stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE NN 1/73 DECEMBER

H. FIOLET

NUMERIEKE INTEGRATIE VAN DIFFERENTIAALVERGELIJKINGEN DOOR MIDDEL VAN PADÉ-BENADERINGEN

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS (MOS) onderwerpen classificatie schema (1970): 65L05

Voorwoord

Dit rapport beschrijft een methode voor het oplossen van beginwaarde-problemen. Het ALGOL 60 programma is getest op de EL X8 computer van het Mathematisch Centrum te Amsterdam.

De schrijver wil zijn dank betuigen aan Dr. P.J. van der Houwen voor diens waardevolle suggesties.

Inhoud

1. Inleiding

Beschouw beginwaarde-problemen van het type

$$(1.1) \qquad \frac{d\vec{u}}{dt} = \vec{H}(t,\vec{u}) \ , \qquad \vec{u}(t_0) = \vec{u}_0 \ ,$$

waarin $\vec{H}$ een vector-functie is van $\vec{u}$ en de variabele t en $\vec{u}$ een functie is van t. Wanneer het niet moeilijk is om hogere afgeleiden van $\vec{u}$ te evalueren kan men voor het oplossen van (1.1) gebruik maken van Taylormethoden. Omdat de gewone Taylormethoden in het algemeen een klein stabiliteitsgebied hebben zijn ze niet geschikt voor het oplossen van stelsels stijve differentiaalvergelijkingen, die juist worden gekarakteriseerd door een grote spreiding in de eigenwaarden van de Jacobiaan $J = (\partial H_i/\partial u_j)$ van het stelsel. Men kan op verschillende manieren het stabiliteitsgebied van deze methoden vergroten (zie b.v. [2], [3]). In dit rapport construeren we enkele formules, die eveneens van hogere afgeleiden gebruik maken en die, om goede stabiliteitseigenschappen te verkrijgen, zijn gebaseerd op rationale Padé-approximaties. Rationale approximatie bij het oplossen van differentiaalvergelijkingen is eerder toegepast door Lambert en Shaw [5], die meer algemeen een aantal zowel impliciete als expliciete meerstapsformules beschouwen. In dit rapport komen slechts expliciete éénstapsformules ter sprake. Bij een aantal formules wordt het principe van de exponentiële aanpassing toegepast. Belangrijke publicaties betreffende deze techniek zijn verschenen na 1963 (o.a. Pope [8], Liniger en Willoughby [7], Lawson [6] en Fowler en Warten [1]).

De afleiding der formules wordt gegeven in de volgende sectie. Sectie 3 bevat de ALGOL 60 tekst en de beschrijving der parameters van twee procedures rat2 en rat. In de laatste sectie worden drie testvoorbeelden behandeld, waarbij ter vergelijking de resultaten van de procedures modified taylor en exponential fitted taylor uit [4] worden vermeld.

2. Formules

In deze sectie wordt een afleiding van de integratieformules gegeven. Veronderstel dat we bij het oplossen van een enkele differentiaalvergelijking

$$\frac{du}{dt} = H(u,t) \ , \qquad u(t_0) = u_0 \ ,$$

zijn gevorderd met integreren tot het tijdstip t_k, waarbij een numerieke oplossing $u_k^*(t_k) = u_k$ is verkregen. Onder de lokaal analytische oplossing op het tijdstip t_k verstaan we nu de analytische oplossing van het beginwaardeprobleem

$$\frac{du}{dt} = H(u,t) \ , \qquad u(t_k) = u_k \ .$$

Deze lokaal analytische oplossing op het tijdstip t_k wordt nu beschouwd als functie van de staplengte τ en genoteerd als $\tilde{u}_k(\tau)$. We benaderen $\tilde{u}_k(\tau)$ door

$$u_k(\tau) = \frac{P_m(\tau)}{Q_m(\tau)} = \frac{\sum_{i=0}^{m} a_i \tau^i}{\sum_{i=0}^{n} b_i \tau^i} \ .$$

Indien $\tilde{u}_k(0)$ bestaat moet gelden $b_0 \neq 0$ en, omdat $u_k(\tau)$ onveranderd blijft bij deling van teller en noemer door b_0, kiezen we steeds $b_0 = 1$.
$u_k(\tau)$ heeft dus n+m+1 vrije coëfficiënten, die als volgt gekozen kunnen worden:

1) Een aantal afgeleiden van $u_k(\tau)$ en $\tilde{u}_k(\tau)$ stemmen overeen als $\tau = 0$, ofwel

$$u_k^{(i)}(0) = \tilde{u}_k^{(i)}(0) \text{ voor } i = 0,1,\ldots,p \ .$$

De i-de afgeleide van $u_k(\tau)$ en $\tilde{u}_k(\tau)$ wordt hierbij genoteerd als $u_k^{(i)}(\tau)$ respectievelijk $\tilde{u}_k^{(i)}(\tau)$. Een formule, die aan deze voorwaarden voldoet heet consistent van de orde p.

2) De overige l = n+m-p coëfficiënten kunnen worden gebruikt om de formule exponentieel aan te passen.

We definiëren nu kort enige begrippen met betrekking tot de stabiliteit van een integratiemethode. Veronderstel dat een éénstapsmethode wordt toegepast op de modelvergelijking:

$$du/dt = \delta u \ ,$$

waarin δ een willekeurig getal is. Voor eenvoudige formules, b.v. Taylorformules, vinden we dan de volgende betrekking:

$$u_k(\tau_k) = R(\tau_k\delta)u_k \ ,$$

waarbij R(z) een functie is, die uitsluitend afhangt van de integratieformule en niet van δ, τ_k of u_k. R(z) noemen we de stabiliteitsfunctie van de integratieformule. Met stabiliteitsgebied wordt bedoeld het gebied

$$\{z \in C \mid |R(z)| < 1\} \ .$$

Wanneer voor willekeurige z in het linkerhalfvlak Re z < 0 geldt $|R(z)| < 1$, spreekt men van A-stabiliteit.

We beschouwen nu een in het voorgaande gedefinieerde integratieformule, waarvan de coëfficiënten slechts door de consistentie-eisen (1) worden bepaald, en die dus een Padé-approximatie is van de lokaal analytische oplossing. De stabiliteitsfunctie, behorende bij deze methode is dan een Padé-benadering van e^z. Om een zo groot mogelijk stabiliteitsgebied te verkrijgen, hebben we in dit rapport rationale Padé-benaderingen gebruikt. Voor het geval, waarin de noemer een lineaire vorm is (n=1) en de graad van de teller willekeurig, vindt men in [5] een algemene vorm van de integratiemethode. We hebben hier slechts formules beschouwd, waarvoor geldt $n \geq m$, vanwege de dan over het algemeen betere stabiliteitseigenschappen. De "diagonale" Padé-benaderingen bijvoorbeeld, zijn zelfs A-stabiel.

Tabel 2.1. geeft een overzicht van de formules, die in dit rapport ter sprake zullen komen.

tabel 2.1. Formules

	m	n	p	l
rat1	1	1	1	1
rat2	1	1	2	0
rat3	1	2	2	1
rat4	1	2	3	0
rat5	2	2	3	1

Vanwege de reeds goede stabiliteitseigenschappen hebben we het aantal coëfficiënten dat gebruikt wordt om de formule exponentieel aan te passen, beperkt tot een enkele coëfficiënt.

De formules worden afgeleid voor een scalar differentiaalvergelijking; oplossing van een stelsel differentiaalvergelijkingen geschiedt door de betreffende formule componentsgewijs toe te passen. De parameters, die bepaald worden door de consistentievoorwaarden zijn dus eigenlijk vectoren, terwijl de door exponentiële aanpassing bepaalde parameter voor iedere component dezelfde waarde heeft.

Een nadeel van het gebruik van deze rationale formules is de ongewenste situatie, die zich voordoet wanneer de noemer van de rationale vorm bijna nul wordt en men weet dat de oplossing op dat tijdstip geen singulariteit heeft. De enige mogelijkheid om uit deze situatie te geraken is een verandering aan te brengen in de staplengte. In de in sectie 3 te geven procedures is een controle op de noemer ingebouwd.

Nu volgt de afleiding der formules, alsmede de bij formule rat2 geconstrueerde stapkeuzestrategie. De eerste, tweede en derde afgeleide van de numerieke oplossing op het tijdstip t_k zullen worden aangeduid met u_k', u_k'' resp. u_k'''.

2.1. rat1

Bij de afleiding van de formule voor rat1 gaan we uit van

$$u_k(\tau) = a_0 \frac{1+a_1\tau}{1+b_1\tau} \quad .$$

De twee consistentievoorwaarden zijn:

$$u_k = u_k(0) = a_0 \ ,$$
$$u_k' = u_k'(0) = a_0(a_1-b_1) \ .$$

Na eliminatie van de coëfficiënten a_0 en a_1 ontstaat de formule

$$u_k(\tau) = u_k + \frac{u_k' \tau}{1+b_1\tau}$$

of

$$(2.1.1) \qquad u_{k+1} \underset{\text{def}}{=} u_k(\tau_k) = u_k + \frac{u_k' \tau_k}{1+b_1\tau_k}$$

De parameter b_1 wordt nu bepaald door aanpassing van de formule in de meest negatieve eigenwaarde δ_k van de Jacobiaan van het op te lossen stelsel. Wanneer we (2.1.1) toepassen op de modelvergelijking $u' = \delta_k u$ vinden we de betrekking

$$u_{k+1} = u_k \left(1 + \frac{\tau_k\delta_k}{1+b_1\tau_k}\right) \ .$$

De waarde voor b_1 wordt nu gevonden uit

$$e^{\tau_k\delta_k} = 1 + \frac{\tau_k\delta_k}{1+b_1\tau_k} \ .$$

Substitutie van b_1 in (2.1.1) levert

$$(2.1.2) \qquad u_{k+1} = u_k + \frac{e^{\tau_k\delta_k}-1}{\delta_k} u_k' \ .$$

We merken op dat door de exponentiële aanpassing het rationale karakter van de formule is verdwenen. In feite is (2.1.2) een eerste orde exponentiëel gefitte Taylormethode.

Bij stelsels differentiaalvergelijkingen, waarbij sterk negatieve eigenwaarden optreden geeft deze methode numeriek slechte resultaten. De term

$e^{\tau_k \delta_k}$ kan bijvoorbeeld kleiner worden dan de nauwkeurigheid waarmee wordt gerekend, waardoor (2.1.2) overgaat in de formule

$$u_{k+1} = u_k - \frac{u_k'}{\delta_k} \quad ,$$

die dan op het gehele stelsel wordt toegepast en in bepaalde componenten tot grote afwijkingen kan leiden. Dit wordt mede veroorzaakt door het feit dat rat1 slechts eerste orde exact is, en we hebben besloten om rat1 niet in de testresultaten op te nemen.

De afleiding der overige formules verloopt analoog aan die van rat1 en daarom geven we nog slechts de basisformule voor $u_k(\tau)$, de consistentievoorwaarden en de verkregen formule met eventueel de coëfficiënt die bepaald wordt door de exponentiële aanpassing.

2.2. rat2

Basisformule:

$$u_k(\tau) = a_0 \frac{1+a_1\tau}{1+b_1\tau} \quad .$$

Consistentievoorwaarden:

$$u_k = u_k(0) = a_0 \quad ,$$
$$u_k' = u_k'(0) = a_0(a_1-b_1) \quad ,$$
$$u_k'' = u_k''(0) = -2a_0b_1(a_1-b_1) = -2b_1u_k' \quad .$$

Formule rat2 :

$$u_{k+1} = u_k + \frac{\tau_k u_k'}{1-\frac{1}{2}\tau_k \frac{u_k''}{u_k'}} \quad .$$

Na toepassing van rat2 op de modelvergelijking $u' = \delta u$ vinden we als stabiliteitsfunctie de A-stabiele (1,1) Padé-benadering van e^z

$$R(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z} \quad .$$

Bij de stapkeuzestrategie voor rat2 zullen we gebruik maken van de volgende benadering van de discrepantie ρ_k, die in [4], sectie 6.3. is afgeleid voor de stapgroottebepaling bij de procedure exponential fitted taylor:

$$(2.2.1) \qquad \rho_k = \tau_k \| u'_k(\tau_k) - u'_{k+1} \| \ .$$

In het bijzonder geldt bij de formule rat2 :

$$u'_k(\tau_k) = \frac{u'_k}{(1-\frac{1}{2}\frac{u''_k}{u'_k}\tau_k)^2} \ .$$

Wegens de tweede orde consistentie geldt bovendien

$$\rho_k = O(\tau_k^3)$$

of

$$\rho_k = c_k \tau_k^3 \ .$$

We nemen vervolgens aan dat c_k weinig verandert bij verschillende integratiestappen zodat bij vervanging van c_k door c_{k-1} ontstaat:

$$\rho_k = c_{k-1}\tau_k^3$$

en dus

$$\tau_k = \tau_{k-1}\sqrt[3]{\frac{\rho_k}{\rho_{k-1}}}$$

Wanneer ρ_k gelijk wordt genomen aan de tolerantie η_k dan vindt men voor de nieuw te nemen staplengte

$$\tau_k = \tau_{k-1}\sqrt[3]{\frac{\eta_k}{\rho_{k-1}}} \ .$$

waarbij

$$(2.2.2) \qquad \rho_{k-1} = \tau_{k-1} \| \frac{u'_k}{(1-\frac{1}{2}\frac{u''_{k-1}}{u'_{k-1}}\tau_{k-1})^2} - u'_k \| \ .$$

In de procedure rat2 is voor $\|\ \|$ in (2.2.2) de euclidische norm gekozen.

2.3. rat3

Basisformule:

$$u_k(\tau) = a_0 \frac{1+a_1\tau}{1+b_1\tau+b_2\tau^2} \quad .$$

Consistentievoorwaarden:

$$u_k = u_k(0) = a_0 \ ,$$
$$u_k' = u_k'(0) = a_0(a_1-b_1) \ ,$$
$$u_k'' = u_k''(0) = 2a_0(b_1^2-b_2-a_1b_1) = -2b_1u_k'-2b_2u_k \ .$$

Formule rat3 :

$$u_{k+1} = u_k + \frac{\tau_k(u_k' - u_k\tau_k b_2)}{1-\frac{u_k''}{2u_k'}\tau_k - \frac{u_k}{u_k'}\tau_k + b_2\tau_k^2} \ ,$$

waarbij

$$b_2 = \frac{\delta_k(1 - \frac{1}{2}z - (1+\frac{1}{2}z)e^{-z})}{\tau_k(1 - z - e^{-z})} \ , \quad z = \tau_k\delta_k \ .$$

Bij het testen van rat3 verkregen we enkele afwijkende resultaten. We geven hierop een korte toelichting.

Exponentiële aanpassing van de eigenwaarde van de Jacobiaan heeft tot gevolg dat de oplossing van de modelvergelijking $u' = \delta u$ exact wordt berekend. Voor een willekeurige constante c beschouwen we nu de vergelijking

(2.3.1) $\quad u' = \delta u + c$

met startwaarde $u(t_k) = u_k$, waarvan de exacte oplossing wordt gegeven door

(2.3.2) $\quad \tilde{u}_{k+1} = \tilde{u}(t_{k+1}) = u_k e^{\delta(t_{k+1}-t_k)} + \frac{c}{\delta}(e^{\delta(t_{k+1}-t_k)} - 1) \ .$

Bij gefitte Taylorformules bijvoorbeeld, is eenvoudig na te gaan, dat ook (2.3.1) exact wordt opgelost. Bij rat3 is dit echter niet het geval. Om een indruk te krijgen van de afwijking passen we rat3 toe op (2.3.1) en vinden

$$u_{k+1} = u_k + \frac{u_k\tau_k(\delta-\tau_k b) + \tau_k c}{1 - \frac{\delta\tau_k}{2} + b\tau_k^2 - \frac{u_k}{\delta u_k+c} b\tau_k} ,$$

met $\tau_k = t_{k+1}-t_k$. Deze vorm kunnen we als volgt splitsen

$$u_{k+1} = u_k + Au_k + \frac{\tau_k c - \frac{b\tau_k c}{\delta(\delta u_k+c)} Au_k}{1 - \frac{\delta\tau_k}{2} + b\tau_k^2 - \frac{u_k}{\delta u_k+c} b\tau_k} ,$$

waarbij geldt

$$A = \frac{\tau_k(\delta - \tau_k b)}{1 - \frac{\delta\tau_k}{2} + b\tau_k^2 - \frac{b\tau_k}{\delta}} .$$

Rat3 is exponentiëel aangepast, hetgeen betekent dat de parameter b zodanig is gekozen dat

$$A = e^{\tau_k\delta} - 1 .$$

Nu kunnen we m.b.v. (2.3.2) een uitdrukking vinden voor de foutterm ϕ, die blijkbaar afhankelijk is van de beginwaarde u_k :

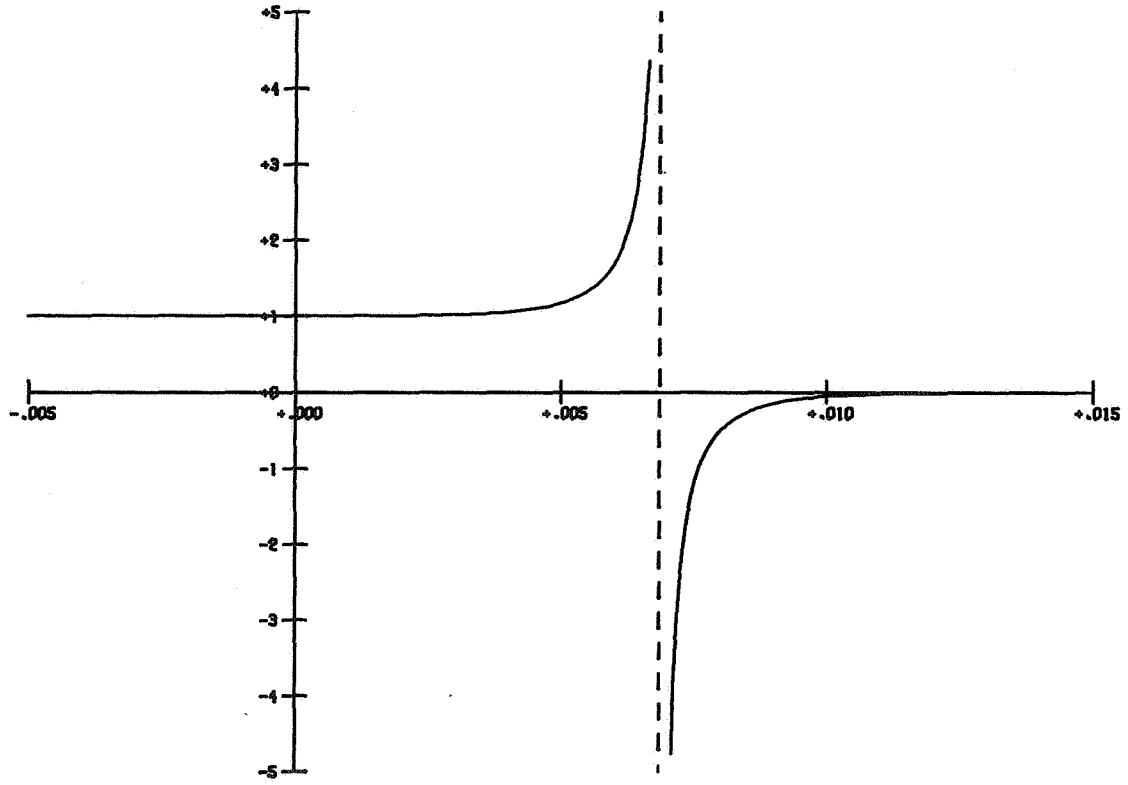
$$(2.3.3) \qquad \phi = \frac{\tau_k c - \frac{b\tau_k c}{\delta(\delta u_k+c)} (e^{\tau_k\delta}-1)u_k}{1 - \frac{\delta\tau_k}{2} + b\tau_k^2 - \frac{u_k b\tau_k}{\delta u_k+c}} - \frac{c}{\delta} (e^{\tau_k\delta} - 1) .$$

Omdat ϕ een tamelijk ondoorzichtige expressie is beschouwen we in het bijzonder de vergelijking

$$u' = -1000(u+1) , \quad u(0) = 0 ,$$

met als exacte oplossing $u = e^{-1000t}-1$.

We starten nu voor verschillende waarden van t_k op deze oplossing en na een integratiestap ter lengte 1 beschouwen we de absolute fout ϕ, die verkregen wordt uit (2.3.3), waarbij $c = \delta = -1000$, $\tau = 1$ en $u_k = e^{-1000t_k} - 1$. In de onderstaande figuur wordt ϕ uitgezet tegen het begintijdstip t_k. We zien dat slechts een redelijke precisie wordt bereikt indien t_k zodanig wordt genomen dat $u_k = -1$. Dit werd bevestigd door enkele numerieke experimenten, waarbij na meerdere integratiestappen een snelle foutenopbouw ontstond, als werd gestart met een geringe afwijking van $u_k = -1$.



Wegens dit falen bij het oplossen van niet-autonome differentiaalvergelijkingen hebben we rat3, evenals rat1, niet bij de testresultaten in sectie 4 opgenomen.

2.4. rat4

Basisformule:

$$u_k(\tau) = a_0 \frac{1 + a_1\tau}{1 + b_1\tau + b_2\tau^2} \quad .$$

Consistentievoorwaarden:

$$u_k = u_k(0) = a_0 \ ,$$
$$u_k' = u_k'(0) = a_0(a_1-b_1) \ ,$$
$$u_k'' = u_k''(0) = 2a_0(b_1^2-b_2-a_1b_1) = -2b_1u_k' - 2b_2u_k \ ,$$
$$u_k''' = u_k'''(0) = 6a_0(a_1(b_1^2-b_2) + 2b_1b_2 - b_1^3) = -3b_1u_k'' - 6b_2u_k' \ .$$

Formule rat4 :

$$u_{k+1} = u_k + \frac{(u_k'-b_2\tau_k u_k)\tau_k}{1+b_1\tau_k+b_2\tau_k^2} \ ,$$

waarbij

$$b_1 = \frac{3u_k''u_k' - u_k'''u_k}{3(u_ku_k''-2(u_k')^2)} \quad \text{en} \quad b_2 = \frac{2u_k'''u_k' - 3(u_k'')^2}{6(u_ku_k''-2(u_k')^2)} \ .$$

Als stabiliteitsfunctie vinden we de A-stabiele (1,2) Padé-benadering

$$R(z) = \frac{1 + \frac{1}{3}z}{1 - \frac{2}{3}z + \frac{1}{6}z^2} \ .$$

2.5. rat5

Basisformule:

$$u_k(\tau) = a_0 \frac{1 + a_1\tau + a_2\tau^2}{1 + b_1\tau + b_2\tau^2} \quad .$$

Consistentievoorwaarden:

$$u_k = u_k(0) = a_0 ,$$
$$u_k' = u_k'(0) = a_0(a_1-b_1) ,$$
$$u_k'' = u_k''(0) = 2a_0(b_1(b_1-a_1)+a_2-b_2) = -2b_1u_k' - 2(b_2-a_2)u_k ,$$
$$u_k''' = u_k'''(0) = 6a_0(2b_1b_2-b_1^3+a_1b_1^2-a_1b_2-a_2b_1) = -3b_1u_k'' - 6b_2u_k' .$$

Formule rat5 :

$$u_{k+1} = u_k + \frac{u_k'\tau_k + (b_1u_k' + \frac{1}{2}u_k'')\tau_k^2}{1 + b_1\tau_k - \frac{(u_k'''+3b_1u_k'')\tau_k^2}{6u_k'}} ,$$

waarbij

$$b_1 = \frac{e^z(\frac{z^2}{6} - 1) + z + 1 + \frac{z^2}{3}}{-\tau_k(e^z(\frac{z}{2} - 1)+ \frac{z}{2} + 1)} , \quad z = \tau_k\delta_k .$$

We merken op dat de bij formule rat3 gesignaleerde moeilijkheden bij rat5 niet optreden, d.w.z. (2.3.1) wordt door rat5 exact opgelost.

3. Procedures

In deze sectie worden de parameters besproken van de procedures rat2 (formule rat2) en rat (formule rat3, rat4 of rat5). Bovendien wordt ook de ALGOL 60 tekst van beide procedures gegeven.

3.1. Heading en parameters van rat2

```
procedure rat2 (t, te, m0, m, u, derivative, hmin, hmax, eta, k,
                output);
integer mo, m, k;
real t, te, hmin, hmax, eta;
array u;
procedure output, derivative;
```

Parameters:

t : <variable>;
t wordt gebruikt als Jensen parameter;
bij een aanroep van rat2 moet t de beginwaarde t_0 van de onafhankelijke variabele hebben;

te : <arithmetic expression>;
de eindwaarde van t;

m0, m : <arithmetic expression>;
indices van de eerste en de laatste vergelijking van het stelsel differentiaalvergelijkingen;

u : <array identifier>;
een één-dimensionaal array u[m0:m];
bij een aanroep van rat2 moet dit array de beginwaarden van $U(t_0)$ bevatten;

derivative : <procedure identifier>;
derivative moet als volgt door de gebruiker worden gegeven:
procedure derivative (i,a); integer i; array a;
<body>;
i neemt de waarden 1 en 2 aan en a is een één-dimensionaal

array a[m0:m]; na uitvoering van deze procedure moet array a de componenten van de i-de afgeleide van U(t) bevatten;

hmin, hmax : <arithmetic expression>;
de minimale resp. maximale staplengte, waarmee wordt geïntegreerd;

eta : <arithmetic expression>;
de gewenste precisie in de resultaten;

k : <variable>;
telt het aantal integratiestappen;
bij de eerste aanroep van rat2 moet k een door de gebruiker gegeven waarde hebben (bijvoorbeeld 0);

output : <procedure identifier>;
deze procedure dient als volgt door de gebruiker te worden gegeven;
procedure output; <body>;
in deze procedure kan men uitvoer vragen van bijvoorbeeld t, u[m0],...,u[m], k;

3.2. De body van rat2

```
procedure rat2(t,te,m0,m,u,derivative,hmin,hmax,eta,k,output);
real t,te,hmin,hmax,eta;
integer m0,m,k;
array u;
procedure derivative,output;
begin
        real hn,x,s,F;
        integer j;
        array f,test,f1,fnm1[m0:m];
        derivative(1,f);
        derivative(2,f1);
        hn:=hmin;
return:
        if t+hn>te then begin hn:=te-t;t:=te end;
        for j:=m0 step 1 until m do test[j]:=2×f[j]/f1[j];
```

```
            for j:=m0 step 1 until m do if abs(hn-test[j])<hn×10-4 then
                                    begin hn:=hn×(1-2×10-4);j:=m0-1 end;
            for j:=m0 step 1 until m do u[j]:=u[j]+hn×f[j]/(1-hn/test[j]);
            t:=if t=te then te else t+hn;k:=k+1;output;
            if t<te then
            begin
                    s:=0;
                    for j:=m0 step 1 until m do fnm1[j]:=f[j];
                    derivative(1,f);
                    for j:=m0 step 1 until m do
                    begin  F:=fnm1[j]/(fnm1[j]-hn×f1[j]/2);
                            x:=abs(fnm1[j]×F×F-f[j]);
                            if x>s then s:=x
                    end;
                    derivative(2,f1);
                    hn:=(eta×hn×hn/s)↑(1/3);
                    if hn>hmax then hn:=hmax;
                    if hn<hmin then hn:=hmin;
                    goto return
            end
    end rat2;
```

3.3. Heading en parameters van rat

```
procedure rat (t, te, m0, m, u, derivative, h, k, output, delta, l);
integer m0, m, k, l;
real t, te, h, delta;
array u;
procedure derivative, output;
```

De parameters, die niet reeds genoemd zijn in 3.1. zijn:

h : <arithmetic expression>;
de door de gebruiker op te geven integratiestaplengte;

delta : <arithmetic expression>;
de meest negatieve eigenwaarde van de Jacobiaan van het stelsel;

delta moet door de gebruiker worden gegeven;

l : <arithmetic expression>;
door mee te geven l = 3, 4 of 5 kan de gebruiker een keuze maken tussen respectievelijk de formules rat3, rat4 of rat5.

De procedure identifier derivative wordt gedefinieerd als in 3.1., waarbij dan bovendien de derde afgeleide moet worden gegeven (i=3).

3.4. De body van rat

```
procedure rat(t,te,m0,m,u,derivative,h,k,output,delta,l);
real t,te,h,delta;
integer m0,m,k,l;
array u;
procedure derivative,output;
begin   real num,den,au,aq,hn,z,expz,d,dh,h2;
        real array q,f,f1,f2[m0:m];
        integer j,count;
        procedure check denominator;
        begin   q[j]:=num/den;
                if abs(den)<10-5 ∧ count<2 then
                begin   au:=abs(u[j]);aq:=abs(q[j]);
                        if (au>1∧aq>au×102)∨(au<1∧aq>102) then
                        begin hn:=hn×.7;j:=m0-1;count:=count+1 end
                end
        end;

return:
        derivative(1,f);
        derivative(2,f1);
        hn:=h;if t+hn>te then hn:=te-t;
        count:=0;
        if l=3 then
        begin   for j:=m0 step 1 until m do
```

```
            begin   if j=m0 then
                    begin   z:=hn×delta;expz:=exp(z);
                            dh:=delta×(1+z/(1-z/(1-1/expz))/2)
                    end;
                    den:=1-f1[j]×hn/(2×f[j])+dh×(hn-u[j]/f[j]);
                    num:=hn×(f[j]-u[j]×dh);
                    check denominator
            end
    end else
    if l=4 then
    begin   derivative(3,f2);
            for j:=m0 step 1 until m do
            begin   if j=m0 then h2:=hn×hn;
                    den:=6×u[j]×f1[j]-12×f[j]×f[j]+hn×(6×f1[j]×f[j]-
                         2×f2[j]×u[j])+h2×(2×f2[j]×f[j]-3×f1[j]×f1[j]);
                    num:=f[j]×hn×(6×u[j]×f1[j]-12×f[j]×f[j])-
                         u[j]×h2×(2×f2[j]×f[j]-3×f1[j]×f1[j]);
                    check denominator
            end
    end else
    begin   derivative(3,f2);
            for j:=m0 step 1 until m do
            begin   if j=m0 then
                    begin z:=hn×delta;expz:=exp(z);
                          d:=(expz×(z×z/6-1)+z+1+z×z/3)/
                             (-hn×(expz×(z/2-1)+z/2+1));
                          h2:=hn×hn
                    end;
                    num:=f[j]×hn+(d×f[j]+f1[j]/2)×h2;
                    den:=1+d×hn-h2×(f2[j]+3×d×f1[j])/(6×f[j]);
                    check denominator
            end
    end;
    for j:=m0 step 1 until m do u[j]:=u[j]+q[j];
```

```
            t:=t+hn;k:=k+1;
            output;
            if t<te then goto return

    end rat;
```

4. Testresultaten

In deze sectie bespreken we drie problemen waarmee we de formules rat2, rat4 en rat5 hebben getest, te weten een enkele niet-lineaire vergelijking, een stelsel niet-lineaire [6] en een stelsel lineaire vergelijkingen [1].

Ter vergelijking zijn deze problemen ook opgelost met de procedures modified taylor en exponential fitted taylor [4], die eveneens gebruik maken van een aantal afgeleiden van U(t). De formules rat2 en rat4 kunnen bijvoorbeeld vergeleken worden met modified taylor, waarbij dan n afgeleide-evaluaties worden gebruikt, n = 2 resp. n = 3. In het geval n = 2 hebben we modified taylor toegepast met de volgende stabiliteitspolynomen:

$$1 + x + \frac{1}{2}x^2 , \quad \beta = 2 ,$$

$$1 + x + \frac{1}{8}x^2 , \quad \beta = 8 ,$$

waarbij β de reële stabiliteitsgrens is.
In het geval n = 3 onderscheiden we:

$$1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 , \quad \beta = 2.54 ,$$

$$1 + x + \frac{1}{2}x^2 + \frac{1}{16}x^3 , \quad \beta = 6.3 ,$$

$$1 + x + \frac{4}{27}x^2 + \frac{4}{729}x^3 , \quad \beta = 18 .$$

Rat5 kan worden vergeleken met exponential fitted taylor. Bij beide procedures wordt exponentieel gefit en worden 3 afgeleiden gebruikt.

4.1. Een eenvoudige niet-lineaire differentiaalvergelijking

Het beginwaardeprobleem

$$(4.1) \qquad \begin{cases} \dot{u} = 100 - u^2 \\ u(0) = 0 \end{cases}$$

heeft als exacte oplossing

$$\tilde{u} = 10 - \frac{20}{e^{20t} + 1} .$$

We kunnen spreken van een mild-stijf probleem, aangezien de eigenwaarde -2u van de vergelijking asymptotisch naar -20 gaat voor $t \geq 0$.

We geven eerst in tabel 4.1. enige resultaten, verkregen met de procedure modified taylor, waarbij gebruik gemaakt is van verschillende stabiliteitspolynomen. De betekenis der parameters is als volgt:

n : het aantal afgeleiden van u(t) ,

p : de orde van de formule ,

η_a, η_r : de opgegeven absolute resp. relatieve tolerantie ,

$-^{10}\log \varepsilon$: het aantal correcte cijfers op het tijdstip $t = 6$

(dus $\varepsilon = \left| \frac{u(6) - \tilde{u}(6)}{\tilde{u}(6)} \right|$) ,

k : het aantal gebruikte integratiestappen .

tabel 4.1. probleem 4.1. met modified taylor

n	p	η_a	η_r	$-^{10}\log \varepsilon$	k
3	3	10^{-3}	10^{-3}	5.6	65
3	2	10^{-4}	10^{-3}	9.3	44
3	1	10^{-3}	10^{-3}	11.1	51
2	2	10^{-4}	10^{-3}	4.1	107
2	1	10^{-3}	10^{-3}	8.1	55

De procedure rat2, waarbij verschillende waarden zijn meegegeven voor de tolerantie η en de minimale staplengte hmin, levert de volgende resultaten:

tabel 4.2. probleem 4.1. met rat2

η	hmin	$-^{10}\log \varepsilon$	k
10^{-3}	.05	6.6	15
10^{-3}	.03	6.4	21
10^{-4}	.05	7.7	17
10^{-4}	.03	7.5	24

Rat2 gebruikt 2 afgeleide-evaluaties en we kunnen deze reusltaten dus vergelijken met de laatste twee rijen van tabel 4.1. en constateren dan een aanzienlijke winst in het aantal integratiestappen.

Omdat bij de procedures rat4 en rat5 geen automatisch stapkeuzemechanisme is afgeleid wordt een staplengte τ meegegeven, die echter in de beginfase klein moet zijn omdat de afgeleide aanvankelijk tamelijk groot is. Tabel 4.3. toont de resultaten van rat4, waarbij τ als volgt is voorgeschreven:

$$\tau = \underline{\text{if}}\ t < t_0\ \underline{\text{then}}\ h\ \underline{\text{else}}\ 6\ .$$

Voor iedere waarde van h wordt in dezelfde kolom achtereenvolgens gegeven het aantal correcte cijfers in t = 6 en het aantal integratiestappen k.

tabel 4.3. probleem 4.1. met rat4

t_0	h					
	.02		.03		.05	
.9	5.8	46	6.1	32	5.7	19
.8	5.0	41	5.0	28	4.8	17
.7	4.1	36	4.2	25	4.0	15
.6	3.2	31	3.2	21	3.1	13
.5	2.3	26	2.4	18	2.3	11

Rat4 is derde orde exact en maakt gebruik van drie afgeleide-evaluaties. Vergelijking met tabel 4.1. toont weer een winst in het aantal integratiestappen, die echter ten koste gaat van de bereikte precisie. Uit de ver-

schillende kolommen h = .02, .03 en .05 van tabel 4.3. blijkt dat vergroting van het aantal integratiestappen hierin weinig verbetering brengt. De formule is echter wel stabiel; wanneer het inschakelverschijnsel goed is gerepresenteerd kan met zeer grote staplengte worden geïntegreerd.

De procedures rat5 en exponential fitted taylor, waarbij de eigenwaarde -2u exponentieel wordt aangepast, leveren beide in een beperkt aantal stappen resultaten in 12 cijfers nauwkeurig. Exponential fitted taylor, aangeroepen met $\eta_a = \eta_r = 10^{-2}$, heeft hiervoor 17 stappen nodig. Rat5 gebruikt 8 stappen, waarbij de staplengte τ is meegegeven volgens

τ = if t < .2 then .05 else 2.

Voor de volledigheid geven we tenslotte de gebruikte procedure derivative en de aanroep van rat5.

```
procedure derivative (i,a); integer i; array a;
begin  if i = 1 then begin c[1]:= y[1];
                           c[2]:= a[1]:= 100 - c[1]*c[1]
                     end;
       if i = 2 then c[3]:= a[1]:= -2*c[1]*c[2];
       if i = 3 then a[1]:= -2*(c[1]*c[3]+c[2]*c[2])
end;

t:= 0; y[1]:= 0; k:= 0;
rat (t, 6, 1, 1, y, derivative, if t < .2 then .05 else 2,
     k, output, -2*y[1],5);
```

4.2. Een stelsel niet-lineaire differentiaalvergelijkingen

We beschouwen het beginwaardeprobleem (zie [6])

$$(4.2)\qquad \begin{cases} \dot{u}_1 = (u_2^2-1)u_1 + u_2(1+u_2) \\ \dot{u}_2 = (-19+u_1^2+2u_1)u_2 - u_1 \\ u_1(0) = -1 \\ u_2(0) = 1 \end{cases}$$

Een analytische oplossing van (4.2) is niet bekend. De referentiewaarden op de tijdstippen t = 1 en t = 12 zijn verkregen door een 5-de orde Runge-Kutta formule toe te passen, waarbij een kleine staplengte werd gebruikt:

$$u_1(1) = -.33063085 \quad , \quad u_2(1) = .01784955 \quad ,$$
$$u_1(12) = -.29946231_{10}\text{-}5 \quad , \quad u_2(12) = .1668846_{10}\text{-}6 \quad .$$

De beide componenten gaan asymptotisch naar nul. Door toepassing van de stelling van Gerschgorin vinden we als benadering voor de eigenwaarden van de Jacobiaan

$$\delta_l = -19 + u_1^2 + 2u_1$$
$$\delta_r = u_2^2 - 1 \, .$$

De betekenis der parameters n, p, η_a, η_r en k in de volgende tabellen (4.4, 4.5, 4.6) is dezelfde als reeds genoemd is in 4.1. Met c_1 en c_2 wordt bedoeld het aantal correcte cijfers van u_1 resp. u_2 op het tijdstip t = 12. Tabel 4.4. geeft resultaten van de procedures modified taylor en exponential fitted taylor, waarbij modified taylor weer is toegepast met verschillende stabiliteitspolynomen. In tabel 4.5. worden de resultaten van rat2 gegeven voor enkele waarden van de tolerantie η, waarbij voor de minimumstap is genomen hmin = .05.

tabel 4.4. probleem (4.2) met modified taylor en exponential fitted taylor

	n	p	η_a	η_r	c_1	c_2	k
modified taylor	3	3	0	10^{-3}	2.8	3.3	111
	3	2	0	10^{-3}	1.2	1.3	84
	2	2	10^{-5}	10^{-3}	1.7	1.8	215
exponential fitted taylor			10^{-8}	10^{-3}	1.3	1.3	160

tabel 4.5. probleem (4.2) met rat2

η	k	c_1	c_2
10^{-8}	200	2.6	2.6
10^{-7}	164	1.8	2.0
10^{-6}	128	2.2	.9

Wanneer we bijvoorbeeld uit tabel 4.5. de resultaten, verkregen met $\eta = 10^{-7}$, vergelijken met modified taylor, n = p = 2, zien we dat rat2 bij dezelfde precisie duidelijk minder integratiestappen gebruikt.

Bij rat4 en rat5 wordt wederom niet met een vaste staplengte geïntegreerd omdat de u_2-component in de beginfase sterk daalt. De stap τ wordt als volgt voorgeschreven:

τ = if t < .5 then τ_1 else τ_2 .

We zien in tabel 4.6. dat bij rat5 door de exponentiële aanpassing met een grotere staplengte kan worden geïntegreerd, maar dat rat4 bij kleine staplengte duidelijk nauwkeuriger is.

tabel 4.6. probleem (4.3) met rat4 en rat5

			rat4		rat5	
τ_1	τ_2	k	c_1	c_2	c_1	c_2
.03	.1	132	3.4	3.4	2.3	2.3
.03	.2	76	2.1	2.6	1.5	1.8
.03	.3	56	.8	1.0	1.3	1.9
.03	.4	47	.5	.5	1.4	2.5
.05	.1	125	2.7	2.7	2.4	2.4
.05	.2	69	2.2	2.5	1.6	1.8
.05	.3	49	.9	1.1	1.3	2.0
.05	.4	40	.4	.7	1.4	2.3

4.3. Een lineair stelsel van Fowler en Warten

Fowler en Warten beschouwen in [1] het volgende beginwaardeprobleem

$$\dot{U} = DU + F \ , \quad U(0) = U_0 \tag{4.3}$$

waarin

$$D = \begin{pmatrix} -500.5 & 499.5 \\ 499.5 & -500.5 \end{pmatrix} \ , \quad F = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \ , \quad U_0 = \begin{pmatrix} -.1 \\ .1 \end{pmatrix} .$$

De analytische oplossing van (4.3) luidt:

$$U = 2(1-e^{-t})\begin{pmatrix} 1 \\ 1 \end{pmatrix} + .1\, e^{-1000t}\begin{pmatrix} -1 \\ 1 \end{pmatrix} .$$

De eigenwaarden van de matrix D zijn -1000 en -1 .

Een toepassing van de procedure exponential fitted taylor staat beschreven in [4]. Modified taylor gebruikt zeer veel integratiestappen, omdat vanwege de kleine eigenwaarde -1000 een kleine staplengte is vereist om toch binnen het stabiliteitsgebied te blijven. Daarom geven we in tabel 4.7. slechts de resultaten van de drie procedures rat2, rat4 en rat5, die zijn toegepast met een staplengte

$$\tau = \underline{\text{if}}\ t < .04\ \underline{\text{then}}\ .001\ \underline{\text{else}}\ h \ .$$

Er is dus weer voor gezorgd dat het inschakelverschijnsel goed wordt gerepresenteerd. Echter bij de 2-de orde formule rat2 bleek het noodzakelijk om het interval, waar met $\tau = .001$ wordt geïntegreerd, te vergroten tot [0,.05]. In de drie kolommen in tabel 4.7. worden achtereenvolgens gegeven het aantal integratiestappen k en $-^{10}\log \varepsilon$, waarbij

$$\varepsilon = \max_{k=1,2} \left| \frac{u_k(10) - \tilde{u}_k(10)}{\tilde{u}_k(10)} \right| .$$

tabel 4.7. probleem (4.3) met rat2, rat4 en rat5

h	rat2		rat4		rat5	
2	56	4.3	46	3.9	46	4.3
1	61	4.5	51	5.1	51	4.5
.5	71	5.1	61	6.1	61	5.1
.2	101	5.8	91	7.3	91	5.8
.1	151	6.4	141	7.8	141	6.4

Bij de A-stabiele formules rat2 en rat4 blijkt inderdaad dat met zeer grote staplengte kan worden gerekend, indien het inschakelverschijnsel goed wordt gerepresenteerd. We constateren dat bij rat5, evenals in het vorige voorbeeld en behalve voor grote staplengte, de resultaten minder nauwkeurig zijn dan bij rat4.

5. Samenvatting

In sectie 1 werden enkele rationale integratieformules afgeleid. Uitvoerig getest zijn de formules rat2 (tweede orde exact), rat4 (derde orde exact) en de exponentieel aangepaste formule rat5 (derde orde exact). De A-stabiliteit van de formules rat2 en rat4 maakt het mogelijk om met een grote staplengte te integreren. Dit wordt bevestigd wanneer de testproblemen ter vergelijking worden opgelost met de procedure modified taylor (zie [4]). Ook het bij rat2 geconstrueerde stapkeuzemechanisme blijkt goed te voldoen. Vergelijking van formule rat5 met de procedure exponential fitted taylor (zie [4]) levert voor rat5 geen betere resultaten op. Door de exponentiële aanpassing van rat5 is namelijk het rationale karakter van de formule verdwenen alsmede natuurlijk de goede stabiliteitseigenschappen.

Bij deze rationale formules treden moeilijkheden op wanneer bij het oplossen van een stelsel vergelijkingen in een der componenten de noemer zeer klein wordt. Als dit wordt geconstateerd kan men een correctie aanbrengen in de integratiestaplengte, maar het is mogelijk dat de moeilijkheid dan terugkeert in een andere component.

Referenties

[1] M.E. Fowler en R.M. Warten,
A numerical integration technique for ordinary differential equations with widely separated eigenvalues, IBM Journal, 537-543, (1967).

[2] P.J. van der Houwen,
One step methods for linear initial value problems I. Rapport TW 119/70, Math. Centrum.

[3] -- One step methods for linear initial value problems II. Rapport TW 122/70, Math. Centrum.

[4] P.J. van der Houwen, P. Beentjes, K. Dekker en E. Slagt,
One step methods for linear initial value problems III, Numerical examples. Rapport TW 130/71, Math. Centrum.

[5] J.D. Lambert en B. Shaw,
On the numerical solution of $y' = f(x,y)$ by a class of formulae based on rational approximation, Math. of Comp., vol. 19, (1965).

[6] J.D. Lawson,
Generalized Runge-Kutta processes for stable systems with large Lipschitz constants, SIAM J. Numer. Anal., vol. 4, no. 3, (1967).

[7] W. Liniger en R. Willoughby,
Efficient integration of stiff systems of ordinary differential equations, IBM Research Report RC 1970, (1967).

[8] D.A. Pope,
An exponential method of numerical integration of ordinary differential equations, Communications of the ACM, Numerical analysis, vol. 6, no. 8, (1963).