

Numerieke methoden
voor stelsels
gewone differentiaalvergelijkingen

Prof. Dr. Marnix Van Daele

Deel III

Runge–Kutta-methoden

Hoofdstuk 9

Expliciete methoden

9.1 Vierdeordemethoden

In het vorige hoofdstuk hebben we de voorwaarden opgesteld opdat een RKM een gegeven orde zou hebben. We komen nu aan het probleem om oplossingen te vinden die aan deze voorwaarden voldoen. In dit hoofdstuk bekijken we enkel de expliciete methoden. Een eerste vraag die men zich kan stellen is : *welke orde kan bereikt worden met een expliciete s-traps methode ?* Om een antwoord te kunnen formuleren, moeten we opnieuw een beroep doen op enkele stellingen die opgesteld en bewezen werden door Butcher.

Stelling 9.1.1 *Als U en V twee 3×3 matrices zijn, zodat*

$$UV = [w_{ij}] = \begin{bmatrix} w_{11} & w_{12} & 0 \\ w_{21} & w_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad (9.1)$$

waarbij $w_{11}w_{22} \neq w_{21}w_{12}$, dan is ofwel de laatste rij van U de nul-rijvector ofwel is de laatste kolom van V de nul-kolomvector, m.a.w. als $e_3 = (0, 0, 1)^T$, dan is ofwel $Ve_3 = 0$, ofwel $U^T e_3 = 0$. \square

Bewijs. Als $\det U \neq 0$ dan impliceert $UVe_3 = 0$ dat $Ve_3 = 0$. Als $\det U = 0$ dan bestaat er een vector $x = (x_1, x_2, x_3)^T \neq 0$ zodat $U^T x = 0$, waardoor ook $V^T U^T x = 0$. Maar (9.1) impliceert dan dat x een veelvoud is van e_3 . Analoge resultaten volgen vertrekkend van de matrix V . \blacksquare

Een tweede stelling levert een bovengrens voor de orde van een expliciete s-traps methode.

Stelling 9.1.2 *De orde van een expliciete s-traps RKM is hoogstens s .* \square

Bewijs. Veronderstel dat de s -traps methode een orde p bezit en beschouw de p -de orde boom $t = [_{p-1}\tau]_{p-1}$. Uit Definitie 8.3.1 volgt dat $\gamma(t) = p!$ en uit Opmerking 8.4.1 dat

$$\psi(t) = \sum_{i, j_1, j_2, \dots, j_{p-2}} b_i a_{ij_1} a_{j_1 j_2} \cdots a_{j_{p-3}, j_{p-2}} c_{j_{p-2}}.$$

Vermits de beschouwde RKMn expliciet zijn, is $a_{ij} = 0$ voor $j \geq i$ en dus volgt er dat $\psi(t) = 0$ tenzij er een rij $i, j_1, j_2, \dots, j_{p-2}$ bestaat van gehele getallen $1, 2, \dots, s$ zodat

$$i > j_1 > j_2 > \cdots > j_{p-2} > 1.$$

Merk wel op dat $j_{p-2} = 1$ niet kan vermits dan $c_{j_{p-2}} = c_1 = 0$. De niet-verdwijnde sequentie met de kleinste i -waarde is aldus

$$p > p-1 > p-2 > \cdots > 2 > 1,$$

Vermits $i \leq s$, moet $p = i \leq s$. ■

Het ligt voor de hand de vraag te stellen of de orde $p = s$ kan bereikt worden voor alle s . In paragraaf 7.3 hebben we de algemene Butcher-matrices afgeleid voor s -traps methoden met $1 \leq s \leq 3$. We hebben daar vastgesteld dat er voor die s -waarden inderdaad methoden bestaan met orde $p = s$.

Laten we hier starten met de studie van de expliciete vier-steps methoden. De bijhorende Butcher-matrix is:

$$\begin{array}{c|cccc} 0 & & & & \\ c_2 & a_{21} & & & a_{21} = c_2 \\ c_3 & a_{31} & a_{32} & & a_{31} + a_{32} = c_3 \\ c_4 & a_{41} & a_{42} & a_{43} & a_{41} + a_{42} + a_{43} = c_4 \\ \hline & b_1 & b_2 & b_3 & b_4 \end{array}$$

De ordevoorwaarden zijn, net als in Voorbeeld 8.4.3, zeer omslachtig om uit te schrijven. Daarom gebruiken we vanaf hier een somnotatie, waarbij we afspreken dat alle sommen van 1 t.e.m. het trapgetal s lopen. Deze conventie in acht nemend kunnen we de acht vierdeordevoorwaarden als volgt noteren:

$$\left. \begin{array}{l} (1) \quad \sum_i b_i = 1 \\ (2) \quad \sum_i b_i c_i = \frac{1}{2} \\ (3) \quad \sum_i b_i c_i^2 = \frac{1}{3} \\ (4) \quad \sum_{i,j} b_i a_{ij} c_j = \frac{1}{6} \\ (5) \quad \sum_i b_i c_i^3 = \frac{1}{4} \\ (6) \quad \sum_{i,j} b_i c_i a_{ij} c_j = \frac{1}{8} \\ (7) \quad \sum_{i,j} b_i a_{ij} c_j^2 = \frac{1}{12} \\ (8) \quad \sum_{i,j,k} b_i a_{ij} a_{jk} c_k = \frac{1}{24}. \end{array} \right\} \quad (9.2)$$

Alle oplossingen vinden van dit stelsel niet-lineaire vergelijkingen is beslist niet eenvoudig. Maar opnieuw kunnen we gebruik maken van ideeën van Butcher om de taak te verlichten. We gebruiken Stelling 9.1.1 en passen het toe op de matrices

$$U = \begin{bmatrix} c_2 & c_3 & c_4 \\ c_2^2 & c_3^2 & c_4^2 \\ \lambda_2 & \lambda_3 & \lambda_4 \end{bmatrix}, \quad V = \begin{bmatrix} b_2 & b_2 c_2 & \mu_2 - \beta_2 \\ b_3 & b_3 c_3 & \mu_3 - \beta_3 \\ b_4 & b_4 c_4 & \mu_4 - \beta_4 \end{bmatrix},$$

waarbij

$$\left. \begin{aligned} \lambda_i &= \sum_j a_{ij} c_j - \frac{1}{2} c_i^2 \\ \mu_j &= b_j (1 - c_j) \\ \beta_j &= \sum_i b_i a_{ij}, \end{aligned} \right\} i, j = 2, 3, 4. \quad (9.3)$$

Als we $UV = [w_{ij}]$ stellen en gebruik maken van de voorwaarden (2), (3) en (5) vinden we dat $w_{11} = \frac{1}{2}$, $w_{12} = \frac{1}{3} = w_{21}$, $w_{22} = \frac{1}{4}$. Evenzo zien we dat

$$\begin{aligned} w_{13} &= \sum_j c_j \left[b_j (1 - c_j) - \sum_i b_i a_{ij} \right] = \frac{1}{2} - \frac{1}{3} - \frac{1}{6} = 0 \text{ wegens (2), (3) en (4)}, \\ w_{23} &= \sum_j c_j^2 \left[b_j (1 - c_j) - \sum_i b_i a_{ij} \right] = \frac{1}{3} - \frac{1}{4} - \frac{1}{12} = 0 \text{ wegens (3), (5) en (7)}, \\ w_{31} &= \sum_i b_i \left[\sum_j a_{ij} c_j - \frac{1}{2} c_i^2 \right] = \frac{1}{6} - \frac{1}{2} \frac{1}{3} = 0 \text{ wegens (4) en (3)}, \\ w_{32} &= \sum_i b_i c_i \left[\sum_j a_{ij} c_j - \frac{1}{2} c_i^2 \right] = \frac{1}{8} - \frac{1}{2} \frac{1}{4} = 0 \text{ wegens (6) en (5)}, \\ w_{33} &= \sum_i \left[\sum_j a_{ij} c_j - \frac{1}{2} c_i^2 \right] \left[b_i (1 - c_i) - \sum_j b_j a_{ji} \right] \\ &= w_{31} - w_{32} - \frac{1}{24} + \frac{1}{2} \frac{1}{12} = 0 \text{ wegens (8) en (7)}. \end{aligned}$$

Aldus is

$$UV = [w_{ij}] = \begin{bmatrix} \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{4} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Vermits $w_{11} w_{22} - w_{21} w_{12} = \frac{1}{8} - \frac{1}{9} \neq 0$ zijn de hypotheses van Stelling 9.1.1 voldaan en geldt dat

$$\text{ofwel } \lambda_i = 0, \quad i = 2, 3, 4 \quad \text{ofwel } \mu_j = \beta_j, \quad j = 2, 3, 4.$$

We tonen nu aan dat de eerste geciteerde voorwaarde onmogelijk kan voldaan zijn. Indien

$$\lambda_2 = \sum_j a_{2j} c_j - \frac{1}{2} c_2^2 = 0$$

dan moet, rekening houdend met $c_1 = 0$ en $a_{2j} = 0$, $j = 2, 3, 4$ noodzakelijk ook $c_2 = 0$. Voor een expliciete methode echter wordt vergelijking (8) van (9.2)

$$b_4 a_{43} a_{32} c_2 = \frac{1}{24}, \quad (9.4)$$

waaruit $c_2 \neq 0$. Het tweede alternatief, $\mu_j = \beta_j$, $j = 2, 3, 4$ moet daarom geldig zijn. Uit (9.3) volgt dan dat

$$\sum_i b_i a_{ij} = b_j (1 - c_j), \quad j = 2, 3, 4. \quad (9.5)$$

Dit heeft twee belangrijke gevolgen. Vooreerst is

$$\begin{aligned} \sum_{i,j} b_i a_{ij} c_j &= \sum_j b_j (1 - c_j) c_j = \frac{1}{2} - \frac{1}{3} = \frac{1}{6} \text{ wegens (2) en (3),} \\ \sum_{i,j} b_i a_{ij} c_j^2 &= \sum_j b_j (1 - c_j) c_j^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12} \text{ wegens (3) en (5),} \\ \sum_{i,j,k} b_i a_{ij} a_{jk} c_k &= \sum_{j,k} b_j (1 - c_j) a_{jk} c_k = \frac{1}{6} - \frac{1}{8} = \frac{1}{24} \text{ wegens (4) en (6).} \end{aligned}$$

De voorwaarden (4), (7) en (8) zijn aldus automatisch voldaan als de overblijvende voorwaarden voldaan zijn; ze kunnen dus geïgnoreerd worden.

Het tweede gevolg is afkomstig uit (9.5) voor $j = 4$ en de vaststelling dat

$$\sum_i b_i a_{i4} = 0$$

wegens $a_{ij} = 0$ als $j \geq i$. Hieruit volgt $b_4(1 - c_4) = 0$ en vermits door (9.4) b_4 niet nul kan zijn volgt hieruit dat $c_4 = 1$. We bekomen aldus dat voor alle vierde-orde 4-traps ERKMn $c_4 = 1$ is.

Als we de bovenstaande eigenschappen in acht nemen, herleidt het stelsel (9.2) zich tot het volgende equivalente stelsel :

$$\left. \begin{array}{l}
 (1) \quad b_1 + b_2 + b_3 + b_4 = 1 \\
 (2) \quad b_2 c_2 + b_3 c_3 + b_4 = \frac{1}{2} \\
 (3) \quad b_2 c_2^2 + b_3 c_3^2 + b_4 = \frac{1}{3} \\
 (4) \quad b_2 c_2^3 + b_3 c_3^3 + b_4 = \frac{1}{4} \\
 (5) \quad b_3 c_3 a_{32} c_2 + b_4 (a_{42} c_2 + a_{43} c_3) = \frac{1}{8} \\
 (6) \quad b_3 a_{32} + b_4 a_{42} = b_2 (1 - c_2) \\
 (7) \quad b_4 a_{43} = b_3 (1 - c_3) \\
 (8) \quad c_4 = 1,
 \end{array} \right\} \quad (9.6)$$

waarbij we bovendien weten dat

$$b_4 a_{43} a_{32} c_2 \neq 0. \quad (9.7)$$

We proberen nu de oplossingen van dit iets eenvoudiger stelsel te zoeken.

Merk vooreerst op dat (9.6), (1)-(4) en (8) aangeven dat b_i en c_i de coëfficiënten zijn van een vierdeordekwadratuurformule met $c_1 = 0$ en $c_4 = 1$. Als we dit stelsel van vier vergelijkingen oplossen, dan kunnen we 4 families van oplossingen onderscheiden die aanleiding geven tot een oplossing van het stelsel (9.2) : één waarbij de vier c_i -waarden verschillend zijn en 3 waarbij twee van de vier c_i -waarden samenvallen.

(1) $0, c_2, c_3$ en 1 zijn allen verschillend.

$$\begin{aligned}
 b_2 &= \frac{2c_3 - 1}{12c_2(1 - c_2)(c_3 - c_2)}, & b_3 &= \frac{1 - 2c_2}{12c_3(1 - c_3)(c_3 - c_2)}, \\
 b_4 &= \frac{6c_2c_3 - 4(c_2 + c_3) + 3}{12(1 - c_2)(1 - c_3)}, & b_1 &= 1 - b_2 - b_3 - b_4.
 \end{aligned}$$

Omdat noch b_3 , noch b_4 0 kunnen zijn, moet bovendien geëist worden dat $c_2 \neq \frac{1}{2}$ en $6c_2c_3 - 4(c_2 + c_3) + 3 \neq 0$.

(2) $c_2 = \frac{1}{2}$ en $c_3 = 0$: $b_1 = \frac{1}{6} - b_3$, $b_2 = \frac{2}{3}$, $b_3 \neq 0$, $b_4 = \frac{1}{6}$.

(3) $c_2 = \frac{1}{2}$ en $c_3 = \frac{1}{2}$: $b_1 = \frac{1}{6}$, $b_2 = \frac{4}{6} - b_3$, $b_3 \neq 0$, $b_4 = \frac{1}{6}$.

(4) $c_2 = 1$ en $c_3 = \frac{1}{2}$: $b_1 = \frac{1}{6}$, $b_2 = \frac{1}{6} - b_4$, $b_3 = \frac{4}{6}$, $b_4 \neq 0$.

Eenmaal b_i en c_i vastgelegd, bekommen we a_{43} uit (9.6), (7). Vervolgens vormen (9.6), (5)-(6) een lineair stelsel van twee vergelijkingen in a_{32} en a_{42} . De determinant van dit stelsel is

$$\begin{vmatrix} b_3 & b_4 \\ b_3 c_3 c_2 & b_4 c_2 \end{vmatrix} = b_3 b_4 c_2 (1 - c_3),$$

en is omwille van (9.7) en (9.6), (7) verschillend van nul. We bekommen finaal a_{21} , a_{31} en a_{41} uit de rijsum-voorwaarde.

Uit het bovenstaande blijkt dat de algemene oplossing bestaat uit een twee-parameter familie en drie één-parameter families. Twee bijzondere keuzen van parameterwaarden van Kutta zijn bijzonder populair geworden:

1. ‘De’ Runge–Kutta methode : geval (3) met $b_3 = \frac{1}{3}$,
2. de 3/8 regel : geval (1) met $c_2 = \frac{1}{3}$ en $c_3 = \frac{2}{3}$.

De corresponderende Butcher-matrices zijn weergegeven in Tabel 9.1 en Tabel 9.2. Beide methoden veralgemenen klassieke kwadratuurformules. De eerste is de meer populaire, de tweede is de meer nauwkeurige.

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

Tabel 9.1: ‘De’ Runge–Kutta-methode

0				
$\frac{1}{3}$	$\frac{1}{3}$			
$\frac{2}{3}$	$-\frac{1}{3}$	1		
1	1	-1	1	
	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Tabel 9.2: 3/8 regel

waarbij λ_i en β_i gegeven zijn door (9.3) maar nu voor $i = 2, 3, 4, 5$ en

$$\nu_j = \frac{1}{2}\beta_j(1 - c_j) - \sum_i \beta_i a_{ij}, \quad j = 2, 3, 4, 5.$$

Door gebruik te maken van de ordevoorwaarden (9.2) en (9.8), tesamen met het feit dat $\beta_5 = 0 = \nu_5$, vinden we dat

$$UV = [w_{ij}] = \begin{bmatrix} \frac{1}{6} & \frac{1}{12} & 0 \\ \frac{1}{12} & \frac{1}{20} & 0 \\ 0 & 0 & 0 \end{bmatrix}. \quad (9.9)$$

Vermits

$$w_{11}w_{22} - w_{21}w_{12} = \frac{1}{120} - \frac{1}{144} \neq 0$$

zijn de hypotheses van Stelling 9.1.1 voldaan en moet ofwel $\lambda_i = 0$, $i = 2, 3, 4$, ofwel $\nu_j = 0$, $j = 2, 3, 4$. Het eerste is onmogelijk : omwille van exact hetzelfde argument als vroeger impliceert $\lambda_2 = 0$ dat $c_2 = 0$ en vermits voorwaarde (9.8), (17) zich herleidt tot

$$b_5 a_{54} a_{43} a_{32} c_2 = \frac{1}{120}, \quad (9.10)$$

volgt hieruit dat $c_2 \neq 0$. Aldus moet $\nu_j = 0$, $j = 2, 3, 4$ waar zijn; maar $\nu_4 = 0$ impliceert

$$\beta_4(1 - c_4) = 2 \sum \beta_i a_{i4} = 2\beta_5 a_{54} = 0,$$

vermits $\beta_5 = 0$ (zie (9.3)). Nu volgt ook uit (9.3) dat $\beta_4 = b_5 a_{54} \neq 0$ omwille van (9.10). Hieruit volgt dat $c_4 = 1$.

We passen Stelling 9.1.1 opnieuw toe maar nu met

$$U = \begin{bmatrix} c_2 & c_3 & c_5 \\ c_2^2 & c_3^2 & c_5^2 \\ \lambda_2 & \lambda_3 & \lambda_5 \end{bmatrix}, \quad V = \begin{bmatrix} \mu_2 & \mu_2 c_2 & (\mu_2 - \beta_2)(1 - c_2) \\ \mu_3 & \mu_3 c_3 & (\mu_3 - \beta_3)(1 - c_3) \\ \mu_5 & \mu_5 c_5 & (\mu_5 - \beta_5)(1 - c_5) \end{bmatrix},$$

waarbij λ_i , μ_j en β_j gegeven zijn door (9.3) voor $i, j = 2, 3, 4, 5$. Gebruik makend van de ordevoorwaarden en het feit dat $c_4 = 1$ verkrijgen we :

$$\mu_4 = b_4(1 - c_4) = 0 \quad \text{en} \quad (\mu_4 - \beta_4)(1 - c_4) = 0.$$

We vinden bovendien dat UV ook de waarde (9.9) bezit. Vermits $c_2 \neq 0$ is $\lambda_2 \neq 0$, zodat de laatste kolom uit UV gelijk is aan 0, waaruit volgt $(\mu_5 - \beta_5)(1 - c_5) = 0$. Vermits $\beta_5 = 0$ bekomen we uit (9.3) dat $\mu_5(1 - c_5) = b_5(1 - c_5)^2 = 0$ en vermits $b_5 \neq 0$ omwille van (9.10), volgt hieruit dat $c_5 = 1$. We hebben dus tot nu toe gevonden dat $c_4 = c_5 = 1$. Beschouwen we nu

$$\sum_{i,j,k} b_i(1 - c_i) a_{ij} a_{jk} c_k = b_5(1 - c_5) \sum_{j,k} a_{5j} a_{jk} c_k + b_4(1 - c_4) a_{43} a_{32} c_2 = 0.$$

Maar de ordevoorwaarden (9.2), (8) en (9.8), (12) leveren dat

$$\sum_{i,j,k} b_i (1 - c_i) a_{ij} a_{jk} c_k = \sum_{i,j,k} b_i a_{ij} a_{jk} c_k - \sum_{i,j,k} b_i c_i a_{ij} a_{jk} c_k = \frac{1}{24} - \frac{1}{30} = \frac{1}{120}.$$

We komen dus tot een contradictie, wat meteen de stelling bewijst. ■

We kunnen deze stelling uitbreiden om aan te tonen dat geen p -traps expliciete methode bestaat van orde p voor $p \geq 5$. Andere zeer gelijkaardige stellingen verbonden met hogere-orde methoden werden ook bewezen. We geven de formulering van deze stellingen, maar de bewijzen vallen buiten het kader van deze nota's.

Stelling 9.2.2 *Er bestaat geen p -traps ERKM van orde $p \geq 5$.* □

Stelling 9.2.3 *Er bestaat geen $p + 1$ -traps ERKM van orde $p \geq 7$.* □

Stelling 9.2.4 *Er bestaat geen $p + 2$ -traps ERKM van orde $p \geq 8$.* □

De algemene vraag welke orde p bereikt kan worden door een s -traps ERKM is nog steeds open. Het volgende is bekend :

p	1	2	3	4	5	6	7	8	9	10
minimale s	1	2	3	4	6	7	9	11	$12 \leq s \leq 17$	$13 \leq s \leq 17$

De constructie van hogere-orde methoden is relatief gecompliceerd. Opnieuw zijn het boek van Butcher en het boek van Hairer *et al.* goede referentiewerken.

9.3 Expliciete methoden – schatting van de LAF

Zoals in paragraaf 1.2.1 opgemerkt is de behandeling van de LAF bij RKMn niet zo uitgebreid ontwikkeld als bvb. bij LMMn.

Een eerste techniek die we wensen te bespreken is *Richardson-extrapolatie*, ook bekend als *the deferred approach to the limit*. Veronderstel dat we gebruik maken van een RKM van orde p om een numerieke oplossing y_{n+1} te bekomen in x_{n+1} . Onder de veronderstelling $y_n = y(x_n)$ volgt uit (8.25) dat de LAF T_{n+1} geschreven kan worden als

$$T_{n+1} = y(x_{n+1}) - \tilde{y}_{n+1} = \Psi(y(x_n)) h^{p+1} + \mathcal{O}(h^{p+2}), \quad (9.11)$$

waarbij $\Psi(y(x_n))$ een functie is van de elementaire differentiaal van orde $p + 1$, geëvalueerd bij $y(x_n)$. Laat ons nu een tweede numerieke oplossing berekenen bij x_{n+1} door dezelfde methode toe te passen met staplengte $2h$, maar vertrekkend bij x_{n-1} . We noteren de aldus bekomen oplossing \tilde{z}_{n+1} , (de tilde wijst op de veronderstelling $y_{n-1} = y(x_{n-1})$). We kunnen dan schrijven

$$\begin{aligned} y(x_{n+1}) - \tilde{z}_{n+1} &= \Psi(y(x_{n-1})) (2h)^{p+1} + \mathcal{O}(h^{p+2}) \\ &= \Psi(y(x_n)) (2h)^{p+1} + \mathcal{O}(h^{p+2}), \end{aligned} \quad (9.12)$$

waarbij we $y(x_{n-1})$ ontwikkelden rond x_n . Door (9.11) af te trekken van (9.12) bekomen we

$$(2^{p+1} - 1) h^{p+1} \Psi(y(x_n)) = \tilde{y}_{n+1} - \tilde{z}_{n+1} + \mathcal{O}(h^{p+2}),$$

waaruit, rekening houdend met (9.11), de volgende schatting volgt voor de PLAF :

$$\text{PLAF} = \frac{\tilde{y}_{n+1} - \tilde{z}_{n+1}}{2^{p+1} - 1}. \quad (9.13)$$

Deze schatting werkt goed in de praktijk en kan succesvol gebruikt worden om de staplengte te controleren, maar is zeer duur in gebruik; als de expliciete RKM s trappen bezit, zijn in het algemeen $s - 1$ bijkomende functie-evaluaties nodig (k_1 werd voor x_{n-1} reeds vroeger berekend).

Deze Richardson-extrapolatietechniek laat tevens toe *automatische controle van de stapgrootte* uit te voeren. Wanneer een stapgrootte h bij de start gekozen is, kan het programma de twee bovengeciteerde berekeningen uitvoeren. De fout volgt dan uit (9.13), nl.

$$\text{fout} = \frac{1}{2^{p+1} - 1} \max_{i=1,2,\dots,m} \frac{\|{}^i\tilde{y}_{n+1} - {}^i\tilde{z}_{n+1}\|}{d_i}.$$

Hierin is d_i een schaalfactor die 1 kan gekozen worden wanneer men in de absolute fouten is geïnteresseerd of daarentegen gelijk aan $\|{}^i\tilde{y}_{n+1}\|$ kan gekozen worden indien men de relatieve fouten onderzoekt. In bepaalde computercodes vindt men soms ook andere keuzes.

Deze fout wordt vergeleken met de gewenste tolerantie tol , d.i. we wensen dat

$$\begin{aligned} \text{fout} &= \beta \text{tol}, & 0 \leq \beta \leq 1, \\ &\approx Ch^{p+1} & \text{wegens (9.11)} \end{aligned} \quad (9.14)$$

De maximale staplengte h_{\max} die we kunnen voorstellen voor een p -de orde methode produceert een foutschatting fout_{\max} , die voldoet aan:

$$\text{fout}_{\max} = \text{tol} \approx Ch_{\max}^{p+1}. \quad (9.15)$$

Uit (9.14) en (9.15) volgt dan dat

$$\beta \approx \left(\frac{h}{h_{\max}} \right)^{p+1},$$

waaruit

$$h_{\max} \approx h \left(\frac{\text{tol}}{\text{fout}} \right)^{1/(p+1)}. \quad (9.16)$$

Er moet hier wel voorzichtig gehandeld worden om dit procédé op een geschikte wijze in een code te verwerken; normaal wordt (9.16) vermenigvuldigd met een veiligheidsfactor fac , nl. $\text{fac} = 0.8, 0.9 \dots$, zodat de fout met een hoge waarschijnlijkheidsgraad aanvaardbaar blijft bij de volgende stap. Verder wordt niet toegelaten dat h te snel toeneemt of afneemt, m.a.w. voor de nieuwe stap introduceert men :

$$h_{\text{nieuw}} = h \min(\text{facmax}, \max(\text{facmin}, \text{fac} \left(\frac{\text{tol}}{\text{fout}} \right)^{1/p+1})), \quad (9.17)$$

met facmax en facmin resp. de maximaal en minimaal toelaatbare toenamefactoren. Als $\text{fout} \leq \text{tol}$ wordt de oplossing geaccepteerd, d.w.z. \tilde{y}_{n+1} wordt aanvaard en er wordt een nieuwe stap gezet met h_{nieuw} als staplengte. Als $\text{fout} \geq \text{tol}$ worden de berekeningen in het punt x_{n+1} herhaald, maar nu met een staplengte h_{nieuw} .

Een andere techniek voor de bepaling van de LAF en de hieruit vloeiende staplengte controle, als functie van de reeds berekende k_i waarden, volgt uit een idee van Merson in 1957. De methode van Merson wordt gedefinieerd door de Butcher-matrix in Tabel 9.3.

0					
$\frac{1}{3}$	$\frac{1}{3}$				
$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$			
$\frac{1}{2}$	$\frac{1}{8}$	0	$\frac{3}{8}$		
1	$\frac{1}{2}$	0	$-\frac{3}{2}$	2	
$\frac{1}{6}$	0	0	$\frac{2}{3}$	$\frac{1}{6}$	

Tabel 9.3: De methode van Merson

Dit is een 5-traps methode; men kan gemakkelijk natrekken dat het een vierdeordemethode is. Merk op dat tengevolge van Stelling 9.2.2 orde 5 niet mogelijk is. Merson stelde voor de PLAF te schatten door

$$\frac{h}{30} (-2k_1 + 9k_3 - 8k_4 + k_5). \quad (9.18)$$

Als dit een geldige schatting voor de PLAF zou zijn, dan zou het toevoegen van (9.18) aan de waarde van y_{n+1} , die gegeven wordt door de methode uit Tabel 9.3, aanleiding geven tot een vijfdeordemethode waarvoor c en A zouden gegeven zijn door Tabel 9.3 en b^T zou zijn:

$$\left[\frac{1}{6}, 0, 0, \frac{2}{3}, \frac{1}{6} \right] + \left[-\frac{1}{15}, 0, \frac{3}{10}, -\frac{4}{15}, \frac{1}{30} \right] = \left[\frac{1}{10}, 0, \frac{3}{10}, \frac{2}{5}, \frac{1}{5} \right]. \quad (9.19)$$

Uiteraard kan dit niet omwille van Stelling 9.2.2 en we moeten dan ook besluiten dat de schatting (9.18) niet correct is. Men kan inderdaad gemakkelijk natrekken dat de 5-traps methode bestaande uit Tabel 9.3 en de b^T , aangepast zoals hierboven, slechts orde 3 bezit; nochtans bezit deze methode orde vijf in het bijzondere geval dat het differentiaalstelsel lineair is met constante coëfficiënten. Alhoewel deze methode van Merson een belangrijke rol gespeeld heeft in de uitstippeling van verdere ontwikkelingen, is het nodig erop te wijzen die niet te gebruiken voor algemene problemen.

De essentie van het Merson-idee bestaat erin RKMn van orde p en orde $p + 1$ af te leiden, die dezelfde verzameling van vectoren $\{k_i\}$ delen. Dit proces is bekend als *inbedden*. Om inbedde methoden te kunnen voorstellen, modificeren we de Butcher-matrix naar de volgende vorm:

De notatie moet als volgt geïnterpreteerd worden : de methode, gedefinieerd door c , A en b^T is van orde p en degene gedefinieerd door c , A en \hat{b}^T bezit orde $p + 1$. Het verschil tussen de

c	A
	b^T
	\hat{b}^T
	E^T

Tabel 9.4:

waarden van y_{n+1} gegenereerd door deze twee methoden is dan een schatting voor de lokale afknottingsfout. De vector E^T is $\hat{b}^T - b^T$, zodat de fout-schatting gegeven wordt door

$$h \sum_{i=1}^s E_i k_i, \quad \text{waarbij } E^T = [E_1, E_2, \dots, E_s].$$

Men heeft de gewoonte aan een ingebedde methode het etiket $(p, p + 1)$ te hechten. Merk op dat de oplossing voor y_{n+1} gegeven door de p -de orde methode gebruikt wordt als de initiale waarde voor de volgende stap, zodat de methode effectief orde p bezit. Men gebruikt soms de $(p + 1)$ -de orde waarde voor y_{n+1} als de initiale waarde voor de volgende stap, in welk geval de methode orde $p + 1$ bezit; men beschrijft dergelijke methode dan als $(p + 1, p)$. De techniek waarbij een hogere-orde methode wordt aangewend, waarbij m.a.w. de foutterm van de lagere-orde methode bij de oplossing wordt gevoegd, staat in de literatuur bekend als *lokale extrapolatie*.

Voorbeeld 9.3.1

Laat ons hier ter verduidelijking van bovenstaande uiteenzetting in verband met ingebedde methoden een zgn. tweede-orde Fehlberg-methode afleiden. De corresponderende uitgebreide Butcher-matrix wordt gegeven in Tabel 9.5. De re-

0			
c_2	a_{21}		
c_3	a_{31}	a_{32}	
	b_1	b_2	b_3
	\hat{b}_1	\hat{b}_2	\hat{b}_3
	E_1	E_2	E_3

Tabel 9.5:

spectievelijke tweede-orde en derde-orde voorwaarden waaraan de optredende coëfficiënten moeten voldoen zijn de volgende

$$\begin{aligned}
 b_1 + b_2 + b_3 &= 1 & \hat{b}_1 + \hat{b}_2 + \hat{b}_3 &= 1 \\
 b_2 c_2 + b_3 c_3 &= \frac{1}{2} & \hat{b}_2 c_2 + \hat{b}_3 c_3 &= \frac{1}{2} \\
 & & \hat{b}_2 c_2^2 + \hat{b}_3 c_3^2 &= \frac{1}{3} \\
 & & \hat{b}_3 a_{32} c_2 &= \frac{1}{6}
 \end{aligned} \tag{9.20}$$

We kiezen $c_2 = 1$ en $b_3 = 0$ om uit de eerste twee vergelijkingen $b_2 = b_1 = \frac{1}{2}$ te bekomen. Er blijven dan vier vergelijkingen over voor vijf onbekenden. We kiezen $c_3 = \frac{1}{2}$ zodat we $\hat{b}_1 = \frac{1}{6}$, $\hat{b}_2 = \frac{1}{6}$ en $\hat{b}_3 = \frac{4}{6}$ halen uit de eerste drie vergelijkingen voor de derde-orde methode. Uit de laatste vergelijking volgt dan $a_{32} = \frac{1}{4}$. De resulterende methode wordt hieronder in de vorm van de Butcher-matrix weergegeven; tevens geven we ter informatie een tweede methode van Fehlberg met één bijkomende trap, maar de coëfficiënten worden zo gekozen dat $a_{4i} = b_i$ voor alle i ; daardoor kan de laatste evaluatie van f in een lopende stap hergebruikt worden voor de eerste evaluatie in de volgende stap (zie ook commentaar bij DOPRI(5,4) methode, aan het einde van de paragraaf). \square

0			
1	1		
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	
	$\frac{1}{2}$	$\frac{1}{2}$	0
	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{4}{6}$
	$-\frac{1}{3}$	$-\frac{1}{3}$	$\frac{4}{6}$

Tabel 9.6: RKF2(3)

0				
$\frac{1}{4}$	$\frac{1}{4}$			
$\frac{27}{40}$	$-\frac{189}{800}$	$\frac{729}{800}$		
1	$\frac{214}{891}$	$\frac{1}{33}$	$\frac{650}{891}$	
	$\frac{214}{891}$	$\frac{1}{33}$	$\frac{650}{891}$	0
	$\frac{533}{2106}$	0	$\frac{800}{1053}$	$-\frac{1}{78}$
	$\frac{13}{981}$	$-\frac{1}{33}$	$\frac{350}{11583}$	$-\frac{1}{78}$

Tabel 9.7: RKF2(3)B

In alle computercodes maakt men gebruik van RK-codes van minstens orde 4. Uit de overwegingen in paragraaf 9.2 volgt dat een vierde-orde ingebodde methode minstens uit zes trappen moet opgebouwd zijn. In de jaren zestig zijn verschillende van die methoden opgesteld. Eén van de populairste is de (4,5) methode RKF45 van Fehlberg. In die klasse zijn de coëfficiënten van de methode zo gekozen dat de moduli van de coëfficiënten van de functies

$F(t)$ die optreden in de PLAF (8.25) klein zijn. Zo'n methode wordt *error tuned* genoemd. De gemodificeerde Butcher-matrix voor die (4,5) methode wordt weergegeven in Tabel 9.8.

Merk wel op dat slechts vijf trappen vereist zijn om de oplossing te bekomen, wanneer de foutschatting niet gewenst is. In vele moderne automatische codes gebaseerd op ingebedde RKMn wordt lokale extrapolatie gebruikt. Ook RKF45 wordt soms uitgevoerd als een (5,4) methode, alhoewel ze daar niet voor ontworpen is, vermits de error tuning uitgevoerd werd op de vierdeordemethode en niet op de vijfdeordeformule. Er bestaan Fehlberg-methoden van ordes t.e.m. acht. Al de Fehlberg-methoden met orde groter dan vier vertonen een uitzonderlijk gebrek, dat we kunnen illustreren voor bijvoorbeeld de 8-traps methode, waarvoor de vectoren c^T en E^T gegeven worden door:

$$\begin{aligned} c^T &= \left[0, \frac{1}{6}, \frac{4}{15}, \frac{2}{3}, \frac{4}{5}, 1, 0, 1 \right] \\ E^T &= \left[-\frac{5}{66}, 0, 0, 0, 0, -\frac{5}{66}, \frac{5}{66}, \frac{5}{66} \right]. \end{aligned} \tag{9.21}$$

Veronderstel dat zulk een methode toegepast wordt op een stelsel waarin f slechts afhangt van x . Dan reduceren de vijfde- en de zesdeordemethoden respectievelijk tot

$$y_{n+1} = y_n + h \sum_{j=1}^8 b_j f(x_n + c_j h), \quad \hat{y}_{n+1} = y_n + h \sum_{j=1}^8 \hat{b}_j f(x_n + c_j h)$$

waaruit $\hat{y}_{n+1} - y_{n+1} = h \sum_{j=1}^8 E_j f(x_n + c_j h) = 0$ wegens (9.21).

0						
$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$			
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$		
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	
	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0
	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$
	$\frac{1}{360}$	0	$-\frac{128}{4275}$	$-\frac{2197}{75240}$	$\frac{1}{50}$	$\frac{2}{55}$

Tabel 9.8: De (4,5) methode van Fehlberg

D.w.z. dat beide methoden identisch dezelfde resultaten geven, wanneer $f(x, y) = f(x)$. Bovendien is de foutschatting in alle gevallen nul, wat de realistische grootte van de actuele LAF ook is. Hieruit kan men afleiden dat zulke methoden misleidende resultaten zullen

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
	$\frac{71}{57600}$	0	$-\frac{71}{16695}$	$\frac{71}{1920}$	$-\frac{17253}{339200}$	$\frac{22}{525}$	$-\frac{1}{40}$

Tabel 9.9: De (5,4) methode van Dormand en Prince

geven, wanneer ze toegepast worden op stelsels $y = f(x, y)$, waarin f veel meer afhankelijk is van x dan van y . Alternatieve ingebedde methoden van orde 5 tot 8, die deze moeilijkheid niet opleveren, werden afgeleid door Verner in 1978.

Ingebedde methoden, die in het bijzonder ontworpen zijn voor gebruik met lokale extrapolatie, zijn ontworpen door Dormand en Prince. In deze methoden is het de hogereordeformule die error tuned is en die de oplossing draagt; het verschil tussen de waarden gegeven door de hogere en de lagere orde methoden wordt gebruikt om de staplengte bij te sturen, alhoewel dit in feite niet langer meer een ware schatting van de lokale afknottingsfout is. Eén van de meest populaire is de (5,4) methode, bekend als DOPRI(5,4), waarvan de Butcher-matrix gegeven is in Tabel 9.9.

We houden ons hier wel aan de notatie van Tabel 9.4 zodat de vector \hat{b}^T die de oplossing draagt degene is die start met $\frac{35}{384}$. Bovenstaande methode bezit zeven trappen; nochtans is de laatste rij van A identisch met de vector \hat{b}^T , en we zien uit wat volgt dat de methode op die wijze effectief slechts zes trappen bevat; d.w.z. als we de vectoren k_i geëvalueerd gedurende de stap van x_n naar x_{n+1} aanduiden als k_i^n , dan hebben we dat

$$k_7^n = f(x_n + h, y_n + h \sum_{j=1}^6 a_{7j} k_j^n)$$

$$\text{waaruit } k_1^{n+1} = f(x_{n+1}, y_{n+1}) = f(x_n + h, y_n + h \sum_{j=1}^6 \hat{b}_j k_j^n) = k_7^n.$$

Er is dus duidelijk geen nood om k_1^{n+1} te berekenen. Methoden met deze eigenschap (zie ook RKF3(2)B) worden soms *FSAL* (*First Same As Last*) methoden genoemd. De meest

gebruikte methode voor het ogenblik is de 8(7) methode van Prince en Dormand, bekend als DOPRI8. Deze code maakt tenvolle gebruik van het idee van lokale extrapolatie.