BIT 33 (1993), 137–150.

CONVERGENCE OF A CLASS OF RUNGE-KUTTA METHODS FOR DIFFERENTIAL-ALGEBRAIC SYSTEMS OF INDEX 2

LAURENT JAY

Université de Genève, Département de mathématiques, Rue du Lièvre 2–4, Case postale 240, CH-1211 Genève 24, Switzerland.
e-mail: jay@cgeuge51.bitnet or: jay@uni2a.unige.ch.

Abstract.

This paper deals with convergence results for a special class of Runge-Kutta (RK) methods as applied to differential-algebraic equations (DAE's) of index 2 in Hessenberg form. The considered methods are stiffly accurate, with a singular RK matrix whose first row vanishes, but which possesses a nonsingular submatrix. Under certain hypotheses, global superconvergence for the differential components is shown, so that a conjecture related to the Lobatto IIIA schemes is proved. Extensions of the presented results to projected RK methods are discussed. Some numerical examples in line with the theoretical results are included.

Subject classifications: AMS(MOS): 65L06.

Key words: Differential-algebraic, index 2, initial value problems, Runge-Kutta methods.

1. Introduction.

Differential-algebraic equations (DAE's) of index 2 arise in many applications, such as in mechanical modelling of constrained systems (see [4, pp. 6–7] or [5, pp. 483–486 & 539–540]). Whereas optimal convergence results for Runge-Kutta (RK) methods with an invertible RK matrix are well-known (see [4, Section 4] and [5, Section VI.7]), this paper is concerned only with RK methods having a singular RK matrix.

The main result of this article (Theorem 5.2 below) proves a conjecture (see [4, pp. 18, 46 & 47] and [5, p. 515]) related to the Lobatto IIIA processes which belong to the class of methods considered in this paper (see Section 2). Its proof necessitates several preliminary results which are collected in Section 3 (properties of the RK coefficients), in Section 4 (existence, uniqueness of the numerical solution, and influence of perturbations), and in Section 5 (estimates of the local error and

Received June 1991. Revised May 1992.

convergence). Extensions of the previous results to projected RK methods are discussed in Section 6. Finally, some numerical experiments are given in Section 7 which illustrate the theoretical results. Let us mention that all the results presented in this paper remain valid for some other types of DAE's (for further details see [4, pp. 5 & 30]).

In this report, we consider the following system of DAE's given in an autonomous and semi-explicit formulation (or Hessenberg form)

$$(1.1)\qquad \begin{aligned} y' &= f(y,z), \qquad & y(x_0) &= y_0 \in \mathbb{R}^n, \\ 0 &= g(y), \qquad & z(x_0) &= z_0 \in \mathbb{R}^m \end{aligned}$$

where the initial values (y_0, z_0) are assumed to be consistent, i.e.,

$$(1.2)\qquad g(y_0) = 0, \qquad (g_y f)(y_0, z_0) = 0.$$

We suppose that f and g are sufficiently differentiable and that

$$(1.3)\qquad (g_y f_z)(y,z) \quad \text{is invertible}$$

in a neighbourhood of the exact solution (*index* 2).

2. The class of Runge-Kutta methods.

One step of an *s-stage Runge-Kutta* (*RK*) method applied to (1.1) reads (see [3], [4, p. 30] or [5, p. 502])

$$(2.1\text{a})\qquad y_1 = y_0 + \sum_{i=1}^{s} b_i k_i, \qquad z_1 = z_0 + \sum_{i=1}^{s} b_i l_i$$

where

$$(2.1\text{b})\qquad k_i = hf(Y_i, Z_i), \qquad 0 = g(Y_i)$$

and the internal stages are given by

$$(2.1\text{c})\qquad Y_i = y_0 + \sum_{j=1}^{s} a_{ij} k_j, \qquad Z_i = z_0 + \sum_{j=1}^{s} a_{ij} l_j.$$

For a RK method we denote $A := (a_{ij})_{i,j}$ the *RK matrix*, $b := (b_1, \ldots, b_s)^T$ the *weight vector*, and $c := (c_1, \ldots, c_s)^T := A\mathbb{1}_s$ the *node vector* where $\mathbb{1}_s := (1, \ldots, 1)^T$. Let $B(p)$, $C(q)$, $D(r)$ be the following *simplifying assumptions* which are related to the construction of such methods

$$B(p)\colon \sum_{i=1}^{s} b_i c_i^{k-1} = 1/k \qquad\qquad k = 1, \ldots, p;$$

$$C(q)\colon \sum_{j=1}^{s} a_{ij} c_j^{k-1} = c_i^k/k \qquad\qquad i = 1, \ldots, s, \quad k = 1, \ldots, q;$$

$$D(r)\colon \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = b_j(1 - c_j^k)/k \qquad j = 1,\ldots,s, \qquad k = 1,\ldots,r.$$

Throughout this paper we are only interested in RK methods with $s \geq 2$ and coefficients satisfying the hypotheses

$H1$: $a_{1j} = 0$ for $j = 1,\ldots,s$;

$H2$: the submatrix $\tilde{A} := (a_{ij})_{i,j\geq 2}$ is invertible;

$H3$: $b_i = a_{si}$ for $i = 1,\ldots,s$, i.e., the method is *stiffly accurate.*

For these methods, the l_j in (2.1) are not well-defined, but in order to define y_1 and z_1, it is sufficient to solve the equivalent nonlinear system (4.2) below and to apply the fourth remark hereafter.

REMARKS. The following results can be easily proven.

1) The definition of c coincides with the condition $C(1)$.
2) $H3$ together with the condition $B(1)$ leads to $c_s = 1$. If in addition $C(q)$ (resp. $D(r)$) is satisfied then $B(q)$ (resp. $B(r+1)$) holds.
3) From $H1$ it follows that $c_1 = 0$, $\quad Y_1 = y_0$, $\quad g(Y_1) = g(y_0) = 0$, $\quad Z_1 = z_0$ in (2.1), and that A is singular.
4) $H3$ implies that $y_1 = Y_s$, $\quad g(y_1) = g(Y_s) = 0$, and $z_1 = Z_s$ in (2.1).

A main advantage of methods verifying $H1$ and $H3$ is that the first stage of one step is equal to the last stage of the previous step which coincides with the current initial value, so that it requires no supplementary computation. The most prominent examples of such methods are given by collocation methods like the Lobatto IIIA schemes whose coefficients $c_1 = 0$, $\quad c_2,\ldots,c_s = 1$ are the zeros of the polynomial of degree s

$$\frac{d^{s-2}}{dx^{s-2}}\left(x^{s-1}(x-1)^{s-1}\right) \tag{2.2}$$

and which fulfil the conditions $B(2s-2)$, $C(s)$, and $D(s-2)$. Due to their symmetry, they are often used for the solution of boundary value problems (see [2]).

3. Properties of Runge-Kutta coefficients.

This section deals with relations of the RK coefficients appearing in the demonstration of Theorem 4.4.

THEOREM 3.1. *Suppose that the hypotheses* $H1$, $H2$ and $H3$ *are satisfied together with the condition* $D(r)$. *For a fixed* $\rho \in \mathbb{N}\backslash\{0\}$, *consider a multi-index* $v = (v_1,\ldots,v_\rho)$ *satisfying* $v_i \geq 1$ *and let* $\alpha \geq 0$. *If* $|v| := \sum_{i=1}^{\rho} v_i \leq r$ *then we have*

$$e_{s-1}^T \tilde{C}^{\alpha}\left(\prod_{i=1}^{\rho} M_i\right)\tilde{A}_1 = e_{s-1}^T\tilde{C}^{\alpha}M_1\ldots M_{\rho}\tilde{A}_1 = 0 \tag{3.1}$$

where we have set

$$\tilde{C} := \operatorname{diag}(c_2,\ldots,c_s), \quad \tilde{A}_1 := (a_{21},\ldots,a_{s1})^T, \quad e_{s-1} := (0,\ldots,0,1)^T. \tag{3.2}$$

The matrices M_i are of the form $\tilde{A}^{\nu_i}$, $\tilde{C}^{\nu_i}$ or $\tilde{A}\tilde{C}^{\nu_i}\tilde{A}^{-1}$ and it is supposed that $M_{\rho} = \tilde{A}C^{\nu_\rho}\tilde{A}^{-1}$.

Remark. Without loss of generality $\rho \le r$ can be assumed.

Proof. In matrix notation, the simplifying assumption $D(r)$ becomes

$$\tilde{b}^T\tilde{C}^{k-1}\tilde{A} = k^{-1}(\tilde{b}^T - \tilde{b}^T\tilde{C}^k) \qquad k = 1,\ldots,r, \tag{3.3a}$$

$$\tilde{b}^T\tilde{C}^{k-1}\tilde{A}_1 = k^{-1}b_1 \qquad k = 1,\ldots,r \tag{3.3b}$$

where $\tilde{b} := (b_2,\ldots,b_s)^T$, and $H3$ reads

$$\tilde{b}^T = e_{s-1}^T\tilde{A} \tag{3.4a}$$

$$b_1 = e_{s-1}^T\tilde{A}_1 = a_{s1}. \tag{3.4b}$$

Multiplying (3.3a) with $\tilde{A}^{-1}$ and using $\tilde{b}^T\tilde{A}^{-1} = e_{s-1}^T$ which follows from (3.4a), we obtain

$$\tilde{b}^T\tilde{C}^k\tilde{A}^{-1} = e_{s-1}^T - k\tilde{b}^T\tilde{C}^{k-1}; \qquad k = 1,\ldots,r. \tag{3.5}$$

Repeated application of (3.3a), (3.4a), and (3.5) to (3.1) shows that this expression is a linear combination of terms $\tilde{b}^T\tilde{C}^{\gamma}\tilde{A}^{-1}\tilde{A}_1$ with $1 \le \gamma \le r$. They all vanish because of

$$\tilde{b}^T\tilde{C}^{\gamma}\tilde{A}^{-1}\tilde{A}_1 = e_{s-1}^T\tilde{A}_1 - \gamma\tilde{b}^T\tilde{C}^{\gamma-1}\tilde{A}_1 = b_1 - b_1 = 0 \tag{3.6}$$

which is a consequence of (3.5), (3.4b), and (3.3b). ■

Lemma 3.2. *Suppose that the hypotheses $H1$, $H2$, and $H3$ hold. Then $R(z)$, the stability function of the method, satisfies at ∞*

$$R(\infty) = -e_{s-1}^T\tilde{A}^{-1}\tilde{A}_1. \tag{3.7}$$

Proof. $R(z)$ is the numerical solution after one step of the method applied to the test equation

$$y' = \lambda y, \qquad y_0 = 1, \tag{3.8}$$

with $z := h\lambda$. By using $H3$ we get $R(z) = y_1 = Y_s = e_s^T(I_s - zA)^{-1}\mathbb{1}_s$. The result follows from

(3.9) $(I_s - zA)^{-1}$

$$= \begin{pmatrix} 1 & 0 \\ (I_{s-1} - z\tilde{A})^{-1} z\tilde{A}_1 & (I_{s-1} - z\tilde{A})^{-1} \end{pmatrix} \xrightarrow{z \to \infty} \begin{pmatrix} 1 & 0 \\ -\tilde{A}^{-1}\tilde{A}_1 & 0 \end{pmatrix}.$$

■

4. Existence, uniqueness and influence of perturbations.

This section is mainly devoted to the demonstration of Theorem 4.4, which is the fundamental result. We first investigate existence and uniqueness of the solution of the nonlinear system (2.1) where (y_0, z_0) are replaced by approximate h-dependent starting values (η, ζ).

THEOREM 4.1. *Suppose that*

(4.1a) $$g(\eta) = 0,$$

(4.1b) $$(g_y f)(\eta, \zeta) = O(h),$$

(4.1c) $(g_y f_z)(y, z)$ is invertible in a neighbourhood of (η, ζ),

and that the RK coefficients verify the hypotheses $H1$ and $H2$. *Then for* $h \leq h_0$ *there exists a locally unique solution to*

$$\left.\begin{aligned} &(4.2a) \quad Y_i = \eta + h \sum_{j=1}^{s} a_{ij} f(Y_j, Z_j) \\ &(4.2b) \quad 0 = g(Y_i) \end{aligned}\right\} \quad i = 1, \ldots, s$$

with $Z_1 := \zeta$ *and which satisfies*

(4.3) $$Y_i - \eta = O(h), \qquad Z_i - \zeta = O(h).$$

REMARKS.

1) $Y_1 = \eta$, implied by $H1$, shows the necessity of (4.1a).
2) The value of ζ in (4.1b) specifies the solution branch of $(g_y f)(y, z) = 0$ to which the numerical solution is close.

We omit the *proof* which can be obtained similarly as in [4, Theorem 4.1] or [5, Chapter VI, Theorem 7.1] covering the case of invertible RK matrix A. ■

Our next result is concerned with the influence of perturbations to (4.2).

THEOREM 4.2. *Let* Y_i, Z_i *be the solution of* (4.2) *and consider perturbed values* $\hat{Y}_i, \hat{Z}_i$ *satisfying*

$$\hat{Y}_i = \hat{\eta} + h \sum_{j=1}^{s} a_{ij} f(\hat{Y}_j, \hat{Z}_j) + h\delta_i \tag{4.4a}$$

$$0 = g(\hat{Y}_i) + \theta_i \qquad i = 1, \ldots, s \tag{4,4b}$$

with $\hat{Z}_1 := \hat{\zeta}$. *In addition to the assumptions of Theorem* 4.1, *suppose that*

$$\hat{\eta} - \eta = O(h), \quad \hat{Z}_i - \zeta = O(h), \quad \delta_i = O(h), \quad \theta_i = O(h^2). \tag{4.5}$$

Then we have for $h \leq h_0$ *the estimates*

$$\|\hat{Y}_i - Y_i\| \leq C(\|\hat{\eta} - \eta\| + h^2 \|\hat{\zeta} - \zeta\| + h\|\delta\| + \|\theta\|), \tag{4.6a}$$

$$\|\hat{Z}_i - Z_i\| \leq \frac{C}{h}(h\|\hat{\eta} - \eta\| + h\|\hat{\zeta} - \zeta\| + h\|\delta\| + \|\theta\|) \tag{4.6b}$$

where $\delta = (\delta_1, \ldots, \delta_s)^T$, $\|\delta\| = \max_i \|\delta_i\|$ *and similarly for* θ.

REMARKS.
1) The conditions (4.5) ensure that all terms $O(\cdot)$ in the proof below are small.
2) The terms containing $\hat{\zeta} - \zeta$ will be computed in detail. This will be justified in the demonstration of Theorem 4.4.
3) We introduce the notation $\Delta\eta = \hat{\eta} - \eta$, $\Delta\zeta = \hat{\zeta} - \zeta$, $Y = (Y_1, \ldots, Y_s)^T$, $\Delta Y = \hat{Y} - Y$, $\|\Delta Y\| = \max_i \|\Delta Y_i\|$ and similarly for the z-component. Over a multiple-vector a tilde '$\sim$' indicates the removal of its first subvector, e.g., $\tilde{Y} = (Y_2, \ldots, Y_s)^T$.

PROOF. $H1$ implies that $Y_1 = \eta$ and $\hat{Y}_1 = \hat{\eta} + h\delta_1$. Therefore we have

$$\Delta Y_1 = \Delta\eta + h\delta_1, \quad \Delta Z_1 = \Delta\zeta \tag{4.7}$$

which proves the statement (4.6) for $i = 1$. Hence from (4.2b) and (4.4b) we deduce that

$$g_y(\eta)\Delta\eta = O(h\|\Delta\eta\| + h\|\delta_1\| + \|\theta_1\|). \tag{4.8}$$

For $i \geq 2$, by subtracting (4.2) from (4.4) we obtain by linearization

$$\begin{aligned} \Delta Y_i = \Delta\eta &+ h \sum_{j=1}^{s} a_{ij} f_y(Y_j, Z_j)\Delta Y_j + h \sum_{j=1}^{s} a_{ij} f_z(Y_j, Z_j)\Delta Z_j \\ &+ h\delta_i + O(h\|\Delta Y\|^2 + h\|\Delta Y\| \cdot \|\Delta Z\| + h\|\Delta Z\|^2), \end{aligned} \tag{4.9a}$$

$$0 = g_y(Y_i)\Delta Y_i + \theta_i + O(\|\Delta Y_i\|^2). \tag{4.9b}$$

It can be noticed that if $f_{zz} = 0$ (f linear in z) the expression $O(h\|\Delta Z\|^2)$ in (4.9a) disappears, but $O(h\|\Delta Y\| \cdot \|\Delta Z\|)$ remains. Therefore we retain all terms permitting to analyse easily this situation (see the first remark after Theorem 4.4). By using tensor notation, (4.9) can be rewritten with the help of (4.7) as

$$\begin{aligned}(4.10a)\qquad \Delta\tilde{Y} &= \mathbb{1}_{s-1}\otimes\Delta\eta + h(\tilde{A}\otimes I_n)\{f_y\}\Delta\tilde{Y} + \tilde{A}_1\otimes(f_z(\eta,\zeta)h\Delta\zeta)\\ &\quad + (\tilde{A}\otimes I_n)\{f_z\}h\Delta\tilde{Z} + h\tilde{\delta} + O(h\,\|\Delta\tilde{Y}\|^2 + h\,\|\Delta\tilde{Y}\|\cdot\|\Delta Z\|\\ &\quad + h\,\|\Delta Z\|^2 + h\,\|\Delta\eta\| + h^2\,\|\delta_1\|),\end{aligned}$$

$$(4.10b)\qquad 0 = \{g_y\}\Delta\tilde{Y} + \tilde{\theta} + O(\|\Delta\tilde{Y}\|^2)$$

where

$$(4.11a)\qquad \{f_y\} := \operatorname{blockdiag}(f_y(Y_2,Z_2),\dots,f_y(Y_s,Z_s)),$$

$$(4.11b)\qquad \{f_z\} := \operatorname{blockdiag}(f_z(Y_2,Z_2),\dots,f_z(Y_s,Z_s)),$$

$$(4.11c)\qquad \{g_y\} := \operatorname{blockdiag}(g_y(Y_2),\dots,g_y(Y_s)).$$

Insertion of the expression (4.10a) into (4.10b) yields

$$\begin{aligned}(4.12)\qquad &-\{g_y\}(\tilde{A}\otimes I_n)\{f_z\}h\Delta\tilde{Z} =\\ &\{g_y\}(\mathbb{1}_{s-1}\otimes\Delta\eta + h(\tilde{A}\otimes I_n)\{f_y\}\Delta\tilde{Y} + \tilde{A}_1\otimes(f_z(\eta,\zeta)h\Delta\zeta) + h\tilde{\delta})\\ &+\tilde{\theta} + O(\|\Delta\tilde{Y}\|^2 + h\,\|\Delta\tilde{Y}\|\cdot\|\Delta Z\| + h\,\|\Delta Z\|^2 + h\,\|\Delta\eta\| + h^2\,\|\delta_1\|).\end{aligned}$$

In view of (4.3) we have

$$(4.13)\qquad g_y(Y_i)a_{ij}f_z(Y_j,Z_j) = a_{ij}(g_yf_z)(\eta,\zeta) + O(h),$$

thus the left matrix of (4.12) can be written as

$$(4.14)\qquad \{g_y\}(\tilde{A}\otimes I_n)\{f_z\} = \tilde{A}\otimes(g_yf_z)(\eta,\zeta) + O(h)$$

and is invertible by $H2$ and (4.1c) if h is sufficiently small. Hence from (4.12), and by the use of (4.8) for (4.15′), we get

$$\begin{aligned}(4.15)\qquad h\Delta\tilde{Z} &= -(\{g_y\}(\tilde{A}\otimes I_n)\{f_z\})^{-1}\{g_y\}\\ &\quad\times(\mathbb{1}_{s-1}\otimes\Delta\eta + h(\tilde{A}\otimes I_n)\{f_y\}\Delta\tilde{Y} + \tilde{A}_1\otimes(f_z(\eta,\zeta)h\Delta\zeta)\\ &\quad + O(\|\Delta\tilde{Y}\|^2 + h\,\|\Delta\tilde{Y}\|\cdot\|\Delta Z\| + h\,\|\Delta\tilde{Z}\|^2 + h\,\|\Delta\eta\|\\ &\quad + h\,\|\Delta\zeta\|^2 + h^2\,\|\delta_1\| + h\,\|\tilde{\delta}\| + \|\tilde{\theta}\|)\\ (4.15')\qquad &= -(\{g_y\}(\tilde{A}\otimes I_n)\{f_z\})^{-1}\{g_y\}(\tilde{A}_1\otimes(f_z(\eta,\zeta)h\Delta\zeta))\\ &\quad + O(h\,\|\Delta\tilde{Y}\| + h\,\|\Delta\tilde{Z}\|^2 + h\,\|\Delta\eta\| + h\,\|\Delta\zeta\|^2 + h\,\|\delta\| + \|\theta\|).\end{aligned}$$

(4.15) inserted into (4.10a) leads to

$$\begin{aligned}(4.16)\qquad \Delta\tilde{Y} &= P_{\tilde{A}}(\mathbb{1}_{s-1}\otimes\Delta\eta + h(\tilde{A}\otimes I_n)\{f_y\}\Delta\tilde{Y} + \tilde{A}_1\otimes(f_z(\eta,\zeta)h\Delta\zeta))\\ &\quad + O(\|\Delta\tilde{Y}\|^2 + h\,\|\Delta\tilde{Y}\|\cdot\|\Delta Z\| + h\,\|\Delta\tilde{Z}\|^2 + h\,\|\Delta\eta\|\\ &\quad + h\,\|\Delta\zeta\|^2 + h^2\,\|\delta_1\| + h\,\|\tilde{\delta}\| + \|\tilde{\theta}\|),\end{aligned}$$

with the following definitions

(4.17) $P_{\tilde{A}} := I_{(s-1)n} - F_z(G_yF_z)^{-1}G_y,\ F_z := (\tilde{A} \otimes I_n)\{f_z\}(\tilde{A} \otimes I_m)^{-1},\ G_y := \{g_y\}.$

We put $F_{z,0} := I_{s-1} \otimes f_z(\eta, \zeta)$ and since the projector $P_{\tilde{A}}$ satisfies $P_{\tilde{A}}F_z = 0$, the term including $\Delta\zeta$ in (4.16) can be expressed as

$$P_{\tilde{A}}(\tilde{A}_1 \otimes (f_z(\eta,\zeta)h\Delta\zeta)) = -P_{\tilde{A}}(F_z - F_{z,0})(\tilde{A}_1 \otimes h\Delta\zeta) = O(h^2 \|\Delta\zeta\|) \tag{4.18}$$

because of $F_z - F_{z,0} = O(h)$. ■

LEMMA 4.3. *In addition to the hypotheses of Theorem 4.1, suppose that the condition $C(q)$ holds and that $(g_yf)(\eta,\zeta) = O(h^\kappa)$ with $\kappa \geq 1$. Then the solution of (4.2), Y_i, Z_i, satisfies*

$$Y_i = \eta + \sum_{m=1}^{\lambda} \frac{c_i^m h^m}{m!} DY_m(\eta) + O(h^{\lambda+1}), \tag{4.19a}$$

$$Z_i = \zeta(\eta) + \sum_{n=1}^{\mu} \frac{c_i^n h^n}{n!} DZ_n(\eta) + O(h^{\mu+1}) \tag{4.19b}$$

where $\zeta(\eta)$ is defined by the condition $(g_yf)(\eta,\zeta(\eta)) = 0$, $\lambda = \min(\kappa+1, q)$, $\mu = \min(\kappa-1, q-1)$ and DY_m, DZ_n are functions composed by derivatives of f and g evaluated at $(\eta, \zeta(\eta))$.

PROOF. By the implicit function theorem we obtain $\zeta(\eta) - \zeta = O(h^\kappa)$. We define $(y(x), z(x))$ the solution of (1.1) which satisfies $y(x_0) = \eta$ and $z(x_0) = \zeta(\eta)$. The exact solution values $\hat{\eta} = \eta = y(x_0)$, $\hat{\zeta} = \zeta(\eta) = z(x_0)$, $\hat{Y}_i = y(x_0 + c_ih)$, $\hat{Z}_i = z(x_0 + c_ih)$ satisfy (4.4) with $\theta_i = 0$ and

$$\delta_i = \frac{h^q}{q!} y^{(q+1)}(x_0)\left(\frac{c_i^{q+1}}{q+1} - \sum_{j=1}^{s} a_{ij}c_j^q\right) + O(h^{q+1}) = O(h^q). \tag{4.20}$$

The difference from the numerical solution (4.2) can thus be estimated with Theorem 4.2, yielding

$$\|Y_i - y(x_0 + c_ih)\| = O(h^{\min(\kappa+2,q+1)}),\ \|Z_i - z(x_0 + c_ih)\| = O(h^{\min(\kappa,q)}). \tag{4.21}$$ ■

THEOREM 4.4. *In addition to the assumptions of Theorem 4.2, suppose that the conditions $C(q)$, $D(r)$ and the hypothesis H3 hold, and that $(g_yf)(\eta,\zeta) = O(h^\kappa)$ with $\kappa \geq 1$. Then we have*

$$\begin{aligned}\hat{Y}_s - Y_s = {} & P(\eta,\zeta)(\hat{\eta} - \eta) \\ & + O(h\|\hat{\eta} - \eta\| + h^{m+2}\|\hat{\zeta} - \zeta\| + h\|\hat{\zeta} - \zeta\|^2 + h\|\delta\| + \|\theta\|),\end{aligned} \tag{4.22a}$$

$$\hat{Z}_s - Z_s = R(\infty)(\hat{\zeta} - \zeta) + O(\|\hat{\eta} - \eta\| + h\|\hat{\zeta} - \zeta\| + \|\delta\| + \|\theta\|/h) \tag{4.22b}$$

where $m = \min(\kappa-1, q-1, r) \geq 0$, R is the stability function, and P is the projector defined under the condition (1.3) by

$$P := I_n - Q, \qquad Q := f_z(g_yf_z)^{-1}g_y. \tag{4.23}$$

REMARKS.

1) If the function f of (1.1) is linear in z we have $m = \min(\kappa, q, r)$. All terms $O(h\,\|\Delta\zeta\|^2)$ in the proof below can be replaced by $O(h^3\,\|\Delta\zeta\|^2)$ coming from the expression $O(h\,\|\Delta\tilde{Y}\|\cdot\|\Delta Z\|)$ of (4.16) (in this case the terms $O(h\,\|\Delta\tilde{Z}\|^2)$ and $O(h\,\|\Delta\zeta\|^2)$ are not present), so that (4.22a) becomes

$$
\begin{aligned}
(4.22\text{a}')\quad \hat{Y}_s - Y_s &= P(\eta,\zeta)(\hat{\eta} - \eta)\\
&\quad + O(h\,\|\hat{\eta} - \eta\| + h^{m+2}\,\|\hat{\zeta} - \zeta\| + h^3\,\|\hat{\zeta} - \zeta\|^2 + h\,\|\delta\| + \|\theta\|).
\end{aligned}
$$

2) The important result consists in the factor h^{m+2} in front of $\|\hat{\zeta} - \zeta\|$ in (4.22a)–(4.22a′).

PROOF. We return to the end of the proof of Theorem 4.2 by taking Lemma 3.3 into account and using the same notations and definitions. According to (4.15′) we have

$$
(4.24)\qquad \Delta Z_s = -e_{s-1}^T\tilde{A}^{-1}\tilde{A}_1\Delta\zeta + O(\|\Delta\eta\| + h\,\|\Delta\zeta\| + \|\delta\| + \|\theta\|/h)
$$

which together with formula (3.7) of Lemma 3.2 proves the statement (4.22b).

(4.22a) remains to be proved. Taking (4.16), computing $(I - hP_{\tilde{A}}(\tilde{A}\otimes I_n)\{f_y\})^{-1}$ by means of the series of von Neumann, and using (4.18), we obtain

$$
\begin{aligned}
(4.25)\qquad \Delta Y_s &= P(\eta,\zeta)\Delta\eta - (e_{s-1}^T\otimes I_m)\left(\sum_{\delta=0}^{m-1} h^\delta (P_{\tilde{A}}(\tilde{A}\otimes I_n)\{f_y\})^\delta\right)\\
&\quad\times P_{\tilde{A}}(F_z - F_{z,0})(\tilde{A}_1\otimes h\Delta\zeta) + O(h\,\|\Delta\eta\| + h^{m+2}\,\|\Delta\zeta\|\\
&\quad + h\,\|\Delta\zeta\|^2 + h^2\,\|\delta_1\| + h\,\|\tilde{\delta}\| + \|\tilde{\theta}\|).
\end{aligned}
$$

With the help of Lemma 4.3, we will develop $P_{\tilde{A}}$ into h-powers. Let us first consider the expression

$$
\begin{aligned}
(4.26)\qquad G_yF_z &= I_{s-1}\otimes(g_yf_z)(\eta,\zeta(\eta))\\
&\quad\times\left(I_{s-1}\otimes I_m + \sum_{0<i+j\le\omega} h^{i+j}(\tilde{C}^i\tilde{A}\tilde{C}^j\tilde{A}^{-1})\otimes D_{ij}(\eta)\right) + O(h^{\omega+1})
\end{aligned}
$$

where $\omega = \mu$ ($\omega = \lambda$ if f is linear in z because $f_z(y,z)$ is independent of z), and the D_{ij} are terms of the same type as the DY_m and DZ_n of Lemma 4.3. Using again the series of von Neumann, we see that its inverse is of the form

$$
\begin{aligned}
(4.27)\quad (G_yF_z)^{-1} &= I_{s-1}\otimes(g_yf_z)^{-1}(\eta,\zeta(\eta))\\
&\quad + \sum_{0<|\alpha|+|\beta|\le\omega} h^{|\alpha|+|\beta|}\left(\sum_{i=1}^{\omega}\tilde{C}^{\alpha_i}\tilde{A}\tilde{C}^{\beta_i}\tilde{A}^{-1}\right)\otimes E_{\alpha\beta}(\eta) + O(h^{\omega+1})
\end{aligned}
$$

where the $E_{\alpha\beta}$ are expressions like the D_{ij}, $\alpha = (\alpha_1,\ldots,\alpha_\omega)$ and $\beta = (\beta_1,\ldots,\beta_\omega)$ are multi-indices in $\mathbb{N}^\omega$. Here the norm of a multi-index $\gamma = (\gamma_1,\ldots,\gamma_\omega)$ is defined by

$|\gamma| := \sum_{i=1}^{\omega} \gamma_i$. If we insert $(G_y F_z)^{-1}$ into the definition of $P_{\tilde{A}}$ and develop G_y and F_z in powers of h, we arrive at

$$(4.28)\quad P_{\tilde{A}} = I_{s-1} \otimes P(\eta, \zeta(\eta)) + \sum_{0 < |\kappa| + |\nu| \le \omega} h^{|\kappa| + |\nu|} \left(\prod_{i=1}^{\omega} \tilde{A}\tilde{C}^{\kappa_i}\tilde{A}^{-1}\tilde{C}^{\nu_i} \right) \otimes H_{\kappa\nu}(\eta) + O(h^{\omega+1})$$

where $H_{\kappa\nu}$ are analogous to the D_{ij}. Further, $\kappa = (\kappa_1, \ldots, \kappa_\omega)$ and $\nu = (\nu_1, \ldots, \nu_\omega)$ are multi-indices in $\mathbb{N}^\omega$.

With these preparations we are now able to prove (4.22a) by developing into h-powers the expression including $\Delta\zeta$ in (4.25). For example we consider the term which corresponds to $\delta = 1$ in the sum entering in (4.25)

$$(4.29)\qquad H := -(e_{s-1}^T \otimes I_m) h P_{\tilde{A}}(\tilde{A} \otimes I_n)\{f_y\} P_{\tilde{A}}(F_z - F_{z,0})(\tilde{A}_1 \otimes h\Delta\zeta).$$

As a consequence of (4.28) we obtain

$$(4.30a)\quad H = h^2 \sum_{1 \le |\kappa| + |\nu| + |\tau| \le m-1} h^{|\kappa| + |\nu| + |\tau|} C_{\kappa\nu\tau} \cdot K_{\kappa\nu\tau}(\eta)\Delta\zeta + O(h^{m+2}\|\Delta\zeta\|)$$

where

$$(4.30b)\qquad C_{\kappa\nu\tau} = e_{s-1}^T \left(\prod_{i=1}^{\omega} \tilde{A}\tilde{C}^{\kappa_{1i}}\tilde{A}^{-1}\tilde{C}^{\nu_{1i}} \right) \tilde{A}\tilde{C}^{\tau_1} \times \left(\prod_{i=1}^{\omega} \tilde{A}\tilde{C}^{\kappa_{2i}}\tilde{A}^{-1}\tilde{C}^{\nu_{2i}} \right) \tilde{A}\tilde{C}^{\tau_2}\tilde{A}^{-1}\tilde{A}_1,$$

and the $K_{\kappa\nu\tau}$ are other expressions like the D_{ij}. Further, $\kappa_j = (\kappa_{j1}, \ldots, \kappa_{j\omega})$, $\nu_j = (\nu_{j1}, \ldots, \nu_{j\omega})$ (where $j = 1, 2$) are multi-indices in $\mathbb{N}^\omega$, and we also have $\kappa = (\kappa_1, \kappa_2)$, $|\kappa| := |\kappa_1| + |\kappa_2|$, $\nu = (\nu_1, \nu_2)$, $|\nu| := |\nu_1| + |\nu_2|$, $\tau = (\tau_1, \tau_2)$ with τ_2 strictly positive. The coefficients $C_{\kappa\nu\tau}$ are of the form (3.1) and by Theorem 3.1 they vanish by virtue of $|\kappa| + |\nu| + |\tau| + 1 \le r$. We thus get $H = O(h^{m+2}\|\Delta\zeta\|)$. All other remaining terms can be treated in a similar way, so that the statement (4.22a) results. ■

5. Local error and convergence.

Theorem 4.4 yields the main component for the convergence proof of RK methods with singular RK matrix A. The rest closely follows the proofs given in [4, Sections 4 & 5] and [5, Sections VI.7 & VI.8]. For convenience of the reader, we present here the final results and give only some indications for their proof. Details are omitted.

We consider one step of a RK method (2.1) with initial values $\eta = y(x)$, $\zeta = z(x)$ on the exact solution and we want to give estimates for the *local error*

$$(5.1)\qquad \delta y_h(x) = y_1 - y(x+h), \qquad \delta z_h(x) = z_1 - z(x+h).$$

THEOREM 5.1. *Assume that the RK coefficients satisfy the conditions $B(p)$, $C(q)$, and $D(r)$, and that the hypotheses $H1$, $H2$, and $H3$ hold. Then we have*

$$\delta y_h(x) = O(h^{\min(p,2q,q+r+1)+1}), \qquad \delta z_h(x) = O(h^q). \tag{5.2}$$

REMARKS.

1) If the function f of (1.1) is linear in z then we get

$$\delta y_h(x) = O(h^{\min(p,2q+1,q+r+1)+1}). \tag{5.2'}$$

2) $p \geq q$ follows from Remark 2) in Section 2.

The *proof* is omitted. The ideas and techniques are similar to those of [4, Lemma 4.3 & Theorem 5.9] and [5, Chapter VI, Lemma 7.4 & Theorem 8.10] which are devoted to the case of invertible RK matrix A. The local error of the y-component can be found by repeated application of simplifying assumptions to the order conditions. ■

THEOREM 5.2. *Consider the differential-algebraic system* (1.1) *of index* 2 *with consistent initial values and the RK method* (2.1). *In addition to the hypotheses of Theorem* 5.1, *suppose further that* $|R(\infty)| \leq 1$ *and* $q \geq 2$ *if* $R(\infty) = 1$. *Then for* $x_n - x_0 = nh \leq Const$, *the global error satisfies*

$$y_n - y(x_n) = \begin{cases} O(h^{\min(p,2q,q+r+1)}) & \text{if } -1 \leq R(\infty) < 1, \\ O(h^{\min(p,2q-1,q+r+1)}) & \text{if } R(\infty) = 1, \end{cases} \tag{5.3a}$$

$$z_n - z(x_n) = \begin{cases} O(h^q) & \text{if } -1 \leq R(\infty) < 1, \\ O(h^{q-1}) & \text{if } R(\infty) = 1. \end{cases} \tag{5.3b}$$

REMARKS.

1) If the function f of (1.1) is linear in z then we have

$$y_n - y(x_n) = \begin{cases} O(h^{\min(p,2q+1,q+r+1)}) & \text{if } -1 \leq R(\infty) < 1, \\ O(h^{\min(p,2q,q+r+1)}) & \text{if } R(\infty) = 1. \end{cases} \tag{5.3a'}$$

The first remark after Theorem 4.4 applies, therefore in the proof below the terms $O(h\,\|\Delta z_n\|^2)$ can be replaced by $O(h^3\,\|\Delta z_n\|^2)$, and $m = \min(q,r)$ if $-1 \leq R(\infty) < 1$ or $m = \min(q-1,r)$ if $R(\infty) = 1$.

2) The theorem remains valid in the case of variable stepsizes with $h = \max_i h_i$, except if $R(\infty) = -1$ the same results as for $R(\infty) = 1$ hold, because in the first part of the proof a perturbed asymptotic expansion of the global error does not exist.

OUTLINE OF THE PROOF. In a first step we can show that global convergence of order $\min(p, q+1)$ for the y-component and of order q (resp. $q-1$) for the z-component if $|R(\infty)| < 1$ (resp. if $|R(\infty)| = 1$) occurs (the second step can be applied with $m = 0$). For the z-component, if $R(\infty) = -1$, this order can be raised to q by considering a perturbed asymptotic expansion of the global error as described in [4, Theorem 4.8] by applying the ideas of [4, Theorem 4.9 & Theorem 3.1].

The second step is again similar to the proof of [5, Chapter VI, Theorem 7.5]. We denote two neighbouring RK solutions by $\{\tilde{y}_n, \tilde{z}_n\}$, $\{\hat{y}_n, \hat{z}_n\}$ and their difference by $\Delta y_n = \tilde{y}_n - \hat{y}_n$, $\Delta z_n = \tilde{z}_n - \hat{z}_n$. With the results of the previous step and by use of $H3$, Theorem 4.4 can be applied with $\delta = 0$ and $\theta = 0$, yielding

$$\Delta y_{n+1} = P_n \Delta y_n + O(h \|\Delta y_n\| + h^{m+2} \|\Delta z_n\| + h \|\Delta z_n\|^2), \tag{5.4a}$$

$$\Delta z_{n+1} = R(\infty)\Delta z_n + O(\|\Delta y_n\| + h \|\Delta z_n\|) \tag{5.4b}$$

where P_n is the projector (4.23) evaluated at $\hat{y}_n$, $\hat{z}_n$, and $m = \min(q-1, r)$ if $-1 \leq R(\infty) < 1$ or $m = \min(q-2, r)$ if $R(\infty) = 1$. By using the techniques of [4, Lemma 4.5], the estimates (5.4) give

$$\|\Delta y_n\| \leq C(\|P_0 \Delta y_0\| + h \|Q_0 \Delta y_0\| + h^{m+2} \|\Delta z_0\|). \tag{5.5}$$

■

The proof of the conjecture stated in [4, pp. 18, 46 & 47] and [5, p. 515] is now a direct consequence of the precedent theorem.

COROLLARY 5.3. *For the s-stage Lobatto IIIA method as applied to the index 2 system* (1.1), *the global error satisfies*

$$y_n - y(x_n) = O(h^{2s-2}), \qquad z_n - z(x_n) = \begin{cases} O(h^s) & \text{if } s \text{ even,} \\ O(h^{s-1}) & \text{if } s \text{ odd.} \end{cases} \tag{5.6}$$

If the stepsizes are not constant, we get

$$y_n - y(x_n) = O(h^{2s-2}), \qquad z_n - z(x_n) = O(h^{s-1}) \tag{5.7}$$

where $h = \max_i h_i$.

PROOF. The proof is obtained by putting $p = 2s-2$, $q = s$ and $r = s-2$ in (5.3). ■

6. Projected Runge-Kutta methods.

For a RK method satisfying $H1$ and $H2$, but which is not stiffly accurate, identical superconvergence results can be obtained if after every step the numerical solution is projected onto the manifold $g(y) = 0$. This projection is necessary, otherwise the method can not be applied: according to $H1$ the numerical values y_n have to satisfy $g(y_n) = 0$. The new class of projected RK methods has been recently introduced in [3] (see also [5, Sections VI.7 & VI.8]). A necessary and sufficient condition in order to extend the results of the paper to these methods is to have

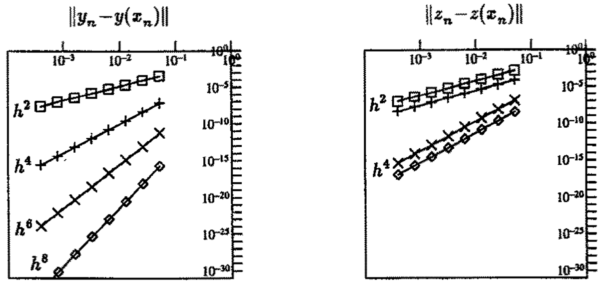

Fig. 7.1. Global errors of the Lobatto IIIA methods with constant stepsizes ($s = 2$: □; 3: $+$; 4: $\times$; 5: ◇).

$$(6.1) \qquad \tilde{b}^T \tilde{A}^{-1} \tilde{C} \mathbb{1}_{s-1} = 1.$$

This implies that $R(\infty)$ remains finite and that z_1 in (2.1a) can be well-defined.

7. Numerical experiments.

To show the relevance of our theoretical results, we have applied the Lobatto IIIA methods ($s = 2, 3, 4, 5$) to the following index 2 problem

$$(7.1) \qquad y_1' = y_1 y_2^2 z^2, \qquad y_2' = y_1^2 y_2^2 - 3 y_2^2 z, \qquad 0 = y_1^2 y_2 - 1,$$

with consistent initial values $y_0 = (1, 1)$, $z_0 = 1$. The exact solution is given by

$$(7.2) \qquad y_1(x) = e^x, \qquad y_2(x) = e^{-2x}, \qquad z(x) = e^{2x}.$$

In Fig. 7.1 we have plotted the global errors at $x_{\text{end}} = 0.1$ as functions of the stepsize h. Logarithmic scales have been used, so that 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. The predicted order of convergence can be observed.

In Fig. 7.2 we have plotted the global errors at $x_{\text{end}} = 0.1$ as functions of h (the stepsizes have been chosen alternatively as $h/3$ and $2h/3$).

Acknowledgements.

I am grateful to E. Hairer for his critical remarks and his suggestions. I wish to thank W. Hundsdorfer for his thorough reading of the manuscript, and S. F. Bernatchez for her careful correction of the English part of this article.

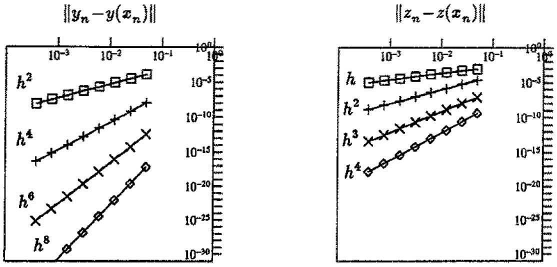

Fig. 7.2. Global errors of the Lobatto IIIA methods with non-constant stepsizes ($s = 2$: □; 3: $+$; 4: $\times$; 5: ◇).

REFERENCES

1. U. M. Ascher & L. R. Petzold: *Projected implicit Runge-Kutta methods for differential-algebraic equations.* SIAM J. Numer. Anal., Vol. 28, pp. 1097–1120, (1991).
2. U. M. Ascher: *Two Families of Symmetric Difference Schemes for Singular Perturbation Problems*, in *Numerical Boundary Values ODEs.* Proceedings of an International Workshop, U. M. Ascher & R. D. Russel editors, Vancouver, Canada, July 10–13, 1984, Progress in Scientific Computing, Vol. 5, pp. 173–191, Birkhäuser, (1985).
3. K. Brenan & L. R. Petzold: *The numerical solution of higher index differential/algebraic equations by implicit Runge-Kutta methods.* SIAM J. Numer. Anal., Vol. 26, pp. 976–996, (1989).
4. E. Hairer, Ch. Lubich & M. Roche: *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods.* Lecture Notes in Mathematics, Vol. 1409, Springer-Verlag, (1989).
5. E. Hairer & G. Wanner: *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems.* Computational Mathematics, Vol. 14, Springer-Verlag, Berlin, (1991).