Half-explicit Runge–Kutta methods with explicit stages for differential-algebraic systems of index 2

M. Arnold*

Abstract

Usually the straightforward generalization of explicit Runge–Kutta methods for ordinary differential equations to half-explicit methods for differential-algebraic systems of index 2 results in methods of order $q \leq 2$ ([8]). The construction of higher order methods is simplified substantially by a slight modification of the method combined with an improved strategy for the computation of the algebraic solution components. We give order conditions up to order $q = 5$ and study the convergence of these methods. Based on the fifth order method of Dormand and Prince a half-explicit Runge–Kutta method of order $q = 5$ is constructed that requires the solution of 6 systems of nonlinear equations per step of integration.

Mathematics Subject Classification (1991): 65L05

Keywords: differential-algebraic systems, half-explicit methods, explicit Runge–Kutta methods

1 Introduction

The numerical solution of differential-algebraic systems requires the solution of nonlinear equations. That is why implicit discretization methods (implicit Runge–Kutta methods, BDF) are very popular for the integration of differential-algebraic systems. However, for non-stiff systems the numerical effort can be substantially decreased if only the algebraic part is discretized implicitly (e. g. [7, 11, 2, 1, 13]).

In the present paper we study the solution of the initial value problem

$$\left.\begin{array}{rcl} y'(t) & = & f(y(t), z(t)) \\ 0 & = & g(y(t)) \end{array}\right\} \ , \ t \in [t_0, t_e] \ , \ y(t_0) = y_0 \ , \ z(t_0) = z_0 \tag{1}$$

by half-explicit Runge–Kutta methods. We suppose that (1) has a solution $y : [t_0, t_e] \to \mathbb{R}^{n_y}$, $z : [t_0, t_e] \to \mathbb{R}^{n_z}$ and that f and g are sufficiently differentiable and satisfy the *index-2 condition*

$$[g_y f_z](\eta, \zeta) \quad \text{non-singular} \tag{2}$$

*University of Rostock, Department of Mathematics, Postfach, D – 18051 Rostock, Germany, E-mail: arnold@mathematik.uni-rostock.d400.de

in a neighbourhood of the trajectory $\{ (y(t), z(t)) : t \in [t_0, t_e] \}$. The initial values y_0, z_0 are assumed to be consistent: $g(y_0) = 0$, $[g_y f](y_0, z_0) = 0$. Because of (2) the differential-algebraic system (1) is of (differential and perturbation) index 2 ([8, p. 3f]). Systems of the form (1) arise in various applications, e. g. as index-2 formulation of model equations for constrained mechanical systems and in the integration of systems of ordinary differential equations (ODEs) with invariants (see Section 4 and [5]).

Half-explicit Runge–Kutta methods for (1) were introduced by Hairer et al. in [8]. They compute the differential solution components y similar to explicit Runge–Kutta methods for ODEs. The algebraic solution components z are defined such that all stage values Y_{ni} remain in the manifold $\{ \eta : g(\eta) = 0 \}$ that is given by the algebraic constraints in (1). In contrast to implicit Runge–Kutta methods and BDF that require the solution of systems of $\geq n_y + n_z$ nonlinear equations the systems of nonlinear equations that have to be solved in half-explicit Runge–Kutta methods are of dimension n_z, only. This approach is studied in detail in [3, 2, 4]. The code HEM5 ([2]) that is based on a fifth order half-explicit method is successfully used in the dynamical simulation of constrained mechanical systems.

A drawback of these half-explicit Runge–Kutta methods (we call them *Type A methods*) is a severe order reduction: for traditional higher order explicit Runge–Kutta methods the order usually drops down to $q = 2$. I. e., to handle the algebraic part of (1) Type A methods lose efficiency in the integration of the differential part. In the present paper we modify the approach of [8] in the first Runge–Kutta stage (*Type B methods*). We prove for a wide class of methods up to classical order $p = 5$ that Type B methods do not suffer from order reduction in the differential components provided that the approximation of the algebraic components is sufficiently good. For methods with $p = s \leq 4$ and for the fifth order method of Dormand and Prince ([6], [9, p. 178f]) this approximation is obtained by at most one additional stage of the half-explicit method. We compared HEDOP5 — a fifth order Type B method based on the explicit Runge–Kutta method of Dormand and Prince — with the fifth order Type A method HEM5 [2]. HEDOP5 needs approximately 25% less numerical effort per step of integration but in various numerical tests the discretization error of HEDOP5 is nevertheless substantially smaller than that of HEM5.

In Section 2 we define Type B methods and prove convergence. In Section 3 we analyse the local error, give order conditions up to order $q = 5$ and construct methods of order $q \leq 4$ and the fifth order method HEDOP5. Finally results of a numerical test are given in Section 4.

2 Type B methods: Definition and convergence

One step of an s-stage half-explicit Runge–Kutta method for (1) is given by ([8, p. 20f])

$$\left.\begin{array}{rcl} Y_{ni} &=& y_n + h\sum_{j=1}^{i-1} a_{ij} f(Y_{nj}, Z_{nj}) \\ 0 &=& g(Y_{ni}) \end{array}\right\} , \qquad (i = i_0(1)\hat{s}+1) \ ,$$
$$\left.\begin{array}{rcl} y_{n+1} &=& y_n + h\sum_{j=1}^{s} b_j f(Y_{nj}, Z_{nj}) \\ 0 &=& g(y_{n+1}) \end{array}\right\} \tag{3}$$

with $Y_{n1} := y_n$, $i_0 := 2$, $\hat{s} := s$ (Type A method). The a_{ij}, b_j denote the coefficients of the method with $a_{ij} := 0$ if $j \geq i$. Throughout the paper we set $c_1 := 0$, $c_i := \sum_j a_{ij}$, $(i = 2(1)\hat{s}+1)$, then all results that are formulated for autonomous systems (1) remain valid in the non-autonomous case if $f(Y_{ni}, Z_{ni})$, $g(Y_{ni})$ is substituted by $f(t_n + c_i h, Y_{ni}, Z_{ni})$, $g(t_n + c_i h, Y_{ni})$. If $a_{i,i-1} \neq 0$, $(i = 2(1)s)$, $b_s \neq 0$, then the method is well-defined ([8, Theorem 4.10]): in the i-th stage $Z_{n,i-1}$ is computed as solution ζ of the system of nonlinear equations

$$0 = g\Big(y_n + h\sum_{j=1}^{i-2} a_{ij} f(Y_{nj}, Z_{nj}) + h a_{i,i-1} f(Y_{n,i-1}, \zeta)\Big) \tag{4}$$

that is locally uniquely solvable if h is sufficiently small and y_n is close to $y(t_n)$ (because of (2) the Implicit function theorem is applicable here). The method defines an approximation y_{n+1} to $y(t_n + h)$, Hairer et al. [8] suggest to use methods (3) with $c_s = 1$ such that $z_{n+1} := Z_{ns}$ gives an approximation to $z(t_n + h)$.

Remark 1

a) *As in the ODE case* the final stage y_{n+1} should give a high order approximation to $y(t_n + h)$ but the first stage vector Y_{n2} gives only a poor approximation to $y(t_n + c_2 h)$. If an explicit Runge–Kutta method is applied with $y_n := y(t_n)$ to the ODE $y' = \varphi(y)$ then $Y_{n2} = y_n + c_2 h\varphi(y_n) = y(t_n) + c_2 h y'(t_n) = y(t_n + c_2 h) + \mathcal{O}(h^2)$ and $\varphi(Y_{n2})$ is therefore a poor approximation to $y'(t_n + c_2 h)$. In higher order methods the influence of Y_{n2} on the final stage vector y_{n+1} is usually kept small setting $b_2 := 0$. (In the fifth order method of Dormand and Prince ([9, p. 178f]) we even have for $i = 2$

$$b_i = 0 \ , \ \sum\nolimits_j b_j a_{ji} = 0 \ , \ \sum\nolimits_j b_j c_j a_{ji} = 0 \ , \ \sum\nolimits_{j,k} b_j a_{jk} a_{ki} = 0 \ .) \tag{5}$$

b) *In contrast to the ODE case* $f(Y_{ni}, Z_{ni})$ is a poor approximation to $y'(t_n + c_i h)$ not only for $i = 2$ but also for $i = 1$ (even if $y_n := y(t_n)$): The Taylor expansion of [8, p. 68f] proves

$$\begin{array}{rcl} f(Y_{n1}, Z_{n1}) &=& f(y_n, Z_{n1}) = f(y(t_n), z(t_n)) + f_z(y(t_n), z(t_n))(Z_{n1} - z(t_n)) + \mathcal{O}(h^2) \\ &=& y'(t_n) + \dfrac{c_2 h}{2}[f_z(-g_y f_z)^{-1} g_{yy}(f,f)](y(t_n), z(t_n)) + \mathcal{O}(h^2) \ , \end{array}$$

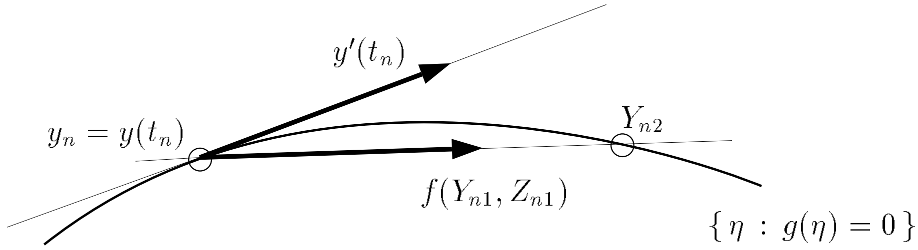


Figure 1: Difference between $f(Y_{n1}, Z_{n1})$ and $y'(t_n)$.

i. e., if the algebraic constraints g are nonlinear ($g_y \neq \text{const}$) then $f(Y_{n1}, Z_{n1}) - y'(t_n) = \mathcal{O}(h)$. Figure 1 illustrates that the large difference between $f(Y_{n1}, Z_{n1})$ and $y'(t_n)$ is caused by the condition $g(Y_{n2}) = 0$. If $Y_{n1} = y(t_n)$ then $y'(t_n)$ lies in the tangential plane of the manifold $\{\,\eta\,:\,g(\eta)=0\,\}$ at $y(t_n)$ since $0 = \frac{d}{dt} g(y(t))|_{t=t_n} = g_y(y(t_n))y'(t_n)$. On the other hand $f(Y_{n1}, Z_{n1}) = \frac{1}{c_2 h}(Y_{n2} - y(t_n))$, i. e. $f(Y_{n1}, Z_{n1})$ is parallel to the vector that joins the points $y(t_n)$, Y_{n2} of this manifold.

Remark 2

a) Brasey and Hairer ([3, 2]) reduce the influence of $f(Y_{n1}, Z_{n1})$ on the final stage setting $b_1 = b_2 = 0$, $\sum_j b_j a_{j1} = \sum_j b_j a_{j2} = 0$ for a method of order $q = 4$ and (5) for $i = 1, 2, 3$ for a method of order $q = 5$. The resulting methods have 5 and 8 stages, respectively, i. e. more stages than in the ODE case (there are explicit Runge–Kutta methods for ODEs of order $p = 4$ with $s = 4$ and of order $p = 5$ with $s = 6$ stages). Another possibility to eliminate the large error term in $f(Y_{n1}, Z_{n1})$ are extrapolation methods [8, p. 49].

b) Figure 1 suggests an alternative definition of half-explicit Runge–Kutta methods: For higher order methods it will in general be important that $f(Y_{n1}, Z_{n1})$ approximates $y'(t_n)$, it might be less important that Y_{n2} lies in the manifold $\{\,\eta\,:\,g(\eta)=0\,\}$. We therefore define $Z_{n1} := z_n$, $Y_{n2} := y_n + c_2 h f(y_n, z_n)$ and need as z_n a sufficiently good approximation of $z(t_n)$. In the first step of integration z_n is available from the (consistent) initial values, approximations for the subsequent steps are obtained adding new stages at the end of each step of integration.

Definition 1 An $\hat{s}$-stage half-explicit Runge–Kutta method of Type B is given by

$$Y_{n1} := y_n \quad , \quad Z_{n1} := z_n \quad , \quad Y_{n2} := y_n + c_2 h f(y_n, z_n)$$

and (3) with $i_0 = 3$, $\hat{s} \geq s$,

$$z_{n+1} := \sum_{i=1}^{\hat{s}} d_i Z_{ni} \quad .$$

Here $d_1, \ldots, d_{\hat{s}}$ and a_{ij}, ($i = 2(1)\hat{s}+1$, $j = 1(1)i-1$), are parameters of the method with $a_{s+1,j} := b_j$, ($j = 1(1)s$), $c_{s+1} := 1$, $a_{i,i-1} \neq 0$, ($i = 2(1)\hat{s}+1$).

In contrast to Type A methods the stage vectors Y_{ni} and y_{n+1} depend now on z_n, i. e. during integration errors are propagated not only in the differential components y but also in the algebraic components z. More precisely: consider vectors $\hat{y}_n$, $\hat{z}_n$, $\hat{Y}_{ni}$, $\hat{Z}_{ni}$, $\hat{y}_{n+1}$, $\hat{z}_{n+1}$, that satisfy $\hat{Y}_{n1} := \hat{y}_n$, $\hat{Z}_{n1} := \hat{z}_n$, $\hat{Y}_{n2} := \hat{y}_n + c_2 h f(\hat{y}_n, \hat{z}_n) + h\delta_2$,

$$\left.\begin{array}{rcl} \hat{Y}_{ni} & = & \hat{y}_n + h \sum_{j=1}^{i-1} a_{ij} f(\hat{Y}_{nj}, \hat{Z}_{nj}) + h\delta_i \\ \theta_i & = & g(\hat{Y}_{ni}) \end{array}\right\} \quad , \qquad (i = 3(1)\hat{s}+1) \ , \tag{6}$$

$\hat{y}_{n+1} := \hat{Y}_{n,s+1}$, $\hat{z}_{n+1} := \sum_i d_i \hat{Z}_{ni}$. It holds (see also [8, Theorem 4.2])

Theorem 1 *Let (y_n, z_n), $(\hat{y}_n, \hat{z}_n)$ be given with*

$$\|y_n - y(t_n)\| + \|z_n - z(t_n)\| = \mathcal{O}(h) \ , \ \|\hat{y}_n - y(t_n)\| + \|\hat{z}_n - z(t_n)\| = \mathcal{O}(h)$$

and

$$\|g(y_n)\| = \mathcal{O}(h^2), \ \|g(\hat{y}_n)\| = \mathcal{O}(h^2), \ \|\delta\| := \max_i \|\delta_i\| = \mathcal{O}(h), \ \|\theta\| := \max_i \|\theta_i\| = \mathcal{O}(h^2) \ .$$

Then there are vectors (Y_{ni}, Z_{ni}), $(\hat{Y}_{ni}, \hat{Z}_{ni})$ with (3), (6) and

$$\|Y_{ni} - y(t_n)\| + \|Z_{ni} - z(t_n)\| = \mathcal{O}(h) \ , \ \|\hat{Y}_{ni} - y(t_n)\| + \|\hat{Z}_{ni} - z(t_n)\| = \mathcal{O}(h)$$

if the stepsize $h \geq 0$ is sufficiently small. These vectors satisfy the estimates

$$\|\hat{Y}_{ni} - Y_{ni}\| \ \leq \ (1 + Ch)\|P(t_n)(\hat{Y}_{ni} - Y_{ni})\| + C\|\theta_i\| \ , \ (i = 3(1)\hat{s}+1), \tag{7}$$

$$\|P(t_n)(\hat{Y}_{ni} - Y_{ni})\| \ \leq \ C(\|\hat{y}_n - y_n\| + h^2\|\hat{z}_n - z_n\| + h\|\delta\| + \|\theta\|), \ (i = 2(1)\hat{s}+1), \tag{8}$$

$$\begin{aligned} \|\hat{Z}_{ni} - Z_{ni} - c_2 w_{i1}(\hat{z}_n - z_n)\| \leq C(\tfrac{1}{h}\|g_y(y(t_n))(\hat{y}_n - y_n)\| + \\ + \|\hat{y}_n - y_n\| + h\|\hat{z}_n - z_n\| + \|\delta\| + \tfrac{1}{h}\|\theta\|) \ , \ (i = 1(1)\hat{s}) \end{aligned} \tag{9}$$

with a constant C that is independent of h, δ and θ. Here P denotes the projector $P(t) := I - [f_z(g_y f_z)^{-1} g_y](y(t), z(t))$ and w_{ij}, $(i, j = 1(1)\hat{s})$ are the elements of the inverse $W := \tilde{A}^{-1}$ of the lower triangular matrix $\tilde{A} := (a_{i+1,j})_{i,j=1(1)\hat{s}}$. $P(t)$ represents a projection into the tangential plane of the manifold $\{\, \eta : g(\eta) = 0 \,\}$, ([8, p. 35 and p. 68]).

Proof Theorem 1 is the counterpart to [8, Theorem 4.2]. The existence of vectors (Y_{ni}, Z_{ni}), $(\hat{Y}_{ni}, \hat{Z}_{ni})$ in a $\mathcal{O}(h)$–neighbourhood of $(y(t_n), z(t_n))$ that satisfy (3), (6) is guaranteed by the Implicit function theorem since $a_{i,i-1} \neq 0$, $(i = 2(1)\hat{s}+1)$, (see also [8, Theorem 4.1]).
Estimate (7) is a consequence of

$$\begin{aligned} g_y(y(t_n))(\hat{Y}_{ni} - Y_{ni}) \ &= \ \int_0^1 \Big(g_y(y(t_n))(\hat{Y}_{ni} - Y_{ni}) - \frac{d}{d\vartheta} g(Y_{ni}^{\vartheta}) \Big) d\vartheta + g(\hat{Y}_{ni}) - g(Y_{ni}) \\ &= \ \int_0^1 \Big(g_y(y(t_n)) - g_y(Y_{ni}^{\vartheta}) \Big) d\vartheta \cdot (\hat{Y}_{ni} - Y_{ni}) + \theta_i \end{aligned} \tag{10}$$

with $Y_{ni}^{\vartheta} := Y_{ni} + \vartheta(\hat{Y}_{ni} - Y_{ni})$ since

$$\begin{aligned}\|\hat{Y}_{ni} - Y_{ni}\| &\leq \|P(t_n)(\hat{Y}_{ni} - Y_{ni})\| + \|(I - P(t_n))(\hat{Y}_{ni} - Y_{ni})\| \\ &\leq \|P(t_n)(\hat{Y}_{ni} - Y_{ni})\| + C(h\|\hat{Y}_{ni} - Y_{ni}\| + \|\theta_i\|) \ .\end{aligned}$$

To prove (8) and (9) we transform (3) and (6) to

$$\hat{Y}_{ni} - Y_{ni} = \hat{y}_n - y_n + h\sum_j a_{ij}\Big((f_y^n + \mathcal{O}(h))(\hat{Y}_{nj} - Y_{nj}) + (f_z^n + \mathcal{O}(h))(\hat{Z}_{nj} - Z_{nj})\Big) + h\delta_i \quad (11)$$

where the upper index "n" indicates that the Jacobians f_y, f_z, g_y are evaluated at $(y(t_n), z(t_n))$. With the notations $Y := (Y_{n2}^T, \ldots, Y_{n,\hat{s}+1}^T)^T$, $Z := (Z_{n1}^T, \ldots, Z_{n,\hat{s}}^T)^T$, ... we thus obtain

$$\begin{aligned}&h((\tilde{A} \otimes [g_y^n f_z^n]) + \mathcal{O}(h))(\hat{Z} - Z) = \\ &\qquad = -(\mathbf{1} \otimes g_y^n)(\hat{y}_n - y_n) + \mathcal{O}(h)\|\hat{Y} - Y\| + (I \otimes g_y^n)(\hat{Y} - Y) + \mathcal{O}(h\|\delta\|)\end{aligned} \quad (12)$$

with $\mathbf{1} := (1, \ldots, 1)^T$ and the Kronecker product $\otimes$. Because of $g(Y_{ni}) = 0$, $g(\hat{Y}_{ni}) = \theta_i$, $(i \geq 3)$ the term $(I \otimes g_y^n)(\hat{Y} - Y)$ can be expressed as (see (10))

$$(I \otimes g_y^n)(\hat{Y} - Y) = \mathcal{O}(\|\theta\|) + \mathcal{O}(h)\|\hat{Y} - Y\| + \boldsymbol{e_1} \otimes g_y^n \cdot (\hat{Y}_{n2} - Y_{n2})$$

with $\boldsymbol{e_1} := (1, 0, \ldots, 0)^T$ and

$$g_y^n \cdot (\hat{Y}_{n2} - Y_{n2}) = g_y^n \cdot (\hat{y}_n - y_n) + c_2 h[g_y^n f_z^n](\hat{z}_n - z_n) + \mathcal{O}(h)(\|\hat{y}_n - y_n\| + h\|\hat{z}_n - z_n\|) \ .$$

So we get in (12)

$$\begin{aligned}&((\tilde{A} \otimes [g_y^n f_z^n]) + \mathcal{O}(h))\Big(\hat{Z} - Z - c_2 W\boldsymbol{e_1} \cdot (\hat{z}_n - z_n)\Big) = \mathcal{O}(\tfrac{1}{h})\|g_y^n \cdot (\hat{y}_n - y_n)\| + \\ &\qquad + \mathcal{O}(1)(\|\hat{y}_n - y_n\| + \|\hat{Y} - Y\| + h\|\hat{z}_n - z_n\|) + \mathcal{O}(1)\|\delta\| + \mathcal{O}(\tfrac{1}{h})\|\theta\| \ .\end{aligned} \quad (13)$$

With (2) the matrix $\tilde{A} \otimes [g_y^n f_z^n]$ is non-singular and the estimates (8), (9) follow from (13) and (11), (note, that $P(t_n)(f_z^n + \mathcal{O}(h)) = \mathcal{O}(h)$ since $P(t)f_z(y(t), z(t)) \equiv 0$). ∎

Theorem 1 is the essential part of the convergence analysis for Type B methods that bounds the global discretization error in terms of the local error:

Definition 2 Let $\hat{y}_{n+1}$, $\hat{z}_{n+1}$ be defined by (6) with $\delta_i = 0$, $\theta_i = 0$, $(i = 2(1)\hat{s} + 1)$ and the analytical solution of (1) at time t_n as $\hat{y}_n$, $\hat{z}_n$: $\hat{y}_n := y(t_n)$, $\hat{z}_n := z(t_n)$. Then

$$\delta y_h(t_n) := \hat{y}_{n+1} - y(t_n + h) \ , \ \delta z_h(t_n) := \hat{z}_{n+1} - z(t_n + h)$$

are called the *local discretization errors* in the differential and algebraic part, respectively.

Theorem 2 *Suppose that in (3)* $a_{i,i-1} \neq 0$, *(*$i = 2(1)\hat{s} + 1$*), that the local discretization errors are of size*

$$\delta y_h(t_n) = \mathcal{O}(h^q) \ , \ P(t_n)\delta y_h(t_n) = \mathcal{O}(h^{q+1}) \ , \ \delta z_h(t_n) = \mathcal{O}(h^{q-1})$$

with some $q \geq 2$ *and that the* contractivity condition

$$|c_2 d^T W \boldsymbol{e_1}| < 1 \tag{14}$$

is satisfied. Then the method is convergent with orders q *and* $q-1$*, respectively, i. e.*

$$\|y_m - y(t_m)\| + h\|z_m - z(t_m)\| = \mathcal{O}(h^q) \quad \textit{for } t_m = t_0 + mh \in [t_0, t_e] \; .$$

Here d *denotes the vector of coefficients* $d := (d_1, \ldots, d_{\hat{s}})^T$*, i. e.* $c_2 d^T W \boldsymbol{e_1} = c_2 \sum_j d_j w_{j1}$ *.*

Proof

a) As in (7) we get $\|y_m - y(t_m)\| \leq (1 + \mathcal{O}(h))\|P(t_m)(y_m - y(t_m))\|$ since $g(y_m) = 0$ and $g(y(t_m)) = 0$. I. e. if $\|P(t_m)(y_m - y(t_m))\| = \mathcal{O}(h^q)$ then convergence with order q in the differential components is guaranteed.

b) Let $\hat{y}_{n+1}$, $\hat{z}_{n+1}$ be defined as in Definition 2 then

$$\begin{aligned} P(t_{n+1})(y_{n+1} - y(t_{n+1})) &= P(t_{n+1})(y_{n+1} - \hat{y}_{n+1}) + P(t_{n+1})\delta y_h(t_n) \\ h(z_{n+1} - z(t_{n+1})) &= h(z_{n+1} - \hat{z}_{n+1}) + \mathcal{O}(h^q) \end{aligned}$$

and $P(t_{n+1})\delta y_h(t_n) = P(t_n)\delta y_h(t_n) + \mathcal{O}(h)\delta y_h(t_n) = \mathcal{O}(h^{q+1})$. In contrast to Type A methods errors are not only propagated in the function values $y_n \approx y(t_n)$ of the differential components but also in the derivatives $f(y_n, z_n) \approx y'(t_n)$. This is similar to the error propagation in Rosenbrock methods that are applied to differential-algebraic equations $B(y)y' = \varphi(y)$ of index 1 ([12]). The statement of the theorem can be proved following the proof of [12, Theorem 4.1] using the inequality

$$\begin{pmatrix} \|P(t_{n+1})(\hat{y}_{n+1} - y_{n+1})\| \\ h\|\hat{z}_{n+1} - z_{n+1}\| \end{pmatrix} \leq \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & \rho + \mathcal{O}(h) \end{pmatrix} \begin{pmatrix} \|P(t_n)(\hat{y}_n - y_n)\| \\ h\|\hat{z}_n - z_n\| \end{pmatrix} \tag{15}$$

with $\rho := |c_2 d^T W \boldsymbol{e_1}|$ that has to be read componentwise. Estimate (15) is obtained straightforwardly from Theorem 1 since $P(t_{n+1}) = P(t_n) + \mathcal{O}(h)$ and $\|y_n - y(t_n)\| \leq (1 + \mathcal{O}(h))\|P(t_n)(y_n - y(t_n))\|$, (see (7)). ■

Remark 3

a) As in [12] the restriction to fixed stepsizes is not essential.

b) In contrast to Type A methods we need consistent initial values for differential *and* algebraic solution components, the statement of Theorem 2 remains valid if $z_0 := z(t_0)$ is substituted by $z_0 := z(t_0) + \mathcal{O}(h^{q-1})$.

c) The contractivity condition $|c_2 d^T W \boldsymbol{e_1}| < 1$ is very important. If $|c_2 d^T W \boldsymbol{e_1}| > 1$ then the numerical solution in general diverges, if $|c_2 d^T W \boldsymbol{e_1}| = 1$ then order reduction may occur.

d) The solution of (1) satisfies $0 = \frac{d}{dt} g(y(t)) = g_y(y(t))y'(t) = [g_y f](y(t), z(t))$. Hence, a numerical solution ζ for the algebraic components $z(t_n)$ can be obtained as solution of $[g_y f](y_n, \zeta) = 0$ if a high order approximation y_n to the differential components $y(t_n)$ has been computed before ([3]). Thus [2, 3] do not focus on a high order of convergence for the algebraic components z, the methods HEM4 and HEM5 that converge in the differential components y with order $q = 4$ and $q = 5$, respectively, have order $q - 2$ for z. As a by-product of our approach Type B methods give a better approximation for the algebraic components z: a method of order q (in y) has a global error of order $\mathcal{O}(h^{q-1})$ in z.

3 Type B methods of order $q \leq 5$

Estimates for the local error are obtained by Taylor expansion using the tree model of [8, Section 5]. This procedure is straightforward but needs a lot of notations. That is why we follow the analysis of the local error of Type A methods ([8, p. 68ff], [3, 2]) as far as possible. The essential difference to Type A methods is in the first stage where we do not have $g(Y_{n2}) = 0$ but

$$Y_{n2} = y(t_n) + c_2 h y'(t_n) \ , \quad Z_{n1} = z(t_n)$$

provided that $(y_n, z_n) = (y(t_n), z(t_n))$. The order conditions can be obtained as in [8] if Formula (5.34b) that reads in our notation

$$\dot{Z}_{ni}(0) = \frac{1}{2} \sum\nolimits_{j,k,l} w_{ij} a_{j+1,k} a_{j+1,l} \cdot (-g_y f_z)^{-1} g_{yy}(f,f) + \sum\nolimits_j a_{ij} \cdot (-g_y f_z)^{-1} g_y f_y f$$

is substituted by

$$\dot{Z}_{ni}(0) = \frac{1}{2} \sum\nolimits_{j,k,l} w_{ij} \hat{a}_{j+1,k} \hat{a}_{j+1,l} \cdot (-g_y f_z)^{-1} g_{yy}(f,f) + \sum\nolimits_j a_{ij} \cdot (-g_y f_z)^{-1} g_y f_y f$$

with $\hat{a}_{j+1,k} := 0$, $(j = 1)$ and $\hat{a}_{j+1,k} := a_{j+1,k}$, $(j > 1)$. Similar modifications are necessary for the higher order terms $\frac{d^p}{dh^p} Z_{ni}\Big|_{h=0}$.

For the differential components y the order conditions for Type B methods are obtained from the ones for Type A methods ([8, (5.38)], [3, Table 1]) by

Algorithm 1 Substitute all terms of the form

$$\sum_{j,k_1,\ldots,k_l} w_{ij} a_{j+1,k_1} \cdots a_{j+1,k_l} \qquad \text{with} \quad l > 1 \quad \text{by} \qquad \sum_{j,k_1,\ldots,k_l} w_{ij} \hat{a}_{j+1,k_1} \cdots \hat{a}_{j+1,k_l} \ .$$

With the usual notation $c_i := \sum_j a_{ij}$ Algorithm 1 substitutes $\sum_j w_{ij} c_{j+1}^l$ by $\sum_j w_{ij} \hat{c}_{j+1}^l$ with $\hat{c}_2 := 0$ and $\hat{c}_i := c_i$, $(i \geq 3,\ l > 1)$.

The slight difference in the order conditions simplifies the construction of higher order half-explicit Runge–Kutta methods essentially:

Example 1 An underlying explicit Runge–Kutta method that satisfies the simplifying condition

$$\mathrm{C}(2) \ : \quad \sum_{j=1}^{i-1} a_{ij} c_j = \frac{1}{2} c_i^2 \ , \quad (i = 3(1)\hat{s} + 1)$$

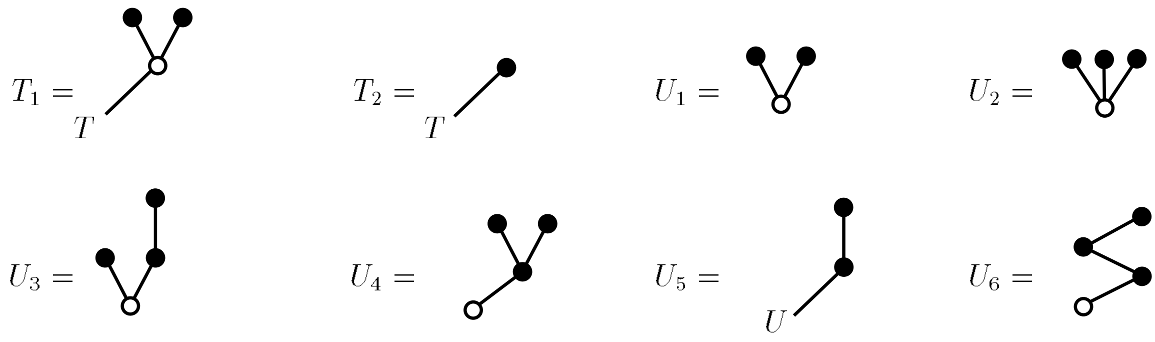


Figure 2: Trees and simplifying conditions.

implies here (note, that $W = \tilde{A}^{-1}$)

$$\mathrm{C}(2)\mathrm{R} \ : \quad \sum_{j=1}^{i} w_{ij}\hat{c}_{j+1}^2 = 2c_i \ , \quad (i = 1(1)\hat{s}) \ .$$

These *reciproque conditions* are substantially simpler than that for Type A methods ([2]) and guarantee that the order condition corresponding to the tree T_1 of Figure 2 is satisfied whenever the order condition corresponding to tree T_2 is satisfied. Here T denotes an arbitrary trunk.

More general: trees that contain a "bushy" tree of order $< r$ with fat root do not give additional order conditions for Type B methods if the underlying explicit Runge–Kutta method satisfies the simplifying condition

$$\mathrm{C}(r) \ : \quad \sum_{j=1}^{i-1} a_{ij}c_j^l = \frac{1}{l+1}c_i^{l+1} \ , \quad (i = 3(1)\hat{s}+1, \, l = 0(1)r-1) \ ,$$

(see Figure 2, U_1 and U_2 are the bushy trees of order 1 and 2 with fat root).

To formulate the main result of this section we introduce another simplifying condition:

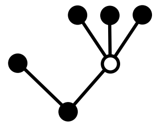
$$\mathrm{D}(1) \ : \quad \sum_{j=i+1}^{s} b_j a_{ji} = b_i(1-c_i) \ , \quad (i = 1(1)s) \ .$$

Theorem 3 *An explicit Runge–Kutta method of order p with $2 \le p \le 5$ yields to a Type B method with local discretization error $\delta y_h(t) = \mathcal{O}(h^{p+1})$ if the simplifying conditions*

- C(2) *if* $p = 3$,
- C(2) *and* D(1) *if* $p = 4$,
- C(3) *and* D(1) *if* $p = 5$

are satisfied.

Proof If $p = 2$ or $p = 3$ and C(2) is satisfied then there are no additional order conditions for Type B methods (see [8, (5.35)–(5.38)], Algorithm 1 and the reciproque condition C(2)R).
If $p = 4$ and C(2) is satisfied then $b_2 = 0$ since $0 = \frac{1}{3} - \frac{1}{3} = \sum_i b_i c_i^2 - 2\sum_{i,j} b_i a_{ij} c_j = \sum_i b_i (c_i^2 - 2\sum_j a_{ij} c_j) = b_2 c_2^2$ and $c_2 = a_{21} \neq 0$. Compared with the ODE case the only additional order condition is ([3, Table 1] and Algorithm 1)



$$\sum\nolimits_{i,j} b_i c_i w_{ij} \hat{c}_{j+1}^3 = \frac{3}{4} \quad . \tag{16}$$

Note, that $W = \tilde{A}^{-1}$ and therefore $\sum_{i,j,k} b_i a_{ij} w_{jk} \hat{c}_{k+1}^3 = \sum_{i,k} b_i \delta_{i,k+1} \hat{c}_{k+1}^3 = \sum_{i>2} b_i c_i^3 = \frac{1}{4}$ and $\sum_{i,j} b_i w_{ij} \hat{c}_{j+1}^3 = \sum_j \delta_{sj} \hat{c}_{j+1}^3 = \hat{c}_{s+1}^3 = c_{s+1}^3 = 1$ such that (16) follows from D(1), (δ_{ij} denotes the Kronecker delta).
If C(3) is satisfied then $\sum_{j,k} w_{ij} \hat{c}_{j+1} a_{j+1,k} c_k = \frac{1}{2} \sum_j w_{ij} \hat{c}_{j+1}^3$, i. e. trees T that contain one of the trees U_1, U_2, U_3 of Figure 2 do not give additional order conditions. Therefore the statement for $p = 5$ can be proved as in the case $p = 4$. ■

The order of the local error in the differential components y is completely defined by the parameters of the underlying s stage explicit Runge–Kutta method since $y_{n+1} = Y_{n,s+1}$. The additional parameters d_i, ($i = 1(1)\hat{s}$) and a_{ij}, ($i = s + 2(1)\hat{s} + 1$, $j = 1(1)i - 1$) (if new stages are added, i. e. if $\hat{s} > s$) are defined such that the order conditions for the algebraic components z are satisfied. As for the differential components the order conditions can be derived straightforwardly from the ones for Type A methods ([3, Table 4]):

Algorithm 2

Apply Algorithm 1 and substitute additionally the leading term w_{sj} by $\displaystyle\sum_{i=1}^{\hat{s}} d_i w_{ij}$.

Theorem 4 *The local error in the algebraic components is of size* $\delta z_h(t) = \mathcal{O}(h^{r+1})$ *with* $0 \le r \le 3$ *if*

$$\sum_{i=1}^{\hat{s}} d_i c_i^l = 1 \ , \quad (l = 0(1)r) \ , \tag{17}$$

$$d_2 = 0 \ , \quad (\textit{if } r \ge 2 \ ) \textit{ and } \quad \sum_{i=1}^{\hat{s}} d_i a_{i2} = \sum_{i,j=1}^{\hat{s}} d_i w_{ij} c_{j+1} a_{j+1,2} = 0 \ , \quad (\textit{if } r \ge 3 \ ) \ , \tag{18}$$

and the coefficients of the method satisfy for $r \ge 1$

- *the simplifying condition* C$(r + 1)$ *or*
- *the simplifying condition* C(r) *and*

$$\frac{1}{r+1} \sum_{i,j=1}^{\hat{s}} d_i w_{ij} \hat{c}_{j+1}^{r+1} = 1 \quad . \tag{19}$$

Proof The order conditions (17) correspond to the "one-leg" trees with fat root and a "bushy" tree of order $\leq r+1$ as branch (see e. g. the tree U_4 of Figure 2 that corresponds to

$$1 = \sum\nolimits_{i,j,k} d_i w_{ij} a_{j+1,k} c_k^2 = \sum\nolimits_{i,k} d_i \delta_{ik} c_k^2 = \sum\nolimits_i d_i c_i^2 \ ).$$

The conditions (18) guarantee that up to order r the order conditions for trees of the form U_5 are satisfied whenever (17) and the simplifying assumption C(r) is satisfied: e. g. the tree U_6 of Figure 2 gives ($r \geq 2$)

$$\sum\nolimits_{i,j} d_i a_{ij} c_j = \frac{1}{2}\sum\nolimits_i d_i c_i^2 + \sum\nolimits_{i\geq 3} d_i (\sum\nolimits_j a_{ij} c_j - \frac{1}{2} c_i^2) - \frac{1}{2} d_2 c_2^2 = \frac{1}{2}\sum\nolimits_i d_i c_i^2 = \frac{1}{2} \ .$$

The "bushy" trees with fat root (see U_1, U_2 in Figure 2) give order conditions

$$\frac{1}{l+1} \sum_{i,j=1}^{\hat{s}} d_i w_{ij} \hat{c}_{j+1}^{l+1} = 1 \ , \quad (l = 1(1)r) \ .$$

These conditions are for $l = 1(1)r-1$ equivalent to (17) if C(r) is satisfied, they are equivalent to (17) for $l = 1(1)r$ if C($r+1$) is satisfied. All the remaining trees contain at a branch leaving a meagre vertex one of the trees U_1, U_2, U_3 of Figure 2. As for the differential components these trees do not give new order conditions if the simplifying condition C(r) is satisfied. ■

To guarantee convergence we have to consider not only order conditions but also the contractivity condition (14) that may be rewritten as

$$\sum\nolimits_{i,j} d_i w_{ij} \hat{c}_{j+1} \in (0, 2) \tag{20}$$

since $c_2 d^T W \boldsymbol{e_1} = \sum_i d_i w_{i1} c_2 = \sum_{i,j} d_i w_{ij} (c_{j+1} - \hat{c}_{j+1}) = 1 - \sum_{i,j} d_i w_{ij} \hat{c}_{j+1}$.

In Example 2 we construct Type B methods up to order $q = 5$ (in the differential components). Table 1 summarizes the essential numerical effort of these methods per one step of integration. Note, that (in contrast to Type A methods) the first stage of a Type B method is explicit and does not require the solution of a system (4) of nonlinear equations. For a Type B method of order q the order of convergence for the algebraic components equals $q-1$. Order and contractivity conditions result in systems of equations that have to be solved w. r. t. the parameters. In Example 2 we transform at first these conditions into systems of *linear* equations that are finally solved using MATHEMATICA [14].

Example 2

a) $q = 2$. An explicit Runge–Kutta method with $p = s = 2$ yields a Type B method of order $q = 2$ if $\hat{s} := s = 2$, $d_i := b_i$, $(i = 1,2)$ since $\sum_i d_i = \sum_i b_i = 1$ and $c_2 d^T W \boldsymbol{e_1} = c_2 \sum_i d_i w_{i1} = c_2 \sum_i b_i w_{i1} = c_2 \delta_{21} = 0$.

b) $q = 3$. An explicit Runge–Kutta method with $p = s = 3$ that satisfies the simplifying condition C(2) has parameters $b_2 = 0$ (since $b_3 a_{32} c_2 = \frac{1}{6}$ and $\sum_i b_i c_i^2 = \frac{1}{3}$), $c_3 = \frac{2}{3}$, $b_1 = \frac{1}{4}$, $b_3 = \frac{3}{4}$, (since $\sum_i b_i c_i^l = \frac{1}{l+1}$, $l = 0,1,2$) and $a_{32} = \frac{2}{9c_2}$ (since $b_3 a_{32} c_2 = \frac{1}{6}$), the parameter $c_2 \neq 0$ remains free ([9, Section II.1]). Then Theorem 4 with $r = q-2 = 1$

	$q=2$	$q=3$	$q=4$	$q=5$
# of stages (ODE)	2	3	4	6
# of stages (Type A)	2	3	5	8
# of stages (Type B)	2	3	5	7
# of systems (4) (Type B)	1	2	4	6

Table 1: Numerical effort of explicit Runge–Kutta methods and half-explicit methods of Type A and B

and the condition $c_2 d^T W \boldsymbol{e_1} = 0$ give a system of 3 linear equations in the unknowns d_1, d_2, d_3, that has the unique solution $d_1 = -2 + \frac{1}{c_2}$, $d_2 = -\frac{1}{c_2}$, $d_3 = 3$. With these parameters the half-explicit method has order $q = 3$, ($\hat{s} = s = 3$).

c) $q = 4$. An explicit Runge–Kutta method with $p = s = 4$ satisfies condition D(1) and thus $c_4 = 1$ (since $b_4(1 - c_4) = \sum b_j a_{j4} = 0$). The methods that fulfill C(2) form a family with the free parameter c_2 and have coefficients $b_2 = \sum b_j a_{j2} = 0$, $c_3 = \frac{1}{2}$, ([9, p. 138]). If $\hat{s} = s = 4$ then the order conditions (17), (19) for $r = q - 2 = 2$ and the reciproque condition C(2)R imply

$$\sum_{i=1}^{4} \gamma_i = 1 \quad , \quad \frac{1}{2}\sum_{i=1}^{4} \gamma_i \hat{c}_{i+1} = 1 \quad , \quad \frac{1}{3}\sum_{i=1}^{4} \gamma_i \hat{c}_{i+1}^2 = 1 \tag{21}$$

with $\gamma_i := \sum_j d_j w_{ji} c_{i+1}$, $(i = 1(1)4)$ since $\sum_i w_{ji} c_{i+1} = \sum_{i,k} w_{ji} a_{i+1,k} = \sum_k \delta_{jk} = 1$. The solutions of (21) are $\gamma_1 = 1$, $\gamma_2 = -4$, $\gamma_3 + \gamma_4 = 4$ (note, that $c_4 = c_5 = 1$). Because of $\gamma_1 = c_2 d^T W \boldsymbol{e_1}$ a method with $\hat{s} = s = 4$, $P(t)\delta y_h(t) = \mathcal{O}(h^5)$, $\delta z_h(t) = \mathcal{O}(h^3)$ can therefore not satisfy the contractivity condition (14). With the same arguments it is proved that the assumptions of Theorem 2 with $q = 4$ can not be satisfied adding a 5th stage with $c_6 = \sum_i a_{6i} = 1$. Thus we add a stage with $c_6 \neq 1$ and define $z_{n+1} := Z_{n5}$, i. e. $d_i := 0$, $(i = 1(1)4)$, $d_5 := 1$. With this definition of z_{n+1} conditions (17), (18) are trivial and the last call of f in time step $t_n \to t_{n+1}$ is identical with the first call of f in time step $t_{n+1} \to t_{n+2}$, ($f(Y_{n,s+1}, Z_{n,s+1}) = f(y_{n+1}, z_{n+1})$). We choose the parameters a_{6i} such that (see C(2) and (19))

$$\sum_j a_{6j} c_j^l = \frac{1}{l+1} c_6^{l+1} \ , \ (l = 0, 1) \ , \quad \frac{1}{3}\sum_j w_{5j} \hat{c}_{j+1}^3 = 1 \ , \quad w_{51} = 0 \tag{22}$$

then the method is convergent with order $q = 4$, (because of $c_2 d^T W \boldsymbol{e_1} = c_2 w_{51} = 0$ Theorem 2 is applicable). Eqs. (22) have a solution with $a_{65} \neq 0$ whenever $c_6 \notin \{0, \frac{1}{2}, 1\}$, free parameters are a_{62} and c_6.
The Jacobian of the system (4) of nonlinear equations is $h a_{i,i-1}[g_y f_z](y(t_n), z(t_n)) + \mathcal{O}(h)$, this motivates a choice of c_6 that results in parameters a_{6j} with large $|a_{65}|$: $c_6 := \frac{1}{2} - \frac{1}{6}\sqrt{3}$. For simplicity we assume $a_{62} = b_2 = 0$ and get

$$a_{61} = \frac{1}{6} - \frac{1}{108}\sqrt{3} \ , \ \ a_{63} = \frac{1}{3} - \frac{4}{27}\sqrt{3} \ , \ \ a_{64} = -\frac{7}{108}\sqrt{3} \ , \ \ a_{65} = \frac{1}{18}\sqrt{3} \ .$$

d) $q = 5$. An explicit Runge–Kutta method of order $p = 5$ has $s \geq 6$ stages, 17 order conditions have to be satisfied. We therefore restrict ourselves to the 5th order method of Dormand and Prince that is known to be a very effective method for ODEs ([9, p. 178f]]. This method has $s = 6$ stages and satisfies the simplifying conditions C(3), D(1). To get a half-explicit method of order $q = 5$ the local error in the algebraic components has to be of order $\delta z_h(t) = \mathcal{O}(h^4)$, i. e. the coefficients d_i, $(i = 1(1)\hat{s})$ have to satisfy the 8 order conditions (17), (18), (19) for $r = 3$, that are linear in d_i. Straightforward computation shows that these 8 conditions can not be fulfilled simultaneously if $\hat{s} = s = 6$, (at least) one new stage has to be added. As for the 4th order method we set $z_{n+1} := Z_{n,s+1} = Z_{n7}$ and define parameters a_{8i} $(i = 1(1)7)$ by the simplifying condition C(3), by the order conditions

$$a_{82} = 0 \ , \ \ \sum\nolimits_j w_{7j} c_{j+1} a_{j+1,2} = 0 \ , \ \ \frac{1}{4} \sum\nolimits_j w_{7j} \hat{c}_{j+1}^4 = 1 \tag{23}$$

and by $0 = d^T W \boldsymbol{e_1} = w_{71}$. The first 6 rows of $\tilde{A}$ contain only parameters of the underlying explicit Runge–Kutta method. Because of the lower triangular structure of $\tilde{A}$ the elements w_{ij}, $(i \leq 6)$ are independent of the additional coefficients a_{8i}, $(i = 1(1)7)$ and we have (note, that $\tilde{A} W = I$)

$$a_{87} w_{7j} = \sum_{i=1}^{7} a_{8i} w_{ij} - \sum_{i=1}^{6} a_{8i} w_{ij} = - \sum_{i=1}^{6} a_{8i} w_{ij} \ , \quad (j = 1(1)6) \ ,$$

$a_{87} w_{77} = 1$. Multiplying $w_{71} = 0$ and the second and third equation in (23) by a_{87} we thus get an equivalent set of conditions that are *linear* in a_{8i}, $(i = 1(1)7)$. For given $c_8 = \sum_j a_{8j}$ we have with $a_{82} = 0$ and C(3) a system of 7 linear equations in 7 unknowns a_{8i} that is uniquely solvable. If $c_8 \notin \{0, 1, \frac{87 \pm 3\sqrt{161}}{170}\}$ then $a_{87} \neq 0$ and we get a half-explicit method of order $q = 5$.

e) $q = 5$, $c_8 = 1$. The code DOPRI5 of Hairer et al. [9, Appendix] uses an additional stage for the explicit Runge–Kutta method of Dormand and Prince to get an embedded method of order 4 for stepsize control. Therefore $Y_{n8} \approx y(t_n + h)$ and $c_8 = 1$. In contrast to methods of order $q = s = 4$ a convergent Type B method of order $q = p = 5$ with $\hat{s} = s + 1 = 7$ and $c_{\hat{s}+1} = 1$ can be constructed: the conditions C(3), $a_{82} = 0$ and $\sum a_{8i} = c_8 = 1$ define a_{81}, a_{83}, a_{84} as (linear) functions of a_{85}, a_{86}, a_{87}. With these parameters the order conditions $\delta z_h(t) = \mathcal{O}(h^4)$ are uniquely solvable w. r. t. d_i, $(i = 1(1)7)$ iff

$$a_{85} = - \frac{243(209 + 5397 a_{86} + 3267 a_{87})}{690272} \qquad \text{and} \qquad 420 a_{86} + 704 a_{87} \neq 55 \ . \tag{24}$$

With these parameters the contractivity condition is satisfied $(c_2 d^T W \boldsymbol{e_1} = -\frac{8}{9})$ and the method is convergent with order $q = 5$. The order of the embedded method is $\hat{q} = 3$ since $\sum_i a_{8i} = 1$, C(3) and (24) imply $\sum_i a_{8i} a_{i2} \neq 0$ such that the order conditions $\sum a_{8i} a_{ij} c_j^2 = \frac{1}{12}$ and $\sum a_{8i} a_{ij} a_{jk} c_k = \frac{1}{24}$ can not be satisfied simultaneously (note, that $\sum_j a_{ij} c_j = \frac{1}{2} c_i^2$ if $i \geq 3$). I. e. in contrast to the ODE case there is not embedded method of order $\hat{q} = 4$ with $\hat{s} = s + 1$ stages.

4 Numerical Test

In the present paper we concentrate on the convergence analysis of Type B methods. To make the methods available for practical computations they have to be combined with stepsize control, dense output etc. as it was done for Type A methods in the code HEM5 by Brasey [2]. Because of the very similar structure of Type A and Type B methods many tricky details of the implementation of Type A methods can be carried over to methods of Type B. In this section we compare the pure discretization schemes applied with fixed stepsize to the model equations of a conservative constrained mechanical system.

Half-explicit methods are expected to be much more efficient than fully implicit methods if a non-stiff differential-algebraic system (1) is of the special form ($y = (q^T, u^T)^T$, $z = \lambda$)

$$\begin{aligned} q'(t) &= u(t) \\ M(q(t))u'(t) &= \hat{f}(q(t), u(t)) - G^T(q(t))\lambda(t) \\ 0 &= G(q(t))u(t) \end{aligned} \tag{25}$$

since Equations (4) are *linear* in this case, (see [2, Figure 6.2] for a comparison of HEM5 and DASSL applied to (25)). Systems of the form (25) arise as model equations of constrained mechanical systems if the original holonomic constraints $\hat{g}(q) = 0$ are substituted by the equations $0 = \frac{d}{dt}\hat{g}(q(t)) = \frac{\partial}{\partial q}\hat{g}(q)q'(t) = G(q)u$ (index-2 formulation, [10, p. 485]). Here q and u denote position and velocity coordinates, the Lagrangian multipliers λ couple the constraints to the dynamical equations, the Jacobian $G(q) := \frac{\partial}{\partial q}\hat{g}(q)$ is supposed to have full rank. $\hat{f}(q,u)$ are the applied forces and $M(q)$ the (symmetric) mass matrix. We require that $M(q)$ is positive definite such that (25) can be transformed to the semi-explicit form (1) multiplying the dynamical equations by $M^{-1}(q)$. (For Type A methods this condition can be relaxed to: $M(q)$ is positive semi-definite and $\xi^T M(q)\xi > 0$ if $G(q)\xi = 0$ and $\xi \neq 0$, [2].)

We compare 4th and 5th order half-explicit methods of Type A and B that are applied with fixed stepsize to the model equations of the Seven body mechanism with $n_y = 2 \cdot 7 = 14$ and $n_z = 6$ ([10, p. 531ff], FORTRAN sources available from CWI Amsterdam through WorldWideWeb at `http://www.cwi.nl/projects/IVPtestset.html`). In this test we are interested in the discretization error and do not take care that the index-3 constraints $\hat{g}(q) = 0$ are satisfied, i. e. we ignore the drift-off effect. In Figure 3 the number of systems (4) that have to be solved during integration is plotted vs. the error at the end of the time interval $[0, 0.03]$. We compare the Type A methods HEM4 ("$*$") and HEM5 ("$\times$") ([3, 2]) with

- the Type B method of order $q = 4$ that is constructed in Example 2c ("$\circ$", $c_2 := \frac{2}{5}$) and
- the Type B method HEDOP5 ("$+$") of order $q = 5$ that is based on the explicit Runge–Kutta method of Dormand and Prince (Example 2d with $c_8 = \frac{1}{2}$, $a_{81} = \frac{84685}{872448}$, $a_{82} = 0$, $a_{83} = \frac{1013965}{2528736}$, $a_{84} = -\frac{38585}{436224}$, $a_{85} = \frac{2072547}{15413248}$, $a_{86} = -\frac{42163}{381696}$, $a_{87} = \frac{17}{256}$; this choice of $c_8 \in [0,1]$ gives rational parameters a_{8i} and a nearly maximum value of $|a_{87}|$).

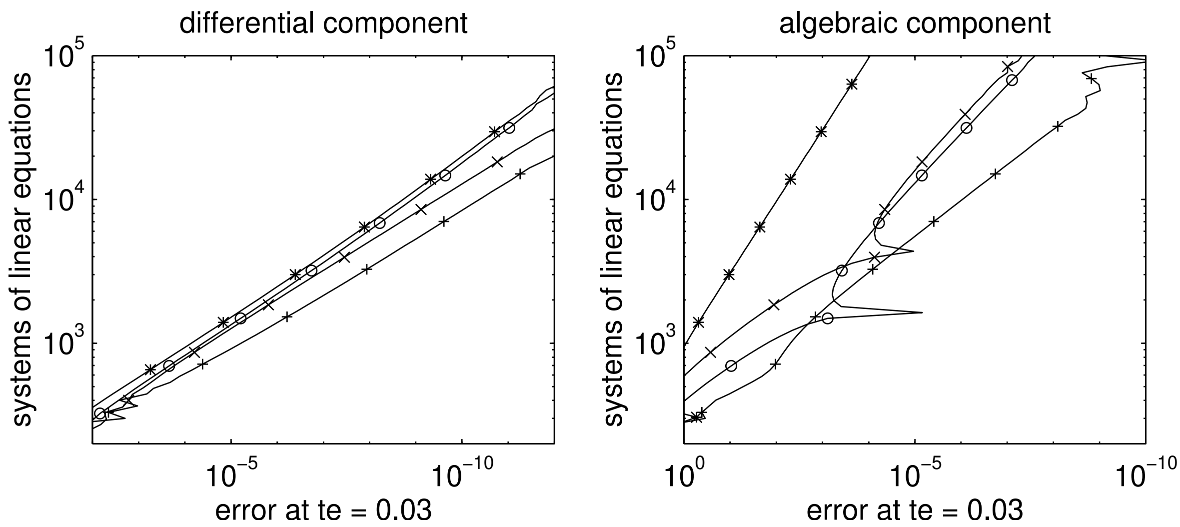


Figure 3: Global discretization error of half-explicit Runge–Kutta methods.

The left plot of Figure 3 shows the error in the first differential component $q_1 = \beta$, the right plot the error in λ_1. For the differential component we see the expected orders of convergence $q = 4$ $(*, \circ)$ and $q = 5$ $(\times, +)$, respectively. Both 4th order methods need approximately the same numerical effort. The 5th order method HEDOP5 is superior to HEM5, it benefits both from the smaller number of stages and from the very small error coefficients of the underlying explicit Runge–Kutta method of Dormand and Prince. While for larger stepsizes the differences between 4th and 5th order methods are negligible the 5th order methods are more efficient to get a numerical solution of high precision.

The global error in the algebraic components is of order $q - 1$ for the Type B methods $(\circ, +)$, of order $r = 2$ for HEM4 $(*)$ and of order $r = 3$ for HEM5 $(\times)$. For two of the methods $(\times, \circ)$ the global error in λ_1 changes sign at some fixed stepsize h^*. This gives an irregular behaviour in the absolute value of the global error that is used as abscissa in the plots of Figure 3.

5 Summary

We constructed a new class of half-explicit methods ("Type B methods") for differential-algebraic systems of index 2 that is based on the half-explicit Runge–Kutta methods of Hairer et al. ([8], "Type A methods"). The main difference to half-explicit Runge–Kutta methods that are known from the literature is the substitution of the first stage by an explicit Runge–Kutta stage. With this modification well known high order explicit Runge–Kutta methods for ODEs can be extended to half-explicit methods for differential-algebraic systems without any order reduction. Thus Type B methods have the potential to combine the advantages of very efficient ODE methods (small error coefficients, dense output formulae etc.) with the advantages of half-explicit methods for differential-algebraic systems. Furthermore the construction of higher order methods is simplified since most of the order conditions coincide with classical order conditions for the underlying explicit Runge–Kutta method.

These new discretization methods are convergent with the same order as the underlying explicit Runge–Kutta method if the local discretization error in the algebraic components is sufficiently small and a contractivity condition is satisfied. For the approximation of the algebraic components the methods can be extended by additional stages. Up to order $q = 3$ we do not need such additional stages, for appropriate methods of order $q \leq 5$ one additional stage is sufficient. For a Type B method that converges with order q in the differential components the global error in the algebraic components has order $q - 1$.

Based on explicit Runge–Kutta methods from the literature Type B methods up to order $q = 5$ are constructed. The 5th order method HEDOP5 that is based on the explicit Runge–Kutta method of Dormand and Prince is seen to be very efficient: it has 7 stages and requires the solution of 6 systems (4) of n_z nonlinear equations per step of integration. In numerical tests for the index-2 formulation of model equations for constrained mechanical systems it is superior to the most efficient Type A methods that are known from the literature (HEM4, HEM5).

Acknowledgement I am grateful to Prof. Dr. K. Strehmel and to J. Wensch (both Halle/S.) for helpful comments on a first version of this manuscript.

References

[1] U.M. Ascher and P. Lin. Sequential regularization methods for higher index DAEs with constraint singularities. I. The linear index-2 case. Technical Report TR–93–24, University of British Columbia, Vancouver, Computer Science Department, 1993, appears in SIAM J. Numer. Anal.

[2] V. Brasey. A half-explicit method of order 5 for solving constrained mechanical systems. *Computing*, 48:191–201, 1992.

[3] V. Brasey and E. Hairer. Half-explicit Runge–Kutta methods for differential-algebraic systems of index 2. *SIAM J. Numer. Anal.*, 30:538–552, 1993.

[4] V. Brasey and E. Hairer. Symmetrized half-explicit methods for constrained mechanical systems. *Appl. Numer. Math.*, 13:23–31, 1993.

[5] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical solution of initial–value problems in differential–algebraic equations.* North-Holland Publ. Co., Amsterdam, 1989.

[6] J.R. Dormand and P.J. Prince. A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.*, 6:19–26, 1980.

[7] C.W. Gear, G.K. Gupta, and B. Leimkuhler. Automatic integration of Euler–Lagrange equations with constraints. *J. Comp. Appl. Math.*, 12&13:77–90, 1985.

[8] E. Hairer, Ch. Lubich, and M. Roche. *The numerical solution of differential-algebraic systems by Runge–Kutta methods.* Lecture Notes in Mathematics, 1409. Springer-Verlag, Berlin Heidelberg New York, 1989.

[9] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations.I. Nonstiff Problems.* Springer-Verlag, Berlin Heidelberg New York, 2nd edition, 1993.

[10] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations.II. Stiff and Differential-Algebraic Problems.* Springer-Verlag, Berlin Heidelberg New York, 1991.

[11] Ch. Lubich. h^2-extrapolation methods for differential-algebraic systems of index 2. *Impact Comput. Sc. Eng.*, 1:260–268, 1989.

[12] Ch. Lubich and M. Roche. Rosenbrock methods for differential-algebraic systems with solution dependent singular matrix multiplying the derivative. *Computing*, 43:325–342, 1990.

[13] A. Ostermann. A class of half-explicit Runge–Kutta methods for differential-algebraic systems of index 3. *Appl. Numer. Math.*, 13:165–179, 1993.

[14] St. Wolfram. *Mathematica. A System for Doing Mathematics by Computer.* Addison Wesley, Redwood City, 2nd edition, 1991.