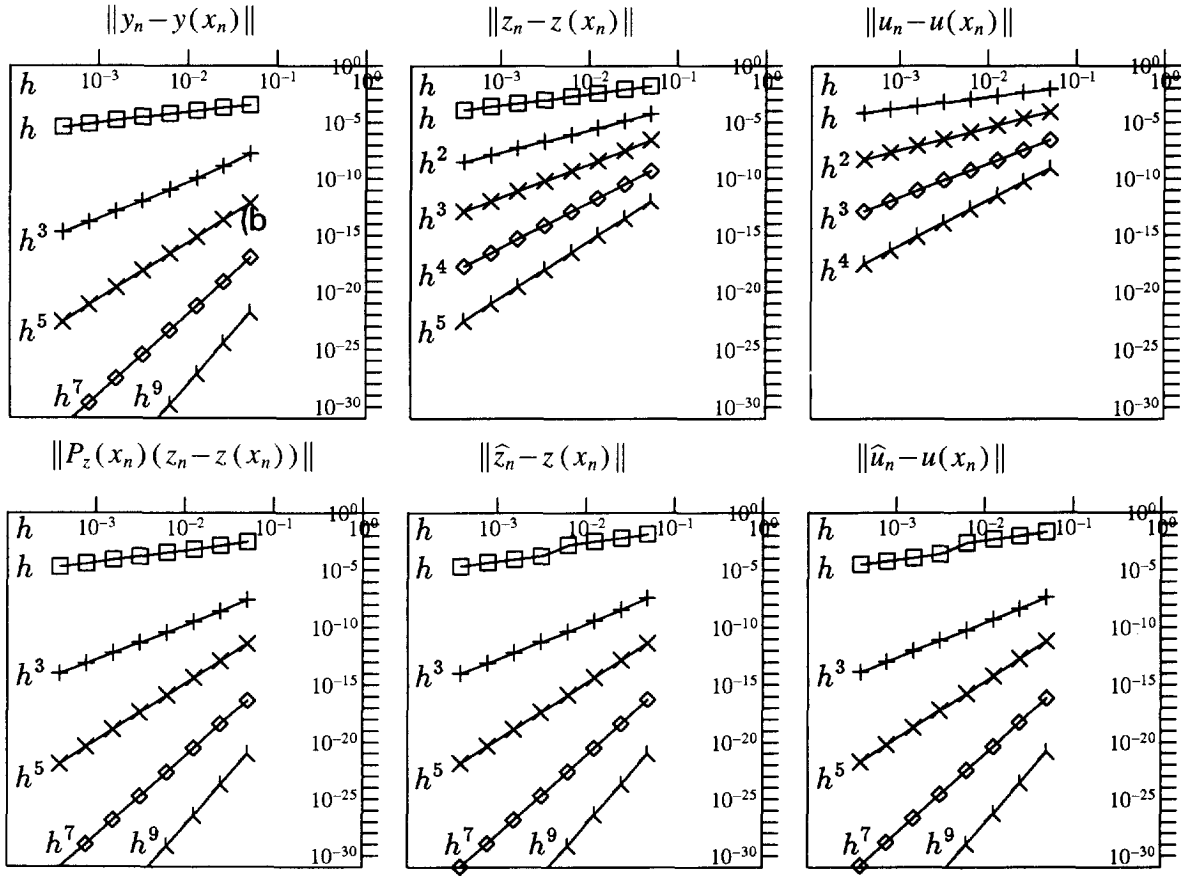

Fig. 1. Global errors of the Lobatto IIIC methods ($s = 2$: □; $s = 3$: +; $s = 4$: ×; $s = 5$: ◇; $s = 6$: ⩚).

we have used logarithmic scales, the curves appear as straight lines of slope k whenever the leading term of the error is $O(h^k)$. This behaviour is indicated in the figures.

Example 7.2. This example is a slight modification of problem (7.1) where the equation for z_2' has been replaced by $z_2' = -y_1 y_2^2 z_2^3 u^2$, so that k becomes nonlinear in u. With the same consistent initial values defined previously, the exact solution is identical to (7.2). The Radau IIA methods ($s = 2, 3, 4, 5$) have been applied to this modified problem and in Fig. 2 the global errors at $x_{\text{end}} = 0.1$ are plotted.

The observed orders of convergence match the theoretical results, showing clearly the importance of the projections described in Section 6 in order to improve the accuracy of the (z, u)-components.

Acknowledgements

I am grateful to E. Hairer for his critical remarks and his many valuable suggestions. I wish to thank S.F. Bernatchez and K. Breaden for their careful reading of the English part of the manuscript.

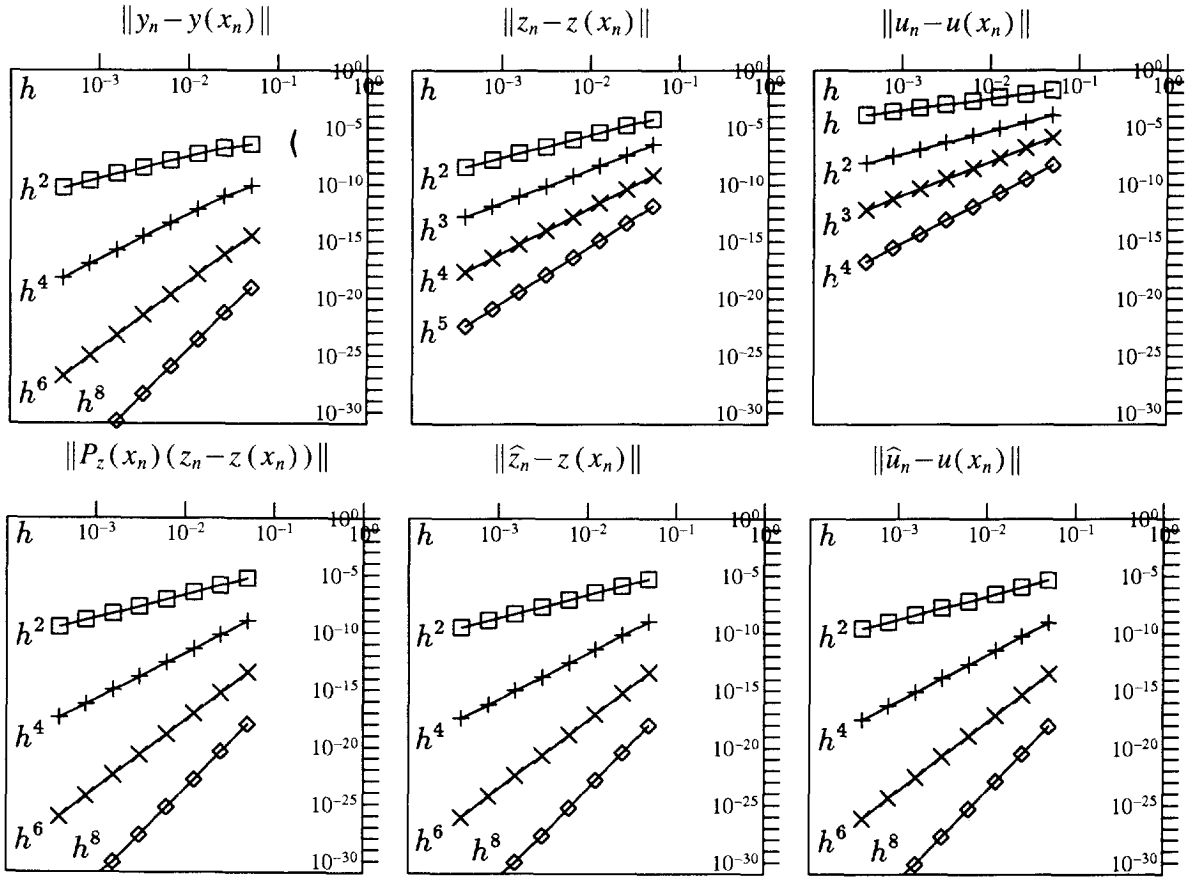

Fig. 2. Global errors of the Radau IIA methods ($s = 2$: □; $s = 3$: +; $s = 4$: ×; $s = 5$: ◇).

References

[1] U.M. Ascher and L.R. Petzold, Projected implicit Runge–Kutta methods for differential-algebraic equations, *SIAM J. Numer. Anal.* 28 (1991) 1097–1120.

[2] K.E. Brenan, S.L. Campbell and L.R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations* (North-Holland, New York, 1989).

[3] K.E. Brenan and B.E. Engquist, Backward differentiation approximations of nonlinear differential/algebraic systems, *Math. Comp.* 51 (1988) 659–676; Supplement *Math. Comp.* 51 (1988) S7–S16.

[4] K. Brenan and L.R. Petzold, The numerical solution of higher index differential/algebraic equations by implicit Runge–Kutta methods, *SIAM J. Numer. Anal.* 26 (1989) 976–996.

[5] J.C. Butcher, Coefficients for the study of Runge–Kutta integration processes, *J. Austral. Math. Soc.* 3 (1963) 185–201.

[6] J.C. Butcher, *The numerical analysis of ordinary differential equations. Runge–Kutta and general linear methods* (Wiley, Chichester, England, 1987).

[7] E. Hairer, C. Lubich and M. Roche, *The Numerical Solution of Differential-Algebraic Systems by Runge–Kutta Methods*, Lecture Notes in Mathematics 1409 (Springer-Verlag, Berlin, 1989).

[8] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I. Nonstiff Problems*, Computational Mathematics 8 (Springer-Verlag, Berlin, 2nd rev. ed., 1993).

[9] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, Computational Mathematics 14 (Springer-Verlag, Berlin, 1991).

[10] L. Jay, Convergence of a class of Runge–Kutta methods for differential-algebraic systems of index 2, *BIT* 33 (1993) 137–150.
[11] L. Jay, Collocation methods for differential-algebraic equations of index 3, *Numer. Math.* 65 (1993) 407–421.
[12] L. Jay, Runge–Kutta type methods for index three differential-algebraic equations with applications to Hamiltonian systems, Ph.D. Thesis, University of Geneva, Switzerland (1994).
[13] C. Lubich, Integration of stiff mechanical systems by Runge–Kutta methods, *Z. Angew. Math. Phys.* 44 (1993) 1022–1053.
[14] A. Ostermann, A class of half-explicit Runge–Kutta methods for differential-algebraic systems of index 3, *Appl. Numer. Math.* 13 (1993) 165–179.
[15] M. Roche, Implicit Runge–Kutta for differential algebraic equations, *SIAM J. Numer. Anal.* 26 (1989) 963–975.