Global Error Estimation for Index 1 and 2 DAE's

Claude-Pierre Jeannerod, Josselin Visconti

► To cite this version:

Claude-Pierre Jeannerod, Josselin Visconti. Global Error Estimation for Index 1 and 2 DAE's. Numerical Algorithms, 1998, 19 (1/4), pp.111-125. 10.1023/A:1019110608075 . hal-03420250

HAL Id: hal-03420250

<https://hal.science/hal-03420250>

Submitted on 9 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Global Error Estimation for Index 1 and 2 DAE's

C.P. Jeannerod and J.Visconti

LMC/IMAG, 51 rue des mathématiques, Grenoble, France

E-mail: Claude-Pierre.Jeannerod@imag.fr, Josselin.Visconti@imag.fr

In this paper, we consider the extension of three classical ODE's estimation techniques (Richardson extrapolation, Zadunaisky's technique and Solving for the Correction) to DAE's. Their convergence analysis is carried out for semi-explicit index 1 DAE's solved by a wide set of Runge-Kutta methods. Experimentation of the estimation techniques with RADAU5 is also presented: their behaviour for index 1 and 2 problems, and for variable step size integration is investigated.

Keywords: Global error estimation, Differential Algebraic Equations, Implicit Runge-Kutta methods, Richardson extrapolation, Zadunaisky's technique, Solving for the Correction.

1. Introduction

Several techniques for estimating the global error in the numerical solution of ODE's

$$y' = f(y), \quad y(0) = y_0, \quad (1)$$

are described in [13]. Stetter's method [15] was extended to stiff problems in [12] and to index 1 and 2 DAE's in [1]. It consists in building specific Runge-Kutta methods with built-in global error estimation. In this paper, we look into techniques of a different type, based on a second integration rather than on the construction of specific methods. Three classical techniques experimented for ODE's in [2] are considered: Richardson extrapolation, Zadunaisky's technique and Solving for the Correction. To our knowledge, they were never extended to DAE's solved by Runge-Kutta methods. Zadunaisky's technique was only considered in [14] for DAE's solved by the linearly implicit Euler method. We give convergence results for the three estimates applied to index 1 problems

$$\begin{aligned} y' &= f(y, z), & y(0) &= y_0, \\ 0 &= g(y, z), & z(0) &= z_0, \end{aligned} \quad (2)$$

with g_z non singular, solved by Implicit Runge-Kutta methods: their extensions are presented in section 2 and convergence results for Zadunaisky's technique and Solving for the Correction are proved in section 3. Among the solvers for initial value problems in DAE's [4,8], we concentrate on RADAU5 [9], based on the 3

stage RADAU IIA method. The three estimation techniques have been implemented in RADAU5: the outcome of our numerical experimentation is described in section 4. We compare the performances of the estimates and give an insight into their behaviour for index 2 problems.

2. Global error estimation for semi-explicit index 1 DAE's

In this section, Richardson extrapolation, Zadunaisky's and Solving for the Correction techniques are first presented for (1) and extended to (2). Let y be the exact solution of (1) and $\{y_n\}$ be its numerical approximation on a constant step size grid $\{t_n = nh\}$ by a Runge-Kutta method of order $p \geq 1$. The global error $E_n = y_n - y(t_n)$ is well known to have an asymptotic h -expansion [10]

$$y_n - y(t_n) = h^p a_p(t_n) + \dots + h^N a_N(t_n) + \mathcal{O}(h^{N+1}) \quad (3)$$

with smooth a_k 's. An estimate $\hat{E}_n$ of E_n is said to be *consistent of relative order* $s > 0$ when

$$\hat{E}_n = E_n + \mathcal{O}(h^{p+s}).$$

The three estimation techniques under consideration are based on this special form of the h -expansion (3), where the coefficients are functions of time only. A similar property is needed for (2).

For the numerical solution of (2), we restrict our attention to Runge-Kutta methods applied with the so-called *direct approach*, i.e.

$$\begin{aligned} y_{n+1} &= y_n + h \sum_{i=1}^s b_i Y'_{ni}, & z_{n+1} &= z_n + h \sum_{i=1}^s b_i Z'_{ni}, \\ Y_{ni} &= y_n + h \sum_{j=1}^s a_{ij} Y'_{nj}, & Z_{ni} &= z_n + h \sum_{j=1}^s a_{ij} Z'_{nj}, \\ Y'_{ni} &= f(t_n + c_i h, Y_{ni}, Z_{ni}), \\ 0 &= g(t_n + c_i h, Y_{ni}, Z_{ni}), \end{aligned}$$

and consistent initial conditions. With the usual settings [9], we also assume that

Assumption 1. The implicit Runge-Kutta method is of classical order p , satisfies $C(q)$ with $p \geq q + 1$, has an invertible coefficient matrix $\mathcal{A}$ and a stability function R satisfying $|R(\infty)| < 1$.

When (2) is solved by such a Runge-Kutta method with constant step size h , the global errors in y and z are also proved to have asymptotic h -expansions ([9], p. 27). However, their valuations may be different (p for the y -component and $r \leq p$ for the z -component), and the h -expansion in z may include perturbations.

However, for all methods with $|R(\infty)| < 1$, Theorem 3.2 in [9] states that the perturbations vanish or exponentially decay as n tends to infinity so that the following lemma is a trivial consequence.

Lemma 2. Suppose that (2) is solved on a grid $\{t_n = nh\}$ by a Runge-Kutta method satisfying Assumption 1. For any integer N , there exists τ such that, for $t_n > \tau$, the asymptotic h -expansions of the global errors in y and z are both of type (3), that is to say unperturbed up to order N .

In practice, this boundary layer only affects the first few steps so that it is numerically insignificant. The above definition for the consistency order of a global error estimate is also valid for the y and z components of (2): an estimate is of relative order s if it coincides with the first s terms in the h -expansion of type (3). These terms can be derived by perturbing either the step size (Richardson extrapolation) or the problem (Zadunaisky's technique and Solving for the Correction).

2.1. Richardson extrapolation (RS)

This technique consists in carrying out two parallel integrations with step sizes h and $h/2$ leading respectively to the numerical values y_n and y_{2n}^* at time t_n . From (3), it is straightforward that the estimate

$$E_n^{RS} = \frac{y_n - y_{2n}^*}{1 - 2^{-p}} \quad (4)$$

coincides with the global error up to order $p+1$. Hence, it is a (relative) first order error estimate for (1). The extension to (2) is straightforward since, according to lemma 2, the first terms in the asymptotic h -expansions are unperturbed for n sufficiently large. Hence, formula (4) with convenient p (IRK convergence order for the variable under consideration) still provides first order estimates of the global errors in y and z , for n sufficiently large.

2.2. Zadunaisky's technique (ZD)

The idea is to build a regularly perturbed problem

$$\hat{y}' = f(\hat{y}) + d_h(t), \quad \hat{y}(0) = y_0 \quad (5)$$

whose solution is known and whose perturbation $d_h(t)$ is small so that the global errors in solving (1) and (5) are close. Zadunaisky [17] proposed to consider P_h defined by $P_h(t) = P_j(t)$ for $t \in [t_{(j-1)m}, t_{jm}]$ with P_j the Lagrange interpolation polynomial of degree $m \geq 1$ of the numerical values $\{y_n, n = (j-1)m \dots jm\}$. In other words, P_h is a piecewise polynomial interpolating the numerical solution $\{y_n\}$ up to order m . Then, setting $d_h(t) = P_h'(t) - f(P_h(t))$, the solution of

Table 1
Maximum relative orders of RS, ZD, SC (n sufficiently large)

Method	Order for ODE's	Order for semi-explicit index 1 DAE's	
		y -component	z -component
IRK method	p	p	r
RS	1	1	1
ZD ($m = 2p$)	p	r	r
SC ($m = 2p$)	p	p	r

problem (5) is P_h . Let $\hat{y}_n$ be the numerical solution of (5) by the same Runge-Kutta method. Zadunaisky's estimate is

$$E_n^{ZD} = \hat{y}_n - P_h(t_n) = \hat{y}_n - y_n. \quad (6)$$

Zadunaisky's technique can naturally be extended to (2) considering again the piecewise interpolation polynomials P_h and Q_h of the numerical solutions $\{y_n\}$ and $\{z_n\}$ up to order m . Then, the associated perturbed problem is

$$\begin{aligned} \hat{y}' &= f(\hat{y}, \hat{z}) + P_h'(t) - f(P_h(t), Q_h(t)), & \hat{y}(0) &= y_0, \\ 0 &= g(\hat{y}, \hat{z}) - g(P_h(t), Q_h(t)), & \hat{z}(0) &= z_0, \end{aligned} \quad (7)$$

and denoting by $\{\hat{y}_n\}$ and $\{\hat{z}_n\}$ its numerical solution with the same Runge-Kutta method, Zadunaisky's estimates are

$$(Ey)_n^{ZD} = \hat{y}_n - P_h(t_n) = \hat{y}_n - y_n, \quad (Ez)_n^{ZD} = \hat{z}_n - Q_h(t_n) = \hat{z}_n - z_n. \quad (8)$$

Consistency orders of (6) and (8) are given in table 1 and proved in section 3.

2.3. Solving for the Correction (SC)

This technique from [13] is based on the same interpolation polynomial P_h as in the previous section. Now, we consider the function $\varepsilon_h = P_h - y$ coinciding with the global error up to order m . This quantity can be computed by solving the differential equation

$$\varepsilon_h' = P_h'(t) - f(P_h(t) - \varepsilon_h), \quad \varepsilon_h(0) = 0. \quad (9)$$

Its numerical solution $\{\varepsilon_n\}$ with the same Runge-Kutta method provides a consistent estimate of the global error

$$E_n^{SC} = \varepsilon_n. \quad (10)$$

The extension to semi-explicit index 1 DAE's is again straightforward. We keep $\varepsilon_h = P_h - y$ and introduce $\zeta_h = Q_h - z$, so that the correction problem associated with (2) is

$$\begin{aligned} \varepsilon_h' &= P_h'(t) - f(P_h(t) - \varepsilon_h, Q_h(t) - \zeta_h), & \varepsilon_h(0) &= 0, \\ 0 &= g(P_h(t) - \varepsilon_h, Q_h(t) - \zeta_h), & \zeta_h(0) &= 0, \end{aligned} \quad (11)$$

and the global error estimates become

$$(Ey)_n^{SC} = \varepsilon_n, \quad (Ez)_n^{SC} = \zeta_n. \quad (12)$$

Convergence orders of (10) and (12) are given in table 1 and proved in section 3.

3. Convergence of ZD and SC for semi-explicit index 1 DAE's

After some preliminary results, we prove the consistency orders shown in table 1 for Zadunaisky's technique and Solving for the Correction applied to (2). Defect correction of type (7) based on the linearly implicit Euler method is studied in [14]: the convergence analysis is carried out directly on (2). For Runge-Kutta methods, it is usual to consider the ODE equivalent to (2), i.e.

$$y' = f(y, G(y)), \quad \text{with } G \text{ defined by } g(y, G(y)) = 0, \quad (13)$$

because the y -component of the numerical solution of (2) can be interpreted as the numerical result for (13) by the same method (see [9], p. 25). Hence, in order to study Zadunaisky's technique and Solving for the Correction for the y -component, we will consider the ODE's equivalent to problems (7) and (11), and compare them to (13). This leads to a higher consistency order for Solving for the Correction than for Zadunaisky's technique. Results for the z -component will then be derived from those of the y -component and are identical for both techniques.

3.1. Preliminary results

The following result [3,6,7] for ODE's was first proven in [7].

Theorem 3. Let $y(t)$ and $\hat{y}(t)$ be the respective solutions of (1) and (5), and y_n and $\hat{y}_n$ be their respective numerical solutions by a Runge-Kutta method of order $p \geq 1$. Assume that the perturbation d_h satisfies for $m \geq r \geq 0$,

$$\forall k \geq 0 \quad d_h^{(k)}(t) = \mathcal{O}(h^{\max(0, \min(r, m-k))}), \quad (14)$$

for all $t \geq 0$. Then, the global errors $E_n = y_n - y(t_n)$ and $\hat{E}_n = \hat{y}_n - \hat{y}(t_n)$ satisfy, for all n ,

$$\hat{E}_n - E_n = \mathcal{O}(h^{\min(m, p+r)}).$$

Remark 4. As a consequence, the order of Zadunaisky's technique for ODE's is given by $E_n^{ZD} - E_n = \mathcal{O}(h^{\min(m, 2p)})$ because the defect $d_h(t) = P_h'(t) - f(P_h(t))$ satisfies (14) with $r = p$ (see [6]). Similar considerations [6] also lead to $E_n^{SC} - E_n = \mathcal{O}(h^{\min(m, 2p)})$.

For the sequel, we need an extension to regularly perturbed problems of type

$$\hat{y}' = f(\hat{y}) + d_h(t, \hat{y}), \quad \hat{y}(0) = y_0. \quad (15)$$

Theorem 5. Let $y(t)$ and $\hat{y}(t)$ be the respective solutions of (1) and (15), and y_n and $\hat{y}_n$ be their respective numerical solutions by a Runge-Kutta method with order $p \geq 1$. Assume that the perturbation d_h satisfies, for some real τ and $m \geq r \geq 0$,

$$\forall k, l \geq 0 \quad \frac{\partial^{k+l} d_h}{\partial t^k \partial y^l}(t, y) = \mathcal{O}(h^{\max(0, \min(r, m-k))}) \quad (16)$$

for $t > \tau$ and any y . Then, the global errors $E_n = y_n - y(t_n)$ and $\hat{E}_n = \hat{y}_n - \hat{y}(t_n)$ satisfy, for $t_n > \tau$,

$$\hat{E}_n - E_n = \mathcal{O}(h^{\min(m, p+r)}).$$

The proof is detailed in [16] and follows the line of that for Theorem 3 [3,7]. The key point is to check that the elementary differentials associated to problems (1) and (15) are close up to a convenient order in h . We will also need the following lemma in the sequel

Lemma 6. Let $x_1, x_2, \hat{x}_1, \hat{x}_2$ be some reals and $F, \hat{F}$ be some smooth functions satisfying $\hat{F}^{(k)}(\hat{x}_1) - F^{(k)}(x_1) = \mathcal{O}(h^r)$, $\forall k \geq 0$. One has

$$\Delta \hat{F} - \Delta F = \mathcal{O}(\Delta \hat{x} - \Delta x) + \mathcal{O}(h^r \Delta x)$$

with $\Delta F = F(x_2) - F(x_1)$, $\Delta \hat{F} = \hat{F}(\hat{x}_2) - \hat{F}(\hat{x}_1)$ and $\Delta x = x_2 - x_1$, $\Delta \hat{x} = \hat{x}_2 - \hat{x}_1$.

This lemma comes from a basic argument on Taylor expansions [16].

3.2. Zadunaisky's technique

Theorem 7. Assume that (2) is solved by a Runge-Kutta method satisfying Assumption 1 with consistent initial conditions. Let p and r be its convergence orders for the y and z components. Zadunaisky's technique, as described in section 2.2 with $\deg(P_h) = \deg(Q_h) = m \geq r \geq 0$, satisfies for sufficiently large n ,

$$(Ey)_n^{ZD} - (Ey)_n = \mathcal{O}(h^{\min(m, p+r)}) \quad \text{with } (Ey)_n = y_n - y(t_n), \quad (17a)$$

$$(Ez)_n^{ZD} - (Ez)_n = \mathcal{O}(h^{\min(m, 2r)}) \quad \text{with } (Ez)_n = z_n - z(t_n). \quad (17b)$$

Proof. Components y and z are considered respectively in parts (a) and (b).

(a) The ODE equivalent to (7) is

$$\hat{y}' = f(\hat{y}, \hat{G}(t, \hat{y})) + P_h'(t) - f(P_h(t), Q_h(t)), \quad \hat{y}(0) = y_0, \quad (18)$$

with $\widehat{G}$ defined by $g(\widehat{y}, \widehat{G}(t, \widehat{y})) - g(P_h(t), Q_h(t)) = 0$. Note that (18) is not Zadunaisky's perturbed problem for (13). Nevertheless, it is a regular perturbation of (13) of the form $\widehat{y}' = f(\widehat{y}, G(\widehat{y})) + d_h(t, \widehat{y})$ with

$$d_h(t, y) = f(y, \widehat{G}(t, y)) - f(y, G(y)) + P_h'(t) - f(P_h(t), Q_h(t)). \quad (19)$$

The idea is to prove that d_h satisfies property (16) for sufficiently large t , so that Theorem 5 leads to the desired result for the y -component. This can be done by the following rather technical analysis.

Properties of Lagrange interpolation ensure that for $t > \tau$,

$$P_h^{(k)}(t) - y^{(k)}(t) = \mathcal{O}(h^{\max(0, \min(p, m+1-k))}), \quad (20a)$$

$$Q_h^{(k)}(t) - z^{(k)}(t) = \mathcal{O}(h^{\max(0, \min(r, m+1-k))}). \quad (20b)$$

Let $\gamma_h(t) = P_h'(t) - f(P_h(t), Q_h(t))$, and $\delta_h(t) = -g(P_h(t), Q_h(t))$ be the two components of the defect in (7). A similar analysis to that in [7] leads to

$$\gamma_h^{(k)}(t) = \mathcal{O}(h^{\max(0, \min(r, m-k))}), \text{ and } \delta_h^{(k)}(t) = \mathcal{O}(h^{\max(0, \min(r, m+1-k))}). \quad (21)$$

Now, let us expand d_h in terms of γ_h and δ_h . Denoting $\Delta G(t, y) = \widehat{G}(t, y) - G(y)$, the expansion of (19) in terms of $\Delta G(t, y)$ reads

$$d_h(t, y) = \gamma_h(t) + d_1(y) \Delta G(t, y) + d_2(y) \Delta G^2(t, y) + \dots + \mathcal{O}(\Delta G^N(t, y)), \quad (22)$$

with smooth d_i 's. By definition, $g(y, G(y)) = 0$ and $g(y, \widehat{G}(t, y)) + \delta_h(t) = 0$ so that $\delta_h(t) = g(y, G(y)) - g(y, \widehat{G}(t, y))$. Hence, we also get an expansion

$$\delta_h(t) = -g_z(y, G(y)) \Delta G(t, y) + e_2(y) \Delta G^2(t, y) + \dots + \mathcal{O}(\Delta G^N(t, y)), \quad (23)$$

with smooth e_i 's and g_z invertible. Inverting (23) and inserting it into (22) then leads to

$$d_h(t, y) = \gamma_h(t) + f_1(y) \delta_h(t) + f_2(y) \delta_h^2(t) + \dots + \mathcal{O}(\delta_h^N(t)), \quad (24)$$

with smooth f_i 's. Finally, considering (24) and its partial derivatives together with (21) proves that d_h satisfies property (16) for $t > \tau$.

- (b) The result for the z -component is derived from that for the y -component, following the proof of Theorem 3.1 in [9].

For stiffly accurate methods, the algebraic order is $r = p$ and constraints are preserved, so that $z_n = G(y_n)$, $\widehat{z}_n = \widehat{G}(t_n, \widehat{y}_n)$. One can check

$$\frac{\partial^k \widehat{G}}{\partial y^k}(t_n, P_h(t_n)) - \frac{\partial^k G}{\partial y^k}(y(t_n)) = \mathcal{O}(h^p), \quad \forall k \geq 0$$

so that (17b) is a direct consequence of (17a) and Lemma 6.

For other methods with $|R(\infty)| < 1$, the algebraic order is $r = q + 1$ (see [9]), and a further analysis, involving the intermediate RK stages, must be carried out. It involves three main stages:

- First, prove

$$\Delta \hat{Y}_n - \Delta Y_n = \mathcal{O}(h^{\min(m, 2r)}), \quad (25)$$

with

$$\begin{aligned} Y(t_n) &= [y(t_n + c_1 h), \dots, y(t_n + c_s h)]^T, & \Delta Y_n &= Y_n - Y(t_n), \\ \hat{Y}(t_n) &= [P_h(t_n + c_1 h), \dots, P_h(t_n + c_s h)]^T, & \Delta \hat{Y}_n &= \hat{Y}_n - \hat{Y}(t_n). \end{aligned}$$

As the assumption C(q) holds, one has

$$\Delta Y_{ni} = (Ey)_n + h \sum_{j=1}^s a_{ij} [F(Y_{nj}) - F(Y_j(t_n))] - \sum_{k=q+1}^{\infty} D_{ki} h^k y^{(k)}(t_n),$$

with $F(y) = f(y, G(y))$ and $D_k = [D_{k1}, \dots, D_{ks}]^T$ a vector of reals depending on the Runge-Kutta coefficients. Similarly, for the perturbed problem,

$$\begin{aligned} \Delta \hat{Y}_{ni} &= (Ey)_n^{ZD} + h \sum_{j=1}^s a_{ij} [\hat{F}(t_n + c_j h, \hat{Y}_{nj}) - \hat{F}(t_n + c_j h, \hat{Y}_j(t_n))] \\ &\quad - \sum_{k=q+1}^{\infty} D_{ki} h^k P_h^{(k)}(t_n), \end{aligned}$$

with $\hat{F}(t, y) = F(y) + d_h(t, y)$ and d_h defined in (19). Moreover, as a consequence of property (16), one has

$$\frac{\partial \hat{F}}{\partial y^k}(t_n, P_h(t_n)) - \frac{\partial F}{\partial y^k}(y(t_n)) = \mathcal{O}(h^r), \quad \forall k \geq 0$$

so that Lemma 6 leads to

$$\begin{aligned} &[\hat{F}(t_n + c_j h, \hat{Y}_{nj}) - \hat{F}(t_n + c_j h, \hat{Y}_j(t_n))] - [F(Y_{nj}) - F(Y_j(t_n))] \\ &= \mathcal{O}(\Delta \hat{Y}_{nj} - \Delta Y_{nj}) + \mathcal{O}(h^r \Delta Y_{nj}) \end{aligned}$$

Relation (20a) also proves

$$\begin{aligned} \sum_{k=q+1}^{\infty} D_k h^k P_h^{(k)}(t) - \sum_{k=q+1}^{\infty} D_k h^k y^{(k)}(t) &= \mathcal{O}(h^{\min(m+1, p+q+1)}) \\ &= \mathcal{O}(h^{\min(m, 2r)}) \quad (r \leq p) \end{aligned}$$

Hence, we have

$$\begin{aligned} \Delta \hat{Y}_n - \Delta Y_n &= [(Ey)_n^{ZD} - (Ey)_n] \mathbb{1} + \mathcal{O}(h(\Delta \hat{Y}_n - \Delta Y_n)) \\ &\quad + \mathcal{O}(h^{r+1} \Delta Y_n) + \mathcal{O}(h^{\min(m, 2r)}). \end{aligned}$$

Finally, $\Delta Y_n = \mathcal{O}(h^{q+1}) = \mathcal{O}(h^r)$ and (17a) give the desired result (25).

- Then, derive

$$\Delta \hat{Z}_n - \Delta Z_n = \mathcal{O}(h^{\min(m, 2r)}), \quad (26)$$

with

$$\begin{aligned} Z(t_n) &= [z(t_n + c_1 h), \dots, z(t_n + c_s h)]^T, & \Delta Z_n &= Z_n - Z(t_n), \\ \hat{Z}(t_n) &= [Q_h(t_n + c_1 h), \dots, Q_h(t_n + c_s h)]^T, & \Delta \hat{Z}_n &= \hat{Z}_n - \hat{Z}(t_n). \end{aligned}$$

One has $Z_n = G(Y_n)$, $\hat{Z}_n = \hat{G}(t_n, \hat{Y}_n)$ and

$$\frac{\partial \hat{G}}{\partial y^k}(t_n, P_h(t_n)) - \frac{\partial G}{\partial y^k}(y(t_n)) = \mathcal{O}(h^r), \quad \forall k \geq 0$$

so that relation (26) is a direct consequence of (25) by Lemma 6 (same argument as above for stiffly accurate methods).

- Now, prove (17b). As the method is of order p and $C(q)$ holds, one has [9]

$$\begin{aligned} (Ez)_{n+1} &= R(\infty)(Ez)_n + b^T \mathcal{A}^{-1} \Delta Z_n \\ &\quad - \sum_{k=p+1}^{\infty} d_k h^k z^{(k)}(t_n) + \sum_{k=q+1}^{\infty} b^T \mathcal{A}^{-1} D_k h^k z^{(k)}(t_n) \end{aligned}$$

with d_k some reals involving the Runge-Kutta coefficients. And, similarly for the perturbed problem,

$$\begin{aligned} (Ez)_{n+1}^{ZD} &= R(\infty)(Ez)_n^{ZD} + b^T \mathcal{A}^{-1} \Delta \hat{Z}_n \\ &\quad - \sum_{k=p+1}^{\infty} d_k h^k Q_h^{(k)}(t_n) + \sum_{k=q+1}^{\infty} b^T \mathcal{A}^{-1} D_k h^k Q_h^{(k)}(t_n) \end{aligned}$$

With the same technique as above for the Taylor expansions, and using (26), one gets

$$(Ez)_{n+1}^{ZD} - (Ez)_{n+1} = R(\infty) \left[(Ez)_n^{ZD} - (Ez)_n \right] + \mathcal{O}(h^{\min(m, 2r)})$$

Since $|R(\infty)| < 1$, this leads to the desired result (17b). $\square$

Remark 8. Consider the special class of DAE's with f and g defined by $f(y, z) = F(y) + \alpha z$ and $g(y, z) = G(y) - z$. Equation (18) then reads

$$\hat{y}' = F(\hat{y}) + \alpha G(\hat{y}) - F(P_h) - \alpha G(P_h),$$

which is exactly Zadunaisky's problem for (13). Hence, for this class of DAE's, (17a) becomes $(Ey)_n^{ZD} - (Ey)_n = \mathcal{O}(h^{\min(m, 2p)})$ (ODE's convergence order). However, (17b) remains unchanged. This class obviously includes linear DAE's.

3.3. Solving for the Correction

Theorem 9. Under the assumptions of Theorem 7, Solving for the Correction, as described in section 2.2 with $\deg(P_h) = \deg(Q_h) = m \geq r \geq 0$, satisfies for sufficiently large n ,

$$(Ey)_n^{SC} - (Ey)_n = \mathcal{O}(h^{\min(m, 2p)}) \quad \text{with } (Ey)_n = y_n - y(t_n), \quad (27a)$$

$$(Ez)_n^{SC} - (Ez)_n = \mathcal{O}(h^{\min(m, 2r)}) \quad \text{with } (Ez)_n = z_n - z(t_n). \quad (27b)$$

Proof. y and z are considered respectively in parts (a) and (b).

(a) The ODE equivalent to (11) is

$$\varepsilon'_h = P'_h(t) - f(P_h(t) - \varepsilon_h, G(P_h(t) - \varepsilon_h)) \quad (28)$$

with G defined by $f(y, G(y)) = 0$, so that it is exactly the correction problem associated to (13). Hence, the convergence result for the y -component is the same as for ODE's (see Remark 4).

(b) The convergence results for the z -component can be derived from those of the y -component with the same technique as for (b) of Theorem 7. The proof is detailed in [16]. □

3.4. Example

Let us illustrate the theoretical results with the nonlinear index 1 system

$$\begin{aligned} y' &= (z + e^t)^2 + y, & y(0) &= -2, \\ 0 &= y - z - e^t, & z(0) &= -1, \end{aligned}$$

solved by the two stages RADAU IA method ($p = 3, r = 2$) with $m = 6$. This system is nonlinear but simple enough to allow us to carry out the integration and the global error estimation with Maple, keeping h as a formal parameter. This leads to the following h -expansions after twelve steps ($n = 12$),

$$\begin{aligned} (Ey)_n^{ZD} - (Ey)_n &\sim -\frac{2009}{26244}h^5, & (Ez)_n^{ZD} - (Ez)_n &\sim -\frac{1}{36}h^4, \\ (Ey)_n^{SC} - (Ey)_n &\sim -\frac{1150604642749}{15496819560}h^6, & (Ez)_n^{SC} - (Ez)_n &\sim -\frac{1}{36}h^4. \end{aligned}$$

These are precisely the results predicted by Theorems 7 and 9: the order of SC is higher than those of ZD for the differential variable. For a stiffly accurate method, $p = r$, so that ZD and SC would lead to the same convergence order for both components.

4. Numerical experiments with RADAU5

In this section, the results of our numerical experimentation of RS, ZD and SC with the code RADAU5 are presented. Among the variety of problems tested, we consider the following set of index 1 and 2 test problems.

- P1** Classical index 1 Pendulum ([9], p. 9): dimension 5 system with four differential variables p, q, u, v and one algebraic variable λ .
- TTA** Two Transistor Amplifier ([9], p. 108): index 1 problem of dimension 8. We focus on the algebraic variable U_8 .
- NL**
$$\begin{cases} \dot{x} = xy/z & x(0) = 1 \\ \dot{y} = 2z & y(0) = 1 \\ 0 = y - x^2 & z(0) = 1 \end{cases}$$
- MK**
$$\begin{cases} \dot{x} = -102x + 100y^2 & x(0) = 1 \\ \dot{y} = -e^{1-z^2} & y(0) = 1 \\ 0 = x - y(1 + y) + \frac{x}{y} & z(0) = 1 \end{cases}$$
- P2** Stabilized index 2 Pendulum ([9], p. 9): dimension 6 system with four differential variables p, q, u, v and two algebraic variables λ, μ .
- DPC** Discharge Pressure Control ([9], p. 116): index 2 problem of dimension 7. We focus on the algebraic variable m .
- RM** Ring Modulator ([9], p. 112, $C_S = 0$): index 2 system of dimension 15. We focus on the algebraic variable U_3 .

The basic Nonlinear system NL and the Modified Kaps problem [5] MK are two nonlinear index 2 problems: x, y are the differential variables and z is the algebraic variable. The solution of NL is (e^t, e^{2t}, e^{2t}) , and the solution of MK is $(e^{-2t}, e^{-t}, \sqrt{1+t})$. The selected problems are mainly index 2 because we want to give an insight into the behaviour of the estimation techniques for index 2 problems. In this paper, convergence results are only given for index 1 problems of type (2), but the situation is also promising for semi-explicit index 2 problems since even then, perturbed asymptotic expansions are proved to exist [9]. Moreover, most of our numerical observations were common to index 1 and 2 problems so that our set of test problems is representative.

The implementation of the three usual estimates is quickly described in section 4.1. All numerical experiments were carried out with double precision on a Sun SPARK station. In order to measure the efficiency of an estimate $\hat{E}_n$ of the global error E_n , we use the criterion

$$\mathcal{E} = -\text{Log} \left| \frac{\hat{E}_n - E_n}{E_n} \right|,$$

that quantifies the number of correct digits in the estimation. For most examples under consideration, no analytic solution is available, so a reference solution is computed with very stringent tolerance. In the sequel, we shall also call *average efficiency* the average of the efficiencies over a wide set of grid points. The

three estimation techniques efficiencies are investigated with constant step size in section 4.2 and with variable step size in section 4.3.

4.1. Implementation in RADAU5

The code RADAU5 is designed for problems up to index 3 of type

$$By' = f(t, y), \quad y(t_0) = y_0, \quad (29)$$

where B is a constant, possibly singular square matrix. The index 1 problem (2) is already in this particular form. In the sequel, the estimation techniques are also experimented when the index of (29) is 2. The following implementation covers both cases.

(RS) For index 1 problems, formula (4) is used with $p = 5$ (convergence order of RADAU5) for all variables. For index 2 problems, it is natural to set $p = 3$ (convergence order of RADAU5) for the algebraic variables.

(ZD) We consider the usual piecewise interpolation polynomial P_h of the numerical solution $\{y_n\}$ of (29), and solve the perturbed problem

$$B\hat{y}' = f(t, \hat{y}) + BP_h'(t) - f(t, P_h(t)), \quad \hat{y}(t_0) = y_0, \quad (30)$$

For problem (2), the perturbed problem (30) coincides with (7).

(SC) We solve the correction problem

$$B\varepsilon_h' = BP_h'(t) - f(t, P_h(t) - \varepsilon_h), \quad \varepsilon_h(t_0) = 0. \quad (31)$$

For problem (2), problem (31) also coincides with (11).

4.2. Experiments with constant step size

Experiments related in this subsection were carried out forcing constant step size in RADAU5. Our aim is to emphasise that ZD and SC are still experimentally consistent for index 2 problems and to compare their efficiencies with those of RS with respect to step size.

Test 1. The index 1 and 2 test problems are integrated on convenient intervals: average (and min-max) efficiencies of ZD and SC are given in table 2.¹ The variables under consideration are algebraic (z for NL and MK, λ for P1 and P2, U_8 for TTA, m for DPC, and U_3 for RM). One can check that, in most cases, ZD and SC provide good estimations for index 1 and 2 problems. Results are similar for other variables.

¹ For problem TTA, the global error presents a series of peaks and in-between zones where the global error is much smaller. In table 2, efficiencies for TTA are considered over zones where the global error is greater than 10^{-11} .

Table 2
Efficiencies of ZD and SC with constant step size h

Pb	Interval	h	ZD		SC	
			$\mathcal{E}_{av}$	$[\mathcal{E}_{min}, \mathcal{E}_{max}]$	$\mathcal{E}_{av}$	$[\mathcal{E}_{min}, \mathcal{E}_{max}]$
P1	$[0, 10]$	2.10^{-2}	3.33	$[0.32, 5.75]$	3.31	$[0.35, 5.74]$
TTA	$[0, 0.2]$	5.10^{-5}	1.85	$[-1.18, 4.24]$	1.86	$[-1.18, 4.61]$
NL	$[0, 5]$	10^{-1}	4.19	$[4.07, 5.17]$	6.34	$[4.18, 8.32]$
MK	$[0, 10]$	2.10^{-1}	2.57	$[0.87, 4.53]$	3.01	$[0.86, 5.04]$
P2	$[0, 10]$	2.10^{-2}	3.57	$[1.19, 5.54]$	3.66	$[1.19, 6.77]$
DPC	$[0, 20]$	2.10^{-1}	5.12	$[0.84, 10.15]$	5.27	$[0.89, 10.15]$
RM	$[0, 5 \cdot 10^{-5}]$	5.10^{-8}	2.99	$[1.19, 5.38]$	2.94	$[1.28, 5.35]$

Test 2. The efficiencies of ZD and SC with respect to step size are investigated for two index 2 problems: results are given in figure 1, together with those of the index 1 pendulum P1, for reference. One can check the consistency of ZD and SC for both algebraic and differential variables: efficiencies significantly increase with step size for both index 1 and 2 problems.

Figure 1. Efficiencies of ZD and SC with respect to step size

Test 3. The efficiencies of ZD and SC are compared to those of RS with respect to step size: results are given in figure 2 for six index 1 and 2 problems. As expected, the convergence of ZD and SC is much faster than for RS. However, we emphasize that for most problems, when the step size is large, ZD and SC provide bad estimations and RS is better.

Figure 2. Efficiencies of RS, ZD, SC with respect to step size for 6 problems

4.3. Experiments with variable step size

Typically, an implicit code like RADAU5 is designed for variable step size integration. The extension of constant stepsize proofs to variable stepsize is closely related to the regularity of the stepsize selection function [11,3]. Here, we only compare experimentally the performances of RS, ZD and SC with variable step size.

Test 4. Integrations of Test 1 are carried out with variable step size: the average (and min-max) efficiencies of RS, ZD and SC are given in table 3 and are to be compared to those of table 2. It turns out that the high order estimates ZD and SC are not reliable with variable step size: for some problems, the efficiency is often negative, which means that no digit is correctly estimated. A further analysis of MTTA, DPC and other problems shows that this is due to two main points. First, the step size varies a lot during the integration so that it affects the accuracy of the interpolation process in ZD and SC. Second, when the step size is large, ZD and SC are not reliable anymore, as emphasised in Test 3. On the other hand, RS still provides consistent estimations for all examples (one to three digits). Finally, the outcome of our variable step size experimentation is that, since the steps are often large, ZD and SC are not reliable, but RS still produces reliable results.

Table 3
Efficiency of RS, ZD and SC with variable step size for six problems

Pb	Tol	RS		ZD		SC	
		$\mathcal{E}_{av}$	$[\mathcal{E}_{min}, \mathcal{E}_{max}]$	$\mathcal{E}_{av}$	$[\mathcal{E}_{min}, \mathcal{E}_{max}]$	$\mathcal{E}_{av}$	$[\mathcal{E}_{min}, \mathcal{E}_{max}]$
P1	10^{-6}	3.08	[1.90, 3.77]	1.48	[0.41, 2.33]	2.60	[1.98, 2.83]
TTA	10^{-6}	0.93	[0.74, 1.09]	-0.73	[-3.49, 1.24]	-0.73	[-3.49, 1.21]
NL	10^{-5}	2.41	[1.23, 2.72]	2.19	[0.54, 2.91]	4.54	[2.34, 5.79]
MK	10^{-6}	1.67	[0.72, 3.06]	0.87	[-0.61, 2.42]	2.51	[-0.46, 5.72]
P2	10^{-6}	1.73	[1.19, 2.55]	0.95	[0.15, 1.92]	2.38	[0.42, 4.08]
DPC	10^{-4}	1.26	[0.93, 2.11]	-0.26	[-6.07, 2.04]	-0.09	[-6.07, 2.2]

5. Conclusions

Richardson extrapolation, Zadunaisky's technique and Solving for the Correction were proved to be consistent estimation techniques for semi explicit index 1 DAE's. While Richardson extrapolation provides a first order estimation for all variables, Zadunaisky's technique and Solving for the Correction are higher order estimates. For stiffly accurate Runge-Kutta methods, they both have relative order p (classical order of the Runge-Kutta method) for all variables. For non stiffly accurate methods, Solving for the Correction is proved to have higher order than Zadunaisky's technique for the differential component. From a practical point of view, the three estimation techniques turned out to still give consistent results for index 2 problems. We also emphasized that in practice, RADAU5 takes large variable steps so that Richardson extrapolation is the only reliable strategy.

References

- [1] S. Adam. *Globale Fehlerschätzung bei der numerischen Behandlung von Algebro-Differentialgleichungen*. PhD thesis, Technischen Universität Dresden, Dresden, July 1996.
- [2] R. Aïd and L. Levacher. Numerical investigations on global error estimation for ordinary differential equations. *J. of Comput. and Appl. Math.*, 82(1-2):21–39, sept. 1997.
- [3] R. Aïd. *Contribution à l'estimation asymptotique de l'erreur globale des méthodes d'intégration numérique à un pas - Application à la simulation des réseaux électriques*. PhD thesis, INPG, Janvier 1998.
- [4] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. North-Holland, 1989.
- [5] J. C. Butcher and P. Chartier. The construction of DIMSIMs for stiff ODEs and DAEs. *Report of the University of Auckland*, 308, 1994.
- [6] J. R. Dormand, R.R Duckers, and P. J. Prince. Global error estimation with Runge-Kutta methods. *IMA Journal of Numerical Analysis*, 4:169–184, 1984.
- [7] R. Frank and C. W. Ueberhuber. Iterated defect correction for Runge-Kutta methods. Technical Report 14/75, Institut für Numerische Mathematik, T. U. Wien, 1975.
- [8] E. Hairer and G.Wanner. *Solving Ordinary Differential Equations II - Second revised edition*. Springer Series in Computational Mathematics, 1996.
- [9] E. Hairer, C. Lubich, and M. Roche. *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*. Lecture Notes in Mathematics, Springer-Verlag, 1989.
- [10] E. Hairer, S.P. Norsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems*. Springer-Verlag, 1987.
- [11] D. J. Higham. Global error versus tolerance for explicit Runge-Kutta methods. *IMA J. Numer. Anal.*, pages 457–480, 1991.
- [12] S. Scholz. On implicit Runge-Kutta methods with a global error estimation for stiff differential equations. *Zeitschrift für Angewandte Mathematik und Mechanik*, 69(8):253–257, 1989.
- [13] R. D. Skeel. Thirteen ways to estimate global error. *Numer. Math.*, 48:1–20, 1986.
- [14] S. Somali and B. Karasözen. Iterated defect correction methods for differential-algebraic equations. *Tr. J. of Mathematics*, 16:230–249, 1992.
- [15] H. J. Stetter. Local estimation of the global discretisation error. *SIAM J. Num. Anal.*, 8:512–523, 1971.
- [16] J.Visconti. Global error estimation for DAE's solved by Runge-Kutta methods, Research Report, 1998.
- [17] P. E. Zadunaisky. On the estimation of error propagated in the numerical solution of a system of ordinary differential equations. *Num. Math.*, 27:21–39, 1976.