stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE ND 2/77 DECEMBER

M. LOUTER-NOOL (RED.)

VERSLAGEN VAN DE ALGEMENE WERKBESPREKINGEN VAN DE AFDELING NUMERIEKE WISKUNDE GEDURENDE DE PERIODE VAN NOVEMBER 1976 TOT EN MET DECEMBER 1977

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).

INHOUD

VOORWOORD

In het kader van de algemene werkbesprekingen van de afdeling Numerieke Wiskunde van het Mathematisch Centrum zijn in de periode van november 1976 tot en met juni 1977 voordrachten gehouden door verscheidene medewerkers van deze afdeling. Het doel van deze werkbesprekingen is de andere leden van de afdeling op de hoogte te stellen van de werkzaamheden op numeriek gebied op het Mathematisch Centrum.

De redactrice wil haar dank betuigen aan degenen, die aan de totstandkoming van dit rapport hebben medegewerkt, met name aan mevr. J.G. Blom en de heren W.J. Gerritsen en F.J. Reckers voor het samenstellen en corrigeren en de dames van de typekamer voor het typen van de manuscripten.

M. Louter-Nool

Algemene Werkbespreking Afdeling Numerieke Wiskunde

Hoofdstuk 1

Exponentieel Aangepaste Gewogen Residuen Methoden

door

P.W. Hemker

1.1. Inleiding

Beschouwd wordt het tweepuntsrandwaardeprobleem

(1.1.1) $\varepsilon y'' + fy' + gy = s,$

waarin f,g en s gladde funkties zijn op [a,b] en $0 < \varepsilon << 1$. De randwaarden zijn: $y(a) = \alpha$ en $y(b) = \beta$. De oplossing van dit soort vergelijkingen bezit, voor kleine waarden van ε, een langzaam en snel variërende component.

Voorbeeld:

Nemen we in (1.1.1) $f \equiv -1$ en $g = s \equiv 0$, dan krijgen we

$$\varepsilon y'' - y' = 0.$$

De analytische oplossing bestaat uit combinaties van de onafhankelijke componenten 1 en $e^{x/\varepsilon}$. Het gedrag van de oplossing is geschetst in figuur 1.1.1.

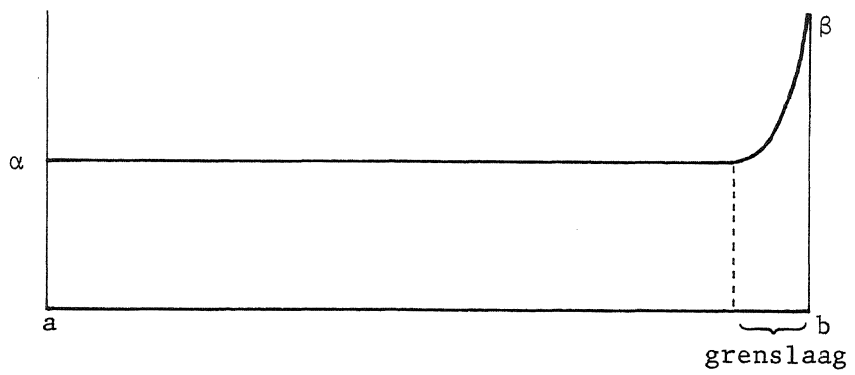


fig. 1.1.1.

De richting van de oplossing van (1.1.1) wordt bepaald door f, zoals blijkt uit de figuren 1.1.2.(a),(b),(c)

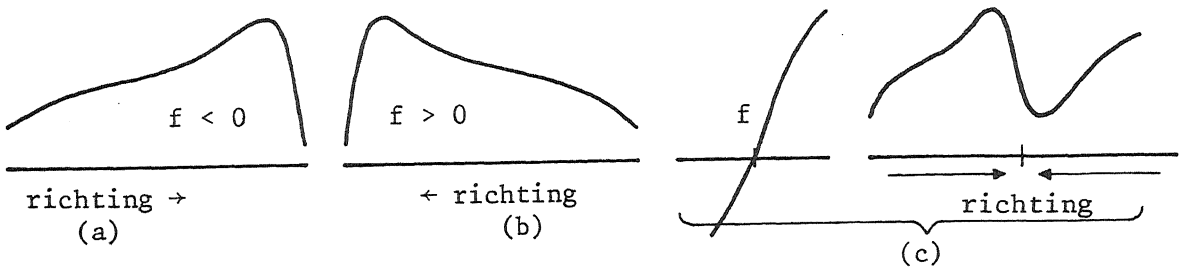


fig. 1.1.2. (a),(b),(c)

De oplossing van (1.1.1) is numeriek moeilijk te berekenen, vanwege het snel variërende karakter in de grenslaag.

1.2. Eindige differentie methode

De eindige differentie methoden kunnen, voor dit soort vergelijkingen, tot slechte resultaten leiden. Dit blijkt uit het volgende

Voorbeeld:

$$\varepsilon y'' + y' = 0,\ y(0) = 0 \text{ en } y(1) = 1.$$

De centrale differentie methode leidt tot de vergelijking

$$(1.2.1) \qquad \varepsilon(y_{i+1}-2y_i+y_{i-1})/h^2 + (y_{i+1}-y_{i-1})/(2h) = 0.$$

De numerieke oplossing wordt gegeven door

$$y_i = \frac{1-\mu^i}{1-\mu^N}, \text{ met } \mu = \frac{2\varepsilon-h}{2\varepsilon+h}$$

en de exacte oplossing door

$$y(x_i) = \frac{1-\nu^i}{1-\nu^N}, \text{ met } \nu = \exp(-h/\varepsilon).$$

In de drie bovenstaande formules is het aantal steunpunten N + 1,

$$h = 1/N,\ x_i = ih.$$

Indien $h << \varepsilon$ dan is de numerieke oplossing een goede benadering voor de analytische; in het geval dat $h = 2\varepsilon$ vindt men als numerieke oplossing $y_i \equiv 1$, $i = 1,\ldots,N$ en $y_0 = 0$.
Geldt dat $\varepsilon << h$ dan kan men twee gevallen onderscheiden: nl., N oneven en even.

In het eerste geval is de numerieke oplossing afwisselend nul en één en in het tweede geval onbegrensd voor oneven i en $\frac{i}{N}$ voor even i.

Een alternatieve oplossingsmethode, wordt als volgt geconstrueerd: wanneer $\varepsilon = 0$, dan krijgen we een gereduceerde vergelijking

$$fy' + gy = s,$$

waarbij we één randvoorwaarde teveel hebben.
Analytische theorie leert ons dat, onder zekere voorwaarden, de oplossing van deze gereduceerde vergelijking met de "goede" randvoorwaarde een benadering is van het gladde gedeelte van de oplossing van het niet gereduceerde probleem. Daarom kunnen we eerst een discretisatie van de gereduceerde vergelijking met de juiste randvoorwaarde construeren en daaraan een correctie-term voor $\varepsilon y''$ toevoegen.
DORR [1970] gebruikt voor de discretisatie van fy' achterwaartse differenties als $f < 0$ en voorwaartse als $f > 0$. Dit levert een O(h)-methode op. Een $O(h^2)$-methode werd geconstrueerd door ABRAHAMSSON [1974] e.a. welke berustte op een discretisatie van de eerste en nulde orde term. Dit gebeurt door middel van de "mid-point rule":

$$fy' \rightarrow \begin{cases} f(x_{i+\frac{1}{2}})\ (y_{i+1}-y_i)/h, & f > 0 \\ f(x_{i-\frac{1}{2}})\ (y_i-y_{i-1})/h, & f < 0 \end{cases}$$

$$gy \rightarrow \begin{cases} g(x_{i+\frac{1}{2}})\ (y_{i+1}+y_i)/2, & f > 0 \\ g(x_{i-\frac{1}{2}})\ (y_i+y_{i-1})/2, & f < 0 \end{cases}$$

Voor de discretisering van het niet-gereduceerde probleem wordt, naast deze discretisering van de eerste en nulde orde term, de discretisatie van de 2e orde term $\varepsilon y''$ als in (1.2.1) gebruikt.

1.3. Gewogen residuen methoden

Het verschil met de vorige methoden is dat de oplossing nu verkregen wordt als een lineaire combinatie van tevoren gekozen funkties.
Probleem (1.1.1) wordt nu herschreven tot

$$(1.3.1) \qquad Ly \equiv -(c_0 y')' + c_1 y' + c_2 y = s$$

op [a,b] met dezelfde randvoorwaarden.
Men kiest uit de oplossingsruimte S een eindige deelruimte S_h, opgespannen door de funkties $\phi_0, \phi_1, \ldots, \phi_N$.
Men kan de oplossing y_h van (1.3.1) dus schrijven als

$$(1.3.2) \qquad y_h = \sum_{j=0}^{N} a_j \phi_j .$$

Een methode (de collocatie methode) om (1.3.1) op te lossen berust op het kiezen van een aantal punten $\{x_i\}_{i=1}^{N-1}$ in het interval (a,b) waarvoor aan de vergelijking $Ly_h = s$ voldaan moet zijn.
Hiermee verkrijgt men het vierkante stelsel

$$\sum_{j=0}^{N} a_j \, L\phi_j(x_i) = s(x_i)$$

dat opgelost dient te worden.
Een tweede methode (de gewogen residuen methode) maakt gebruik van het inprodukt in de $C^2(a,b)$

$$(1.3.3) \qquad (Ly-s,\psi) = 0.$$

We merken op dat per definitie geldt

$$(u,v) = \int_a^b u(x)v(x)dx.$$

Als men (1.3.3) partieel integreert, dan krijgt men

$$B(y,\psi) = (s,\psi)$$

waarin geldt

(1.3.4) $$B(u,v) = (c_0u',v') + (c_1u',v) + (c_2u,v).$$

Na de partiële integratie zijn de eisen voor de funkties y minder beperkend. Immers, in (1.3.3) komt geen tweede afgeleide meer voor. Het op te lossen probleem is nu

$$B(y_h,v_h) = (s,v_h)$$

waarin y_h gedefinieerd is als in (1.3.2) en $v_h \in V_h \subset V$, waarbij V een toetsruimte voorstelt.

Er kunnen twee stellingen geformuleerd worden, die betrekking hebben op de schatting van de globale fout.

Stelling 1.3.1

Zijn S en V Hilbertruimten met inprodukten $(.,.)_S$ en $(.,.)_V$ en zij $B: S * V \longrightarrow \mathbb{R}$, B een bilineaire vorm, zodat

(1.3.1.1) $$\exists\, c_1 > 0 \quad \forall\, u \in S,\ v \in V : |B(u,v)| \leq c_1 \|u\|_S \cdot \|v\|_V$$

(1.3.1.2) $$\exists\, c_2 > 0 \quad \forall\, u \in S,\ \exists\, v \in V,\ v \neq 0 : |B(u,v)| \geq c_2 \|u\|_S \|v\|_V$$

(1.3.1.3) $$\forall\, v \in V,\ v \neq 0 \quad \exists\, u \in S : |B(u,v)| > 0$$

dan geldt:

$$\forall\, f \in V' \qquad \exists\,!\; u_0 \in S : B(u_0,v) = f(v),\ \forall\, v \in V.$$

Hierin is V' de duale ruimte van V.

Stelling 1.3.2

Laten de voorwaarden uit stelling 1.3.1. gelden en kies

$$S_h \subset S \text{ en } V_h \subset V \text{ zodat}$$

(1.3.2.1) $\exists\, D(S_h,V_h) = \text{constant},\ \forall\, u_h \in S_h,\ \exists\, v_h \in V_h,\ v_h \neq 0$:

$$|B(u_h,v_h)| \geq D(S_h,V_h)\ \|u_h\|_S\ \|v_h\|_V$$

(1.3.2.2) $\forall\, v_h \in V_h,\ v_h \neq 0 \qquad \exists\, u_h \in S_h : |B(u_h,v_h)| > 0$

dan geldt:

$$\forall\, f \in V' \qquad \exists\,!\; u_{0h} : B(u_{0h},v_h) = f(v_h)$$

en

(1.3.2.3) $$\|u_0 - u_{0h}\|_S \leq \left[1 + \frac{C_1}{D(S_h,V_h)}\right] \inf_{w \in S_h} \|u_0 - w\|_S.$$

De bewijzen kunnen gevonden worden in BABUŠKA en AZIZ [1972].

1.4. Exponentieel aangepaste gewogen residuen methoden

Bij de exponentieel aangepaste gewogen residuen methoden worden exponentiële funkties gebruikt in de ruimte V_h. Waarom het zinvol is om exponentiële funkties te gebruiken, wordt aangetoond aan de hand van Greense funkties.

De oplossing van het tweepuntsrandwaardeprobleem (1.3.1) wordt gegeven door

$$y(x) = -\int_a^b G(x,\xi)\, s(\xi)\, d\xi,$$

waarbij de Greense funktie $G(x,\xi)$ geconstrueerd kan worden uit twee oplossingen ϕ_1, ϕ_2 van $L^T\phi = 0$.
Laten ϕ_1 en ϕ_2 gedefinieerd zijn op $[a,b]$ met

$$\phi_1(a) = 0 \quad \text{en} \quad \phi_1'(a) = 1$$

$$\phi_2(b) = 0 \quad \text{en} \quad \phi_2'(b) = 1$$

dan is

$$G(x,\xi) = \begin{cases} \dfrac{\phi_1(\xi)\,\phi_2(x)}{F}, & \xi < x \\[2ex] \dfrac{\phi_1(x)\,\phi_2(\xi)}{F}, & \text{anders} \end{cases}$$

waarbij $F = c_0(x)(\phi_1(x)\phi_2'(x) - \phi_1'(x)\phi_2(x))$.

Met behulp van deze Greense funktie kan een puntsgewijze schatting van de fout gemaakt worden:

Als (1.3.1.1) geldt dan

$$|(y-y_h)(x)| \le C_1 \|y-y_h\|_S . \|G(x,.) - v_h\|_V, \quad \forall\, v_h \in V_h.$$

Hieruit kan afgeleid worden dat in de steunpunten geldt:

$$|(y-y_h)(x_i)| \leq k \cdot \|y-y_h\|_S \cdot \inf_{v_h \in V_h} \|G(x_i,.) - v_h\|_V$$

Een afschatting van de globale fout $\|y-y_h\|_S$ wordt gevonden in (1.3.2.3). De puntsgewijze fout kan extra klein gemaakt worden door V_h zo te kiezen dat $\inf_{v_h \in V_h} \|G(x_i,.)-v_h\|_V$ klein wordt.

Voorbeeld:

De Greense funktie corresponderend met de vergelijking

$$Ly = \varepsilon y'' - y' = 0 \quad \text{op} \quad [0,1]$$

is $G(x,\xi) =$ *if* $x > \xi$ *then* $\dfrac{(1-e^{-\xi/\varepsilon})\,(1-e^{(1-x)/\varepsilon})}{F}$ *else*

$$\frac{(1-e^{-x/\varepsilon})\,(1-e^{(1-\xi)/\varepsilon})}{F}$$

met $F = -e^{-x/\varepsilon}(e^{1/\varepsilon}-1)$

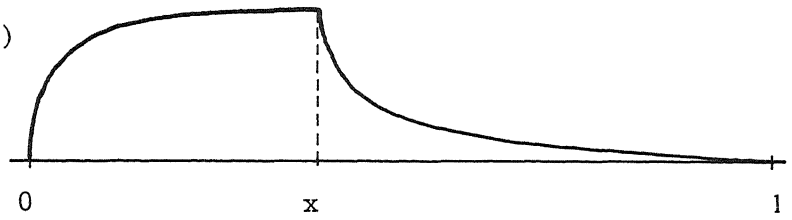


fig. 1.4.1.

Hieruit blijkt dat, althans voor de vergelijking $\varepsilon y'' + fy = 0$, met constante coefficiënten, een puntsgewijs exacte benadering van de oplossing gevonden wordt, wanneer de juiste exponentiële functies in de ruimte V_h worden opgenomen.
Afsluitend kan vermeld worden dat de oplossingsruimte S_h bestaat uit "stuksgewijze" polynomen en de toetsruimte V_h uit stuksgewijze exponentiële funkties en, als er nog vrijheid over is, stuksgewijze polynomen.

Stuksgewijze funkties zijn (hier) continue funkties, als polynoom of exponentiële funktie gedefinieerd op elk interval $[x_i,x_{i+1}]$, $i = 0,\ldots,N-1$.

REFERENTIES

ABRAHAMSSON, L.R., H.B. KELLER & H.O. KREISS [1974], *Difference approximations for singular perturbations of systems of ordinary differential equations*, Num. Math. 22, pp. 367-391.

BABUSKA, I. & A.K. AZIZ [1972], *Survey lectures on the mathematical foundations of the finite element method*, in: AZIZ ed. pp. 1-359.

DORR, F.W. [1970], *The numerical solution of singular perturbations of boundary-value problems*, SIAM J. Num. An. 1, pp. 141-146.

Algemene Werkbespreking Afdeling Numerieke Wiskunde

Hoofdstuk 2

Vergelijking van iteratieve methoden, verkregen d.m.v. inbeddingstechnieken

door

C. den Heijer

2.1. Probleem

Zij $F : \mathbb{R}^n \to \mathbb{R}^n$, een niet-lineaire operator. Inwendig product $(x,y) = \sum_{i=1}^{n} x_i y_i$, $\|\cdot\|$ Euclidische norm. We willen oplossen : $F(x) = 0$, en zij x^* de oplossing. Dit kan bijvoorbeeld met de methode van Newton:

$$\begin{aligned} & x_0 \text{ gegeven} \\ & x_{k+1} = x_k - F'(x_k)^{-1} . F(x_k) \qquad k = 0,1,\ldots \end{aligned} \tag{2.1.1}$$

Deze methode faalt echter, als x_0 "ver" van x^* vandaan ligt. Als x_0 "dicht" bij x^* ligt, dan is de methode van Newton erg goed en snel. ($F'(x^*)$ moet inverteerbaar zijn).

2.2. Continueringsmethoden

Inbeddingsmethoden vormen een onderdeel van de z.g. continueringsmethoden. Deze gaan als volgt te werk. Zij $x_0 \in \mathbb{R}^n$ één of ander punt. Dan definieert een continueringsmethode een curve $x(t)$, $t \in J$ ($J=[0,1]$ of $[0,\infty)$) met $x(0) = x_0$ en $x(1) = x^*$ (of $x(\infty) = x^*$).

Enige voorbeelden:

(2.2.1) Zij $f(x) = Ax + G(x) + \tilde{y}$, A lineair.

x_0 de oplossing van $Ax + \tilde{y} = 0$ (makkelijk).
Dan is x(t) de oplossing van $Ax + t\,G(x) + \tilde{y} = 0, \quad x(1) = x^*,$

$$Ax + (1-e^{-t})G(x) + \tilde{y} = 0, \quad x(\infty) = x^*,$$

(2.2.2) Bekende iteratieve methode:

$$x_{k+1} = x_k - h_k A(x_k)\, F(x_k) \qquad \text{(quasi - Newton methode)}$$

Dit is op te vatten als de methode van Euler, toegepast op d.v.

$$\dot{x}(t) = -\, A(x(t))\cdot F(x(t)) \qquad \text{(cont. quasi - Newton methode)}$$
$$x(0) = x_0$$

(2.2.3) Een inbeddingsmethode:
Zij $F : \mathbb{R}^n \to \mathbb{R}^n$, willekeurig; $x_0 \in \mathbb{R}^n$ gegeven. Dan: x(t) is de oplossing van

$$H(t,x) \equiv (1-t)\,(F(x) - F(x_0)) + t\,F(x) = 0$$

zodat geldt:

$$x(0) = x_0,\ x(1) = x^*.$$

We hebben a.h.w. F(x) = 0 ingebed in een "schaar" van vergelijkingen.

Opmerking:

Bij het gebruik van een continueringsmethode rijst de vraag of er wel een curve bestaat van x_0 naar x^* en of deze uniek is.

Bijvoorbeeld in (2.2.3). In dat geval kunnen we de volgende stelling formuleren:

Stelling 2.2.1.

Als $F'(x)$ niet-singulier is $\forall x \in \mathbb{R}^n$ en $\|F'(x)^{-1}\| \leq \beta$, dan is $\forall x_0 \in \mathbb{R}^n$ de curve $x(t)$ gedefinieerd en uniek met $x(1) = x^*$. De curve is continu en differentieerbaar en er geldt:

$$\begin{cases} \dot{x}(t) = - F'(x(t))^{-1} F(x_0) & \forall t \in [0,1] \\ x(0) = x_0. \end{cases}$$

2.3. Inbeddingsmethoden

Type (2.2.3) is een bijzonder geval van een inbeddingsmethode. Zij x_0 gegeven,

$$H(t,x) \text{ willekeurig, } t \in [0,1],\ x \in \mathbb{R}^n$$

zó dat

$$H(0,x_0) = 0$$
$$H(1,x) \equiv F(x).$$

Laat $x(t)$ de oplossing zijn van $H(t,x) = 0$, $t \in [0,1]$. Er geldt dan:

$$x(0) = x_0, \qquad x(1) = x^*.$$

Als H_t en H_x bestaan en $H(t,x(t)) = 0$, dan geldt:

$$\frac{d}{dt} H(t,x(t)) = 0.$$

Hieruit volgt:

$$H_t(t,x(t)) + H_x(t,x(t)) \cdot \dot{x}(t) = 0.$$

Neem aan dat $H_x(t,x(t))^{-1}$ bestaat, dan geldt:

$$\begin{cases} \dot{x}(t) = - H_x(t,x(t))^{-1} \cdot H_t(t,x(t)) & t \in [0,1] \\ x(0) = x_0. \end{cases}$$

Samenvattend:

We zoeken de curve x(t), die voldoet aan

(i) $H(t,x(t)) = 0, \quad t \in [0,1]$

(ii) $\dot{x}(t) = -H_x(t,x(t))^{-1} \cdot H_t(t,x(t))$
$x(0) = x_0.$

Oplossingsmethoden:

(2.3.1) Vergelijking (i) oplossen voor $0 < t_1 < \ldots\ldots < t_N = 1$, m.b.v. bijv. de methode van Newton (discrete inbedding).

(2.3.2) De d.v. (ii) oplossen m.b.v. een Runge-Kutta methode (evt. meerstapsmethode) (continue inbedding).

(2.3.3) Een kombinatie van (2.3.1) en (2.3.2).

We zullen dit illustreren aan de hand van het volgende
Voorbeeld:

(2.3.4) Zij $F : \mathbb{R}^n \to \mathbb{R}^n$ en

$$\begin{aligned} &\|F'(x) - F'(y)\| \leq \gamma \cdot \|x-y\| && \forall x,y \in \mathbb{R}^n \\ &\|F'(x)^{-1}\| \leq \beta && \forall x \in \mathbb{R}^n \end{aligned}$$

(dan is F een homeomorfisme).

Zij $x_0 \in \mathbb{R}^n$ gegeven

(2.3.5) $H(t,x) \equiv (1-t)\,(F(x) - F(x_0)) + t.F(x) \equiv F(x) - (1-t)F(x_0).$

Hieruit volgt

$$H_t = F(x_0),$$
$$H_x = F'(x),$$

wat leidt tot de volgende d.v.

$$\dot{x}(t) = - F'(x(t))^{-1} \cdot F(x_0)$$
$$x(0) = x_0.$$

Stelling 2.3.1 (Newton-Mysovski; zie b.v. [1])

Zij F een funktie, die aan (2.3.4) voldoet en $x_0 \in \mathbb{R}^n$ een gegeven vektor. Dan geldt, als

$$\beta^2\gamma\|F(x_0)\| < 2$$

dat de methode van Newton convergeert naar de oplossing x^* van $F(x) = 0$ vanuit x_0.

Stelling 2.3.2 (Avila; zie [2])

Zij $H : [0,1] \times \mathbb{R}^n \to \mathbb{R}^n$, gedefinieerd als in (2.3.5). Zij $x : [0,1] \to \mathbb{R}^n$ de continue oplossing van $H(t,x) = 0$. Dan geldt, als

$$0 < \Delta t = \frac{1}{N} \leq \frac{1}{2\beta^2\gamma\|F(x_0)\|}$$

en

$$x_k = x_{k-1} - H_x(k\Delta t, x_{k-1})^{-1} \cdot H(k\Delta t, x_{k-1})$$
$$\text{voor } k = 1,2,\ldots,N-1;\ x_0 = x(0)$$

$$x_k = x_{k-1} - F'(x_{k-1})^{-1} \cdot F(x_{k-1})$$
$$= x_{k-1} - H_x(1, x_{k-1})^{-1} \cdot H(1, x_{k-1}) \quad \text{voor } k \geq N,$$

dat $\{x_k\}$ naar x^* konvergeert.

Deze stelling doet een uitspraak over een oplossingsmethode van het type (2.3.1).

Stelling 2.3.3 (zie [2])

Laat F voldoen aan (2.3.4). Als

$$0 < \Delta t = \frac{1}{N} < \frac{4}{\beta^2\gamma\|F(x_0)\| \ [\exp(\beta^2\gamma\|F(x_0)\|) - 1]}$$

dan genereert het iteratieproces

$$x_k = x_{k-1} - \Delta t \, F'(x_{k-1})^{-1} \cdot F(x_0) \qquad k = 1,2,\ldots,N-1$$
$$x_k = x_{k-1} - \quad F'(x_{k-1})^{-1} \cdot F(x_{k-1}) \qquad k \geq N$$

een naar x^* konvergerende rij.

Deze stelling doet een uitspraak over een oplossingsmethode van het type (2.3.2).

Opmerking bij (2.3.3):
Zij

$$x(t) = -F'(x(t))^{-1} \cdot F(x_0)$$
$$x(0) = x_0$$

Deze d.v. kan opgelost worden met een Runge-Kutta methode met een stap-keuze-mechanisme. Zo'n mechanisme is meestal gebaseerd op een schatting van de lokale fout en dient ervoor te zorgen, dat de benaderingen niet al te ver van de ware oplossing af komen te liggen. Bij bovenstaande d.v. is dit heel eenvoudig te kontroleren.

Nl. zij $y_k \approx x(t_k)$, de benadering dan geldt $F(x(t_k)) = (1 - t_k)\, F(x_0)$.
Als $\|F(y_k) - (1-t_k)\, F(x_0)\| > \varepsilon$, dan
ôf h verkleinen in het R.K. proces
ôf uitgaand van y_k of y_{k-1} met de methode van Newton
$H(t_k,x) = 0$ oplossen.

2.4. Geschiedenis

Continueringsmethoden zijn al lang geleden toegepast in verschillende takken van de wiskunde (zie b.v. [3])

Schwarz (1869)
Hadamard (1906)
Bernstein (1906).

Bij deze toepassingen was niet de curve zelf interessant, maar alleen de existentie. Het betrof hier ook existentiebewijzen voor bepaalde problemen.

2.4.1 Numerieke toepassingen

Eerste toepassing was van LAHAYE (1934), bij het oplossen van een enkele vergelijking.

DAVIDENKO (1953) voerde de oplossingsmethode m.b.v. differentiaalvergelijkingen in, waarna er vele toepassingen zijn gekomen:

integraalvergelijkingen matrix-inversie determinant berekening matrix-eigenwaardenprobleem stelsels vergelijkingen	Davidenko (zie b.v. [4])
tweepuntsrandwaardeprobleem	Roberts & Shipman en Bosarge

GAVURIN (1958) was de eerste, die methode (2.2.2) gebruikte.
Hiermee bewees hij de existentie van de oplossing. Zie ook [1], hoofdstuk 7.5, voor meer referenties.

2.5. Convergentiestralen

Inbeddingsmethoden zijn ook op een andere manier te gebruiken. Namelijk als volgt:
Zij H(t,x) een inbedding, zodat geldt:

$$H(0,x_0) = 0$$
$$H(1,x) \equiv F(x).$$

Om onze gedachten te bepalen nemen we aan:

$$H(t,x) = (1-t)\, K\,(x,x_0) + t\, F(x) \qquad t \in [0,1];\ x \in \mathbb{R}^n$$

en

$$K(x,x) = 0 \qquad \forall x \in X$$

(bijv. $K = F(x) - F(y)$, $K = x - y$ of $K = L(x) - L(y)$) en L een willekeurige afbeelding.
We kunnen wéér de d.v. construeren:

$$H_t(t,x) + H_x(t,x) \cdot \dot{x}(t) = 0$$
$$H(t,x) = 0,$$

waaruit we kunnen vormen:

$$H_t(t,x) + H_x(t,x) \cdot \dot{x}(t) + g(t) \cdot H(t,x) = 0,$$

waarbij $g : [0,1] \to \mathbb{R}$ een gegeven funktie is; de d.v. is dan:

$$\begin{cases} \dot{x}(t) = -H_x(t,x))^{-1}\, [H_t(t,x(t)) + g(t) \cdot H(t,x(t))] \\ x(0) = x_0. \end{cases}$$

Deze d.v. kunnen we m.b.v. een Runge-Kutta-methode oplossen en vinden dan

$$x_1 \approx x(1) = x^*, \qquad x_1 = G(x_0).$$

Dit proces kunnen we herhalen met x_1 i.p.v. x_0:

$$x_2 \approx x(1), \qquad x_2 = G(x_1), \text{ etc.}$$

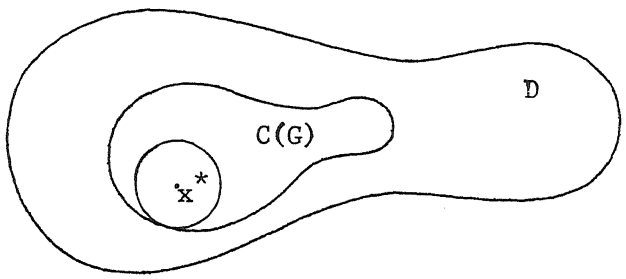
M.a.w. we hebben een iteratieve methode geconstrueerd. De hierboven genoemde G is een funktie van x, F, K, g en de gekozen Runge-Kutta methode.

We willen nu de verschillende iteratie-processen

$$\{x_{k+1} = G(x_k)\} ,$$

waarbij K, g en de Runge-Kutta-methode verschillen, met elkaar gaan vergelijken. Dit kunnen we bijvoorbeeld als volgt doen:

(2.5.1) Neem een vaste F_0, $F_0 : \mathbb{R}^n \to \mathbb{R}^n$, zodat $F_0(x) = 0$ een unieke oplossing x^* heeft. Voor gegeven K, g en Runge-Kutta-methode is
$G(x) = G(x, F_0, K, g, \text{Runge-Kutta-methode})$
$G : D \to \mathbb{R}^n$, $D \subset \mathbb{R}^n$



Het convergentiegebied van G wordt gedefinieerd door:
$C(G) = \{x_0 \mid \text{voor } \{x_k\} \text{ met } x_{k+1} = G(x_k),\ k = 0,\ldots;\ x_k \to x^*\}$.
Zij $B(x^*,\rho)$ het grootste bolletje, dat nog juist bevat is in C(G).
Noem de straal van dit bolletje de convergentiestraal.
Stel dan: hoe groter de convergentiestraal van een G, voor een gegeven kombinatie van K, g en de Runge-Kutta-methode, hoe "beter" deze kombinatie is.

Nadeel: alles wordt vastgepind op één funktie, (nl. F_0).

Een tweede mogelijkheid.

(2.5.2) Zij $F^* = \{F \mid F : \mathbb{R}^n \to \mathbb{R}^n,\ F(x) = 0,$ unieke oplossing $x^*\}$.
Bepaal voor gegeven kombinatie (K, g, Runge-Kutta-methode) voor iedere $F \in F^*$ de convergentiestraal van de geassocieerde G.
Associeer het infinum van de zo verkregen getallenverzameling met de kombinatie (K, g, Runge-Kutta-methode). Dit getalletje zou een aardig vergelijkings-kriterium zijn, ware het niet dat voor zekere $F \in F^*$ de geassocieerde G niet bestaat. (Dit geeft dus weinig informatie). Hieruit volgt dat de convergentiestraal nul is.

(2.5.3) Deze problemen zijn als volgt te omzeilen: pas (2.5.2) toe op een deelklasse $F_1 \subset F^*$.
Dus voor gegeven kombinatie (K, g, Runge-Kutta-methode) en voor alle $F \in F_1$ kunnen we dan de convergentiestraal van G bepalen (r(G)). Neem het infinum van deze getallenverzameling en associeer dat met (K, g, Runge-Kutta-methode).

Hoe groter het getalletje is voor deze kombinatie (K, g, Runge-Kutta-methode) hoe "beter" deze kombinatie is.
Laten we ons beperken tot de volgende deelverzameling F_1 van F^*.
Zij gegeven $\beta, \gamma > 0$,

$$F < \beta, \gamma > = \{F \mid F : \mathbb{R}^n \to \mathbb{R}^n,\ F(x) = 0,\ \text{unieke oplossing } x^*,$$
$$\|F'(x) - F'(y)\| \le \gamma \|x-y\|,\ \|F'(x^*)^{-1}\| \le \beta\}.$$

en K "netjes".
Neem (K, g, Runge-Kutta-methode). Wil een (K, g, Runge-Kutta-methode) kombinatie in aanmerking komen voor een goede beoordeling, dan zal toch minstens de convergentiestraal van G positief moeten zijn en dit betekent dat voor x_0 dichtbij x^*, $\{x_k\}$ naar x^* moet convergeren, oftewel het proces moet lokaal convergent zijn.

Stelling 2.5.1.

Als $F \in F < \beta, \gamma >$ en K netjes, dan zijn de volgende drie uitspraken equivalent.

1. Voor gegeven g en elke Runge-Kutta-methode is het iteratieve proces geassocieerd met G, lokaal kwadratisch convergent.
2. Voor elke Runge-Kutta-methode en elke g is het iteratieve proces, geassocieerd met G, lokaal monotoon convergent.
3. $F'(x^*) = K_1(x^*,x^*)$ $\quad (K_1(x,y) = \frac{\partial}{\partial x} K(x,y))$.

Voorbeeld:

$$K(x,y) = F(x) - F(y)$$
$$= F'(y)(x-y)$$

Opmerking:

Als $K = F(x) - F(y)$, $g \equiv 0$, dan levert de methode van Euler met $h=1$, de methode van Newton.

Stelling 2.5.2.

De convergentiestraal van de methode van Newton (t.o.v. $\bar{F} < \beta, \gamma >$) is $\frac{2}{3\beta\gamma}$.

Stelling 2.5.3.

Laat $F \in \bar{F} < \beta, \gamma >$ en (K, g, Runge-Kutta-methode) zó dat

$$G(x) = y_m(x)$$

met

$$y_1(x) = x_0 - \alpha_1 F'(x_0)^{-1} \cdot F(x_0)$$
$$y_{k+1}(x) = y_k(x) - \alpha_{k+1} F'(y_k(x))^{-1} \cdot F(y_k(x)),$$

voor $k = 1,2,\ldots,m$ en $\alpha_i \in (0,1)$ $i = 1,2,\ldots,m$, met $\alpha_{m+1} = 1$ (dit is mogelijk), dan heeft (K, g, Runge-Kutta-methode) een grotere convergentiestraal dan de methode van Newton.

2.6. Opmerkingen

Zij $F : \mathbb{R} \to \mathbb{R}$ en $F'(x) \neq 0 \quad \forall x \in (x^*-\varepsilon, x^*+\varepsilon)$

en F een "nette" funktie.
Volgens de methode van Newton vinden we een x_1 door

$$x_1 = x - F'(x)^{-1} F(x).$$

In $\mathbb{R}$ geldt, dat de richting van de convergentie afhankelijk is van het teken van $(x-x^*)$. Indien

$$(x-x^*) > 0 \Rightarrow F'(x)^{-1} \cdot F(x) > 0$$
$$(x-x^*) < 0 \Rightarrow F'(x)^{-1} \cdot F(x) < 0.$$

Nu bestaat het vermoeden, dat dit voor "nette" funkties in $\mathbb{R}^n$ ook geldt, d.w.z.

$$(x-x^*,\ F'(x)^{-1} \cdot F(x)) > 0$$

waarbij onder "nette" funkties verstaan wordt, funkties die tenminste aan de volgende eisen voldoen:

(2.6.1) $F'(x)^{-1}$ bestaat voor alle $x \in B(x^*,\varepsilon)$ en
(2.6.2) $F'(x)$ is Lipschitz continu in x, voor alle $x \in B(x^*,\varepsilon)$.

Voor het volgende type funkties is het bovenstaand vermoeden bewezen.
Pas een eindige differentie-methode toe op het tweepuntsrandwaarde probleem:

$$u'' + f(t,u) = 0$$
$$u(0) = a,\ u(1) = b.$$

Dit komt neer op het oplossen van een vergelijking $F(x) = 0$, waarbij

$$F : \mathbb{R}^n \to \mathbb{R}^n$$
$$F(x) = Ax + \phi(x).$$

Met wat eisen op $f(t,u)$ kunnen we aantonen:

$$(x-x^*,F'(x)^{-1} \cdot F(x)) > 0 \quad (x^* \text{ is de unieke oplossing van } F(x) = 0).$$

Verder merken we op:
Zij F een funktie van $\mathbb{R}^n \to \mathbb{R}^n$, F voldoet aan (2.6.1) en (2.6.2), dan geldt voor $x_1 = x - F'(x)^{-1} \cdot F(x)$, met $x \in B(x^*, \varepsilon)$:

$$0 \approx x_1 - x^* \text{ (wegens kwadratische convergentie)},$$

$$\Rightarrow 0 \approx \|x_1 - x^*\|^2 = \|x-x^*\|^2 + \|F'(x)^{-1} \cdot F(x)\|^2 - 2(x-x^*, F'(x)^{-1} \cdot F(x))$$

$$\Rightarrow (x-x^*, F'(x)^{-1} \cdot F(x)) \approx \tfrac{1}{2} \|x-x^*\|^2 + \tfrac{1}{2} \|F'(x)^{-1} \cdot F(x)\|^2.$$

Beschouw nu die funkties F waarvoor geldt:

$$(2.6.3) \qquad (x-x^*, F'(x)^{-1} \cdot F(x)) \geq a\{\|x-x^*\|^2 + \|F'(x)^{-1} \cdot F(x)\|^2\}$$

met

$$x \in \mathbb{R}^n, \qquad a \leq \tfrac{1}{2}.$$

Voor $a > \frac{1}{2}$ is (2.6.3) een onmogelijke eis.
Als $a \in (0.4, \frac{1}{2}]$, dan is er convergentie met de methode van Newton voor alle startwaarden in $\mathbb{R}^n$.
Als $a \leq 0.4$ dan bestaan er funkties F, die aan (2.6.3) voldoen, zodat voor zekere x_0, er geen convergentie optreedt.

Bepalen we x_k via een tussenberekening:

$$y_k = x_{k-1} - \alpha F'(x_{k-1})^{-1} \cdot F(x_{k-1}),$$
$$x_k = y_k - F'(y_k)^{-1} \cdot F(y_k), \qquad \text{voor } k = 1,2,\ldots,$$

dan blijkt $\alpha = \frac{1}{2}\sqrt{2}$ de beste keuze te zijn: er geldt dan het volgende:

als $a \in (\frac{1}{4}\sqrt{2}, \frac{1}{2}]$, dan is er convergentie voor iedere x_0 in $\mathbb{R}^n$.
als $a \leq \frac{1}{4}\sqrt{2}$, dan bestaan er weer funkties F die aan (2.6.3) voldoen zodat voor zekere x_0 er geen convergentie optreedt.

Voor $\alpha \in [0,1]$ en $\alpha \neq \frac{1}{2}\sqrt{2}$ bestaat er een $a \in (\frac{1}{4}\sqrt{2}, \frac{1}{2}]$, zodanig dat er weer funkties zijn, die aan (2.6.3) voldoen, waarbij geen convergentie optreedt.

REFERENTIES

[1] ORTEGA, J.M. & W.C. RHEINBOLDT, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York and London, 1970.

[2] AVILA, J.H. Jr., *The feasibility of continuation methods for nonlinear equations*, SIAM J. Numer. Anal. vol 11 (1974) 102-122.

[3] FICKEN, F, *The continuation method for functional equations*, Comm. Pure Appl. Math. 4 (1951) 435-456.

[4] RALL, L.B., *Davidenko's method for the solution of nonlinear operator equations*, The Univ. of Wisconsin, Mathematics Research Center, MRC Techn. Summary Rept. 948, October 1968.

Algemene Werkbespreking Afdeling Numerieke Wiskunde

Hoofdstuk 3

Algoritmen voor het oplossen van stelsels niet-lineaire vergelijkingen

door

J.C.P. Bus

3.1. Inleiding

We beschouwen een niet-lineaire funktie

(3.1.1) $\quad F: \mathbb{R}^n \to \mathbb{R}^n.$

We stellen ons het probleem een punt $x^* \in \mathbb{R}^n$ te bepalen zodat

$$F(x^*) = 0.$$

Het is duidelijk dat niet altijd een oplossing van dit probleem bestaat, en evenzo kunnen er meerdere, of zelfs oneindig veel, oplossingen bestaan. Dit maakt de konstruktie van algoritmen voor het bepalen van een oplossing van een niet-linear stelsel aanzienlijk gekompliceerd.

We zullen daarom eerst voorwaarden voor existentie en éénduidigheid formuleren. Deze zijn gebaseerd op resultaten uit RHEINBOLDT [1968] en ORTEGA & RHEINBOLDT [1970].

Vervolgens zal een "globale" konvergentiestelling worden gegeven voor een ruime klasse van Newton-achtige algoritmen voor de berekening van oplossingen van niet-lineaire stelsels. Deze stelling is een uitbreiding van een stelling van DEUFLHARD [1974a], [1974b] en vormt de basis voor een aantal Newton-achtige algoritmen, die zullen worden vergeleken aan de hand

van een aantal testresultaten.

3.2. Notaties en definities

Zij weer $F: \mathbb{R}^n \to \mathbb{R}^n$. We zullen de afgeleide van F noteren als $J(x) = F'(x)$, deze wordt de jacobiaan van F genoemd.

Voor een willekeurige $A \subset L(\mathbb{R}^n)$, met A niet-singulier, definiëren we:

levelfunktie: $T(x;A) = (AF(x),AF(x)), \quad x \in \mathbb{R}^n,$

levelset: $S(x_0;A) = \{x \mid x \in \mathbb{R}^n,\ T(x_0;A) \geq T(x;A)\}.$

Met $\hat{S}(x_0;A)$ wordt de boogsamenhangende deelverzameling van $S(x_0;A)$ bedoeld, die x_0 bevat. We zullen in het vervolg gebruik maken van drie hier te definiëren standaardeigenschappen van de funktie.

(I) We zeggen dat de funktie *eigenschap I* heeft op een open verzameling $D \subset \mathbb{R}^n$ als:

$J(x)$ bestaat en kontinu is op D,

$J^{-1}(x)$ bestaat op D,

er bestaan positieve getallen ϕ en β zodat $\|J(x)\| \leq \phi$, $\|[J(x)]^{-1}\| \leq \beta$ voor alle $x \in D$.

(II) We zeggen dat de funktie *eigenschap II* heeft (Lipschitz kontinue jacobiaan) op een open verzameling $D \subset \mathbb{R}^n$ als

er een positief getal γ bestaat zodat

$\|J(x) - J(y)\| \leq \gamma\|x - y\|$, voor alle $x,y \in D$.

(III) We zeggen dat de funktie *eigenschap III* heeft voor een zekere niet-singuliere $A \subset L(\mathbb{R}^n)$ en een zekere $x_0 \in \mathbb{R}^n$ als $\hat{S}(x_0;A)$ kompakt is.

Tenslotte kombineren we deze drie eigenschappen als volgt: Beschouw het probleem een oplossing te vinden van de vergelijking $F(x) = 0$, bij gegeven beginschatting $x_0 \in \mathbb{R}^n$. Dit probleem noteren we als $\{F,x_0\}$.

We noemen $\{F,x_0\}$ *goed gedefinieerd* op een open verzameling $D \in \mathbb{R}^n$ als

a) er een niet-singuliere $A \subset L(\mathbb{R}^n)$ bestaat zodat F eigenschap III heeft voor deze A en x_0;

b) $\hat{S}(x_0;A) \subset D$ en F heeft eigenschap I op D.

De reden voor deze definitie zal blijken in de volgende paragraaf.

Tenslotte nog enige notaties:

Als F de eigenschappen I, II en III heeft, dan zijn de getallen ϕ, β en γ en een matrix A gegeven; we schrijven dan:

$$\beta(x) = \|J^{-1}(x)\|, \quad \alpha(x) = \beta^2\gamma\|F(x)\|,$$

$$\kappa(A) = \|A\|\|A^{-1}\|, \quad \kappa(x) = \|J(x)\|\|[J(x)]^{-1}\|,$$

$$\kappa(x;A) = \|AJ(x)\|\|[AJ(x)]^{-1}\|.$$

3.3. Existentie

De volgende stelling kan worden bewezen.

STELLING 3.3.1. *Als het probleem* $\{F,x_0\}$ *goed gedefinieerd is op een open verzameling* $D \in \mathbb{R}^n$, *dan bestaat er een éénduidige oplossing* $x^* \in \hat{S}(x_0;A)$ *van de vergelijking* $F(x) = 0$.

BEWIJS. Dit bewijs is gebaseerd op het bewijs van de lokale versie van de stelling van Hadamard (zie ORTEGA & RHEINBOLDT [1970], st. 5.3.11) en de kontinuatie eigenschap. Het zal hier achterwege worden gelaten. □

3.4. Approximatie van de jacobiaan

Als F eigenschap I en II heeft voor een zekere open konvexe verzameling D, dan geldt voor alle $x,y \in D$:

$$\|F(y) - F(x) - J(x)(y-x)\| \leq \frac{1}{2}\gamma\|y-x\|^2.$$

Wanneer we nu een benadering B(x) van J(x) willen berekenen dan lijkt het dus redelijk te eisen dat de verkregen benadering voldoet aan

$$B(x)(y-x) = F(y) - F(x), \quad y \neq x, \quad y,x \in \mathbb{R}^n. \tag{3.4.1}$$

Hiermee is echter slechts de beperking van B(x) tot de deelruimte opgespannen door y - x bepaald. Zij nu $\Delta \subset L(\mathbb{R}^n)$ gegeven en niet-singulier. Definieer

$$B(x) = \Gamma(x)\Delta^{-1} \tag{3.4.2}$$

met

$$\Gamma(x) = (F(x+\Delta e^1) - F(x), \ldots, F(x+\Delta e^n) - F(x))$$

dan is $B(x)$ éénduidig bepaald en de volgdende stelling geldt:

STELLING 3.4.1. *Stel* F *heeft eigenschappen* I *en* II *voor een zekere open verzameling* D. *Stel dat* $\Delta = (h^1, \ldots, h^n) \subset L(\mathbb{R}^n)$ $(h^i \in \mathbb{R}^n, h^i \neq 0)$ *voldoet aan*

$$\left\| \left(\frac{h^1}{\|h^1\|}, \ldots, \frac{h^n}{\|h^n\|} \right)^{-1} \right\| \leq \sigma$$

voor zekere $\sigma > 0$. *Dan bestaat er voor alle* $x \in D$ *een* $r > 0$, *zodat als* $\|\Delta\| < r$ *en* $B(x)$ *is gedefinieerd door (3.4.2) dan*

$$(3.4.3) \qquad \|B(x) - J(x)\| \leq \frac{1}{2} \sqrt{n}\, \sigma\gamma \max_i \|h^i\|.$$

BEWIJS. Zie ORTEGA & RHEINBOLDT [1970]. □

VOORBEELD. (voorwaartse differentie benadering).

$$\Delta = \operatorname{diag}(h^{(1)}, h^{(2)}, \ldots, h^{(n)}) \qquad (h^{(i)} \in \mathbb{R},\ i = 1, \ldots, n)$$

dan geldt $\sigma = 1$ in stelling 3.4.1.

Een methode die eveneens is gebaseerd op (3.4.1) is de z.g. *modifikatie methode*. Zij gegeven een benadering $B(y)$ van $J(y)$ voor zekere $y \in \mathbb{R}^n$. Beschouw voor willekeurige $v \in \mathbb{R}^n$ met $v^T(F(x) - F(y)) \neq 0$ en zekere $x \in \mathbb{R}^n$ de benadering van $J(x)$:

$$(3.4.4) \qquad B(x) = B(y) + \frac{[F(x) - F(y) - B(y)(x-y)]v^T B(y)}{v^T B(y)(x-y)} .$$

Dan voldoet $B(x)$ aan (3.4.1).
Kiezen we v zodanig dat $B^T(y)v = x - y$ dan verkrijgen we, wat wel wordt genoemd, Broyden's update formule.
De volgende stelling kan worden gegeven (DENNIS [1971]).

STELLING 3.4.2. *Stel* F *heeft de eigenschappen* I *en* II *voor zekere open konvexe verzameling* $D \subset \mathbb{R}^n$. *Stel* B_0 *is een approximatie van* $J(x_0)$ *voor zekere* $x_0 \in D$. *Stel* $\{x_i\}_{i=0}^{n+1} \subset D$ *is gegeven en* $\{B_i\}_{i=0}^{n+1}$ *is gedefinieerd door (3.4.4). Dan geldt:*

$$\|B_{n+1} - J(x_{n+1})\| \leq \Big(\prod_{i=0}^{n} \xi_i\Big)\|B_0 - J(x_0)\|$$

$$+ \gamma \sum_{j=0}^{n} \Big(\prod_{i=-1}^{n-j-1} \xi_i\Big)\Big(1+\frac{1}{2}\xi_j\Big)\|x_{j+1} - x_j\|,$$

met $\xi_{-1} = 1$ en

$$\xi_i = \frac{\|x_{i+1} - x_i\|\,\|v_i^T B_i\|}{|v_i^T B_i (x_{i+1} - x_i)|}$$

OPMERKING. Voor Broyden's update formule geldt dat $\xi_i = 1$, $i = 0,\ldots,n$.

3.5. Newton-achtige methoden met demping

We beschouwen de klasse van iteratieve methoden gedefinieerd door

$$x_{k+1} = x_k - \lambda_k B_k^{-1} F_k, \quad k = 0,1,2,\ldots, \tag{3.5.1}$$

bij gegeven $x_0 \in \mathbb{R}^n$ en waarbij $F_k = F(x_k)$ en λ_k is een scalar met $0 < \lambda_k \leq 1$. Dergelijke methoden noemen we *Newton-achtige methoden* en als niet $\lambda_k = 1$ $(k = 0,1,\ldots)$ dan spreken we van een *methode met demping*. We verkrijgen Newton's methode als we kiezen

$$\lambda_k = 1, \quad B_k = J(x_k) \quad (k = 0,1,2,\ldots).$$

We zeggen dat een methode *dalend* is met betrekking tot een niet-singuliere $A \in L(\mathbb{R}^n)$ als geldt:

$$T(x_{k+1};A) \leq T(x_k;A), \quad k = 1,2,\ldots . \tag{3.5.2}$$

We kiezen nu een vaste niet-singuliere $A \in L(\mathbb{R}^n)$. Definieer

$$\theta_k = \beta(x_k)\|B_k - J(x_k)\|. \tag{3.5.2}$$

Dan geldt de volgende stelling.

STELLING 3.5.1. *Stel* F *heeft eigenschap* I *en* II *op een zekere open verzameling* D. *Stel voor zekere* $k > 0$ *geldt:*

$$\theta_k < 1/(1+\kappa(A)). \tag{3.5.4}$$

Dan bestaat B_k^{-1} *en is er een* λ_k $(0 < \lambda_k \leq 1)$ *zodat als* x_{k+1} *wordt gedefinieerd door (3.5.1),*

$$T(x_{k+1};A) \leq T(x_k;A).$$

De volgende uitbreiding van een stelling van DEUFLHARD [1974b] geeft een scherper resultaat.

STELLING 3.5.2. *Stel* $\{F,x_0\}$ *is een goed gedefinieerd probleem op een open verzameling* $D \subset \mathbb{R}^n$ *en* F *heeft eigenschap* II *op* D. *Stel* $x_k \in \hat{S}(x_0;A)$ *voor zekere* $k > 0$. *Stel*

$$2\theta_k < \min(1/(1+\kappa(A)),\ 1/(1+\kappa(x;A))). \tag{3.5.5}$$

Dan bestaat B_k^{-1} *en*

$$T(x_k + \lambda B_k^{-1}F_k;A) \leq t_k^2(\lambda;A)T(x_k;A) \tag{3.5.6}$$

voor

$$0 \leq \lambda \leq \min(1,\mu_k(A)), \tag{3.5.7}$$

waarbij

$$t_k(\lambda;A) = 1 - \lambda\psi(\lambda\alpha(x_k)) \tag{3.5.8}$$

$$\mu_k(A) = \min\left(1, \frac{\rho_k}{\alpha(x_k)}\right) \tag{3.5.9}$$

$\psi(t)$ *een kwadratische funktie is met een positieve wortel* ρ_k, *een negatieve wortel* ρ_k' *en* $\psi(t) \geq 0$ *voor* $\rho_k' \leq t \leq \rho_k$.

OPMERKING. Deuflhard geeft een speciaal geval van deze stelling, en wel het geval dat $B_k = J(x_k)$. Een gevolg van stelling 3.5.2 is de volgende "globale" konvergentiestelling voor dalende Newtonachtige methoden met demping.

STELLING 3.5.3. *Stel* $\{F,x_0\}$ *is een goed gedefinieerd probleem op een open verzameling* $D \subset \mathbb{R}^n$ *en* F *heeft eigenschap* II *op* D. *Stel* B_k *is gedefinieerd en niet-singulier voor* $k = 0,1,\ldots$. *Zij* $\{x_k\}$ *gedefinieerd door (3.5.1) en stel* θ_k *voldoet aan (3.5.5),*

$$\lambda_{min} \leq \lambda_k \leq \bar{\mu}_k(A) \tag{3.5.10}$$

met

$$\bar{\mu}_k(A) = \min(1,\ \mu_k(A) - \lambda_{min}) \tag{3.5.11}$$

voor zekere vaste λ_{min} *met*

$$0 < \lambda_{min} \leq \min(1, 1/(2\alpha(x_0)\kappa(A)(2m + \sqrt{m}))) \tag{3.5.12}$$

en

$$m = \max_{x \in \hat{S}(x_0;A)} \kappa(x;A). \tag{3.5.13}$$

Dan geldt:

a) $\{x_k\} \subset \hat{S}(x_0;A)$

b) $T(x_k;A) > 0 \Rightarrow T(x_{k+1};A) < (1 - \lambda_{min}^2/8)T(x_k;A)$

$$T(x_{k+1};J^{-1}(x_k)) < (1 - \lambda_{min}^2/8)T(x_k;J^{-1}(x_k))$$

c) *Er bestaat een unieke oplossing* $x^* \in \hat{S}(x_0;A)$ *met*

$$\lim_{k\to\infty} x_k = x^*; \quad F(x^*) = 0.$$

OPMERKING. We geven twee manieren om gebruik te maken van stelling 3.5.3 ter bepaling van een geschikte λ_k.

1. Men kan door halveren proberen een λ_k te vinden zodat voor zekere $\mu (0 < \mu < 1)$ geldt:

$$T(x_{k+1};A) < (1 - \mu)T(x_k;A);$$

kiezen we $\mu = \lambda_{min}^2/8$, dan is aan deze voorwaarde voldaan voor alle λ_k die aan (3.5.10) voldoen. Kiezen we niet λ_{min} gelijk aan de bovengrens gegeven door (3.5.12) maar hoogstens de helft daarvan, dan kan de gezochte λ_k door halvering worden verkregen en verkrijgt men een konvergent proces. Deze methode is te vergelijken met de methode van GOLDSTEIN & PRICE [1967] voor minimaliseren van funkties, zie ook FLETCHER [1970].

2. De maximale waarde voor $\psi(s)$ in stelling 3.5.2 wordt verkregen voor

$$s = s_k^* = \frac{1}{3}\left[-\eta_k + \sqrt{\eta_k^2 + \frac{6[1-(\kappa(x;A)+1)\theta_k]}{\kappa(x;A)}}\right], \quad \text{met}$$

$$\eta_k = (1+\theta_k^2)/(1-\theta_k)$$

De keus $A = J^{-1}(x_k)$ levert dan, wat we noemen de optimale waarde voor λ_k:

$$\lambda_k^* = \min\left(1, \frac{1}{3\alpha(x_k)}\left[-\eta_k + \sqrt{\eta_k^2 + 6(1-2\theta_k)}\right]\right). \tag{3.5.14}$$

Deze keuze voldoet aan (3.5.10) zodat hiermede ook een konvergent proces wordt verkregen. Gebruik van (3.5.14) vereist berekening van $\alpha(x_k)$ en θ_k.

3.6. Enige praktische Newton-achtige methoden

Op basis van de theoretische resultaten in de vorige paragrafen kunnen we enige praktische Newton-achtige methoden construeren. Beschouwen we formule (3.5.1) dan rest ons nog de invulling van de methoden ter bepaling van λ_k en van B_k $(k = 0,1,\ldots)$. We zullen eerst 3 methoden bespreken voor berekening van B_k.

Newton's algoritme (met demping). We kiezen hier $B_k = J(x_k)$. Dus $\theta_k = 0$ en voor de gegeven F is altijd aan (3.5.5) voldaan.

Diskrete Newton algoritme (met demping). B_k wordt nu gelijk aan de voorwaartse differentiebenadering in x_k gekozen (vgl. §3.4). Dan geldt volgens stelling 3.4.1.

$$\|B_k - J(x_k)\| \leq \frac{1}{2}\sqrt{n}\,\gamma\,h_k, \quad k = 1,2,\ldots \tag{3.6.1}$$

waarbij h_k een vaste staplengte is voor alle koördinaatrichtingen $(h_k \neq 0)$.

Als de voorwaartse differenties numeriek worden berekend dan moet rekening worden gehouden met het wegvallen van cijfers bij het aftrekken van bijna gelijke funktiewaarden. Noteren we de numeriek, met rekenprecisie ε, berekende benadering met $fl_\varepsilon(B_k)$ dan geldt

$$(3.6.2) \qquad \|fl_\varepsilon(B_k)-J(x_k)\| \leq \sqrt{n}\left(\frac{1}{2}\gamma h_k+\frac{2(\delta_1+\varepsilon)\|F_k\|+2\delta_2}{h_k}\right),$$

waarbij δ_1 en δ_2 zo zijn gekozen dat

$$(3.6.3) \qquad \|fl_\varepsilon(F(x)) - F(x)\| \leq \delta_1\|F(x)\|+\delta_2.$$

Uit (3.6.2) volgt dat de meest geschikte keuze voor h_k de waarde is, die beide termen in het rechterlid gelijk maakt. Dus

$$(3.6.4) \qquad h_k = 2\sqrt{((\delta_1+\varepsilon)\|F_k\|+\delta_2)/\gamma}$$

en

$$(3.6.5) \qquad \|fl_\varepsilon(B_k) - J(x_k)\| \leq \sqrt{n}\ \gamma h_k.$$

Modifikatiemethode

We gaan hierbij uit van Broydens's update formule (vgl.(3.4.4)). Dan geldt wegens stelling 3.4.2 en de opmerking daarna, voor gegeven B_k:

$$\|B_{k+i} - J(x_{k+i})\| \leq \|B_k - J(x_k)\| + \frac{3}{2}\gamma \sum_{j=k}^{k+i-1} \|x_{j+1} - x_j\|.$$

Dus

$$(3.6.6) \qquad \theta_{k+i} \leq \theta_k + \frac{3}{2}\beta\gamma \sum_{j=k}^{k+i-1} \|x_{j+1}-x_j\|.$$

In de praktijk zal deze methode worden gekombineerd met berekening met behulp van voorwaartse differenties. Als kriterium om gebruik te kunnen maken van de update formule kunnen we ongelijkheid (3.5.5) nemen. Dus zij B_k een voorwaartse differentiebenadering (numeriek) dan kunnen we voor $i = 1,\ldots,m$, B_{k+i} berekenen met de modifikatiemethode als

$$(3.6.7) \qquad \theta_k + \frac{3}{2}\beta\gamma \sum_{j=k}^{k+i-1} \|x_{j+1}-x_j\| \leq \min(1/(1+\kappa(A)),\ 1/(1+\kappa(x_{k+i};A)))$$

anders wordt B_{k+i} met voorwaartse differenties berekend en beginnen we opnieuw.

Voor de bepaling van λ_k gaven we, in de opmerking aan het eind van §3.5, al twee methoden.

Halvering

Bepaal λ_k door halvering zodat geldt, voor zekere μ $(0 < \mu << 1)$:

$$T(x_{k+1};A) < (1-\mu)T(x_k;A). \tag{3.6.8}$$

Schatting van de optimale stap

We kiezen $\lambda_k = \lambda_k^*$ (zie (3.5.14)).

Voor een aantal van de bovengenoemde zaken, is voor de berekening, de waarde van grootheden zoals β, γ en

$$e_k = \|B_k - J(x_k)\| \tag{3.6.9}$$

nodig. Deze zijn over het algemeen niet beschikbaar en we zullen daarom een methode om deze te benaderen bespreken.

Noteren we

$$d_k = -B_k^{-1}F_k, \quad \hat{d}_k = -B_{k-1}^{-1}F_k, \quad \hat{\beta}_k = \|B_k^{-1}\|. \tag{3.6.10}$$

Dan geldt:

$$\begin{aligned} \|\hat{d}_k - d_k\| &\leq \|B_k^{-1}\|\|B_k - B_{k-1}\|\|\hat{d}_k\| \\ &\leq \hat{\beta}_k\|\hat{d}_k\|[\gamma\|x_k - x_{k-1}\| + e_k + e_{k-1}]. \end{aligned}$$

Met de definitie $\hat{\alpha}_k = \hat{\beta}_k\gamma\|\hat{d}_k\|$ krijgen we

$$\hat{\alpha}_k \geq \frac{\|\hat{d}_k - d_k\|}{\|x_k - x_{k-1}\|}\frac{1}{1+E_k} = \tilde{\alpha}_k, \tag{3.6.11}$$

met

$$E_k = \frac{e_k + e_{k-1}}{\gamma\|x_k - x_{k-1}\|}. \tag{3.6.12}$$

Stellen we

$$e_i = c_i\gamma \qquad i = 1,2,\ldots \tag{3.6.13}$$

waarbij c_i bekend is (afhankelijk van de gekozen benadering B_i van $J(x_i)$), dan krijgen we

$$(3.6.14) \qquad E_k = \frac{c_k + c_{k-1}}{\|x_k - x_{k-1}\|} .$$

Dus E_k kan worden berekend en daarmee $\tilde{\alpha}_k$. Wegens de definitie van $\hat{\alpha}_k$ geldt

$$(3.6.15) \qquad e_k = \frac{c_k \hat{\alpha}_k}{\hat{\beta}_k \hat{d}_k}$$

zodat ook e_k kan worden berekend. Voor θ_k kunnen we de volgende benadering geven:

$$(3.6.16) \qquad \theta_k \approx \hat{\theta}_k = \frac{\hat{\beta}_k e_k}{1+\hat{\beta}_k e_k},$$

en voor α_k:

$$(3.6.17) \qquad \alpha_k \approx \tilde{\tilde{\alpha}}_k = (1-\hat{\theta}_k)(1-\hat{\theta}_{k-1})\tilde{\alpha}_k.$$

De grootheden in (3.6.15), (3.6.16) en (3.6.17) worden gebruikt bij de berekening van de verschillende grootheden in de algoritmen.

We beschouwen nu de volgende algoritmen:

A. Newton's algoritme

$B_k = J(x_k)$, $\lambda_k = 1$, $k = 0,1,\ldots$.

B. Newton's algoritme met halvering

$B_k = J(x_k)$, λ_k berekend door halvering (zie (3.6.8)).

C. Newton's algoritme met optimale staplengtekeuze

$B_k = J(x_k)$, $\lambda_k = \lambda_k^*$ (zie (3.5.14)), waarbij de geschatte waarden voor α_k worden gebruikt, $\theta_k = 0$.

D. Diskrete Newton algoritme

B_k berekend met voorwaartse differenties en h_k volgens (3.6.4), $\lambda_k = 1$.

E. Diskrete Newton algoritme met optimale staplengte strategie

B_k wordt berekend met voorwaartse differenties en h_k volgens (3.6.4), $\lambda_k = \lambda_k^*$ (zie 3.5.14).

F. Een Newton-achtige algoritme verkregen door kombinatie

B_k wordt berekend met voorwaartse differenties (h_k met (3.6.4)), of door modifikatie van B_{k-1}^{-1} met Broyden's update formule voor de inverse matrix. Als kriterium wordt (3.6.7) gebruikt met $\kappa(A) = \kappa(x_{k+i};A) = 1$, $\lambda_k = \lambda_k^*$ (zie (3.5.14)).

3.7. Numerieke resultaten en konklusies

We zullen hier wat resultaten geven van een aantal kleine testproblemen. Voor definitie van deze problemen zie BUS [1975].

We hebben ter vergelijking de algoritme P toegevoegd; deze is equivalent aan A, behalve dat nu het stelsel wordt opgelost met singuliere waardenontbinding. In de tabellen duidt "it" op het aantal benodigde iteraties, "fev" op het aantal komponenten van de funktievektor dat is geëvalueerd. De doorgekruiste hokjes duiden op falen.

Opvallend is dat de methoden zonder demping (A, D en P) minder falen dan die met demping. Er zijn verder enige speciale gevallen aan te wijzen:

1. funktie (2,2,1) start met een exact singuliere jacobiaan. De voorwaartse differentiebenadering is echter niet singulier;
2. funktie (8,2,1) en (24,2,0) geeft bij halvering van de eerste stap ook een exact singuliere jacobiaan; deze halvering gebeurt ook in C, omdat bij de eerste stap nog geen optimale λ kan worden berekend.

Laten we methode B, als meest falende, buiten beschouwing, dan constateren we dat, afgezien van bovenbeschreven uitzonderingen, de problemen zich beperken tot de problemen (3,2,0) en (3,2,1), (6,2,2), (7,2,1), (11,2,3), (20,2,0), (26,2,0) en (27,2,0) en in mindere mate voor funktie 10. Vier van deze 8 problemen worden door geen van de methoden opgelost. De overige worden niet opgelost door methoden met demping. Het is echter waard te vermelden dat ook de methoden zonder demping bij deze problemen relatief veel iteraties vergen. De indruk bestaat dat slechts door nogal willekeurige stappen in het begin, uiteindelijk toevallig een punt wordt bereikt van waaruit wel konvergentie mogelijk is.

Probleem			A		B		C		D		E		F		P	
fn	n	cn	it	fev	it	fev	it	fev	it	fev	it	fev	it	fev	it	fev
1	2	0	1	4	1	4	1	4	1	8	1	8	1	8	1	4
1	3	0	6	21	6	24	6	24	6	75	6	78	7	54	6	21
1	5	0	17	90	7	70	9	70	17	515	9	295	7	160	19	90
2	2	0	24	50	8	20	8	20	24	146	8	52	8	44	24	50
2	2	1	1	2	1	2	0	2	24	146	7	82	8	68	9	20
2	2	2	9	20	8	20	8	20	9	56	8	52	9	46	9	20
2	2	3	14	30	25	302	7	28	14	86	7	56	8	42	14	30
3	2	0	42	86	26	300	10	40	34	206	34	278	18	372	42	86
3	2	1	22	46	27	302	9	44	20	122	40	306	22	366	22	46
3	2	2	5	12	4	12	4	12	5	32	4	28	4	24	5	12
3	2	3	16	34	25	302	5	26	16	98	5	46	6	36	16	34
4	2	0	4	10	4	10	4	10	4	26	4	26	5	20	4	10
4	2	1	5	12	3	14	3	14	5	32	3	26	4	24	5	12
5	3	0	7	24	6	27	6	27	7	87	6	81	9	72	7	24
6	2	0	6	14	5	16	5	16	6	38	5	36	8	30	6	14
6	2	1	6	14	5	16	5	16	6	38	5	36	7	28	6	14
6	2	2	101	202	27	302	15	50	50	302	48	300	27	304	100	202
6	2	3	11	24	25	302	6	26	11	68	6	50	9	40	11	24
7	2	0	12	26	12	26	14	30	12	74	13	80	13	76	12	26
7	2	1	15	32	26	304	5	44	15	92	46	304	2	88	15	32
7	2	2	13	28	13	30	16	36	13	80	15	94	15	86	13	28
8	2	0	2	6	2	6	3	8	2	14	3	20	3	20	2	6
8	2	1	2	6	2	6	1	6	3	20	41	250	44	266	2	6
8	2	2	2	6	5	24	6	20	3	20	5	38	6	36	2	6
8	2	3	1	4	1	4	1	4	1	6	1	6	1	6	1	4
9	2	0	2	6	2	6	2	6	3	20	3	20	2	14	2	6
9	2	1	2	6	2	6	2	6	2	14	2	14	3	16	2	6
10	2	0	6	14	21	54	28	58	7	44	29	176	28	158	6	14
10	2	1	2	6	20	54	28	64	4	26	26	164	29	160	2	6
10	2	2	5	12	18	46	24	52	6	38	22	136	23	130	5	12
11	2	0	15	32	15	32	15	32	14	86	15	92	15	92	15	32

Probleem			A		B		C		D		E		F		P	
fn	n	cn	it	fev	it	fev	it	fev	it	fev	it	fev	it	fev	it	fev
11	2	1	13	28	13	28	13	28	13	80	13	80	13	80	13	28
11	2	2	15	32	15	32	16	34	15	92	16	98	16	98	15	32
11	2	3	17	36	25	302	12	46	17	104	45	300	20	354	17	36
12	4	0	19	80	19	80	19	80	19	384	19	384	19	384	19	80
14	2	0							4	26	4	26	5	20		
14	3	0							2	51	4	51	4	33		
14	4	0							6	124	5	108	7	84		
14	5	0							5	155	4	130	5	110		
16	5	0	4	25	4	25	4	25	4	125	4	125	7	90	4	25
19	2	0	5	12	5	14	5	14	5	32	5	34	6	28	5	12
20	2	0	101	102	27	306	17	46	50	302	31	304	22	384	100	202
21	3	0	2	9	2	9	2	9	2	27	2	27	2	27	3	12
22	3	0	101	303	9	30	9	30	9	111	9	111	14	99	9	30
23	2	0	2	6	2	6	2	6	3	20	3	20	3	16	2	6
24	2	0	5	12	2	6	1	6	5	32	2	34	6	28	5	12
25	4	0	14	60	14	60	14	60	14	284	14	284	14	284	14	60
26	2	0	1	2	1	2	0	2	0	6	0	6	0	6	1	2
27	2	0	1	2	1	2	0	2	0	6	0	6	0	6	1	2
28	5	0	3	20	3	20	3	20	3	95	3	95	4	75		
29	5	0	4	25	4	25	4	25	4	125	4	125	4	75		
30	3	0	5	15	25	453	7	42	11	135	7	105	9	66		

Vergelijking van methode D, E en F geeft echter aanleiding om wel gebruik te maken van demping en ook van de modifikatiemethode. Daarnaast moet dan worden gezocht naar een methode om al na enkele iteraties te ontdekken dat het probleem moeilijk is om vervolgens speciale aktie te ondernemen. Bij gebruik van E en F kan men al snel ontdekken dat een probleem moeilijk is, omdat in deze situatie de benadering van α_k zeer snel oploopt.

Een voorzichtige voorlopige konklusie:

methode F gekombineerd met een speciale methode voor bijzondere situaties geeft een waardevolle methode voor het oplossen van stelsels niet-lineaire vergelijkingen.

REFERENTIES

BUS, J.C.P. [1976], *A comparative study for solving nonlinear equations,* Mathematisch Centrum, NW 25/76.

DENNIS, J.E. [1971], *On the convergence of Broyden's method for nonlinear systems of equations,* Math. o Comp. 25, p.559-567.

DEUFLHARD, P. [1974a], *A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with application to multiple shooting,* Num. Math. 22.

DEUFLHARD, P. [1974b], *A relaxation strategy for the modified Newton-method,* in: Bulirsch, R., W. Oettli & J. Stoer (eds.), Conference on optimization and optimal control, Oberwolfach, Springer.

FLETCHER, R. [1970], *A new approach to variable metric algorithms,* Comp. J. 13, p.317-322.

GOLDSTEIN, A.A. [1967], *An effective algorithm for minimization,* Num. Math. 10, p.184-189.

ORTEGA, J.M. & W.C. Rheinboldt [1970], *Iterative solution of nonlinear equations in several variables,* Academic Press.

RHEINBOLDT, W.C. [1968], *A unified convergence theory for a class of iterative processes,* Num. An. 5, p.42-63.

Algemene Werkbespreking Afdeling Numerieke Wiskunde

Hoofdstuk 4

Integratiemethoden voor het oplossen van semi-gediscretiseerde partiële differentiaalvergelijkingen

door

K. Dekker

4.1. Inleiding

De laatste jaren valt een groeiende belangstelling voor programmatuur voor het oplossen van tijds-afhankelijke partiële differentiaalvergelijkingen (in het vervolg TDPDE te noemen) waar te nemen. De komst van steeds snellere en grotere rekentuigen is niet vreemd aan dit verschijnsel: de mogelijkheid om de vaak zeer bewerkelijke TDPDE's met algemene numerieke methoden op te lossen, begint nu een serieus alternatief te worden voor eventueel vroeger gebruikte sterk probleem gerichte methoden. Als recente publicaties waarin programmapakketten voor TDPDE's beschreven worden, noemen we SINCOVEC and MADSEN [13], POLAK [11] en SCHRIJER [12].

Ook de afdeling numerieke wiskunde van het Mathematisch Centrum houdt zich bezig met onderzoek naar het creëren van programmatuur voor TDPDE's. Dit verslag stelt zich ten doel een schets te geven van de deelproblemen die hierbij aan de orde kunnen komen, waarbij de onderwerpen die reeds onderzocht worden op het Mathematisch Centrum, uitvoeriger belicht zullen worden.

4.2. Globale structuur van een TDPDE-oplosser

Een van de kernvragen bij de constructie van een TDPDE-oplosser is, op welke wijze de tijd en ruimte discretisatie gerealiseerd moet worden.

page 40
missing

Een TDPDE-oplosser, opgevat als een semi-discretisatiemethode gekoppeld aan een integratiemethode, is meestal slechts geschikt voor één bepaalde klasse. Zo is TEDDY2 [11] ontworpen voor twee-dimensionale parabolische problemen

$$(4.3.1) \qquad u_t = a_1 u_{xx} + a_2 u_{yy} + b_1 u_x + b_2 u_y + cu + e,$$

waarin de coëfficiënten a_1, a_2, b_1, b_2, c en e functies van t,x,y en u mogen zijn, POST [12] voor impliciete, één-dimensionale, zelf geadjungeerde vergelijkingen

$$(4.3.2) \qquad a(t,x,u,u_x,u_t,u_{xt})_x = f(t,x,u,u_x,u_t,u_{xt})$$

en PDEONE [13] voor expliciete, één-dimensionale vergelijkingen

$$(4.3.3) \qquad u_t = f(t,x,u,u_x,(D(t,x,u)u_x)_x).$$

Uiteraard spelen de eigenschappen van de semi-discretisator en de integrator een belangrijke rol bij de bepaling van deze klassen. Een expliciete integrator zal minder geschikt zijn voor impliciete vergelijkingen, en voor 2e orde vergelijkingen zal een speciale integrator voor 2e orde ODE het meest efficiënt zijn. De dimensie van het probleem is van belang voor de semi-discretisatie, zeker indien deze op de eindige elementenmethode gebaseerd is.
Een volledig TDPDE-pakket zal daarom moeten bestaan uit een aantal semi-discretisatiemethoden en een aantal integratiemethoden, die alle gekoppeld kunnen worden en tesamen een voldoend groot aantal klassen bestrijken.

Het pakket dat het Mathematisch Centrum voor ogen staat, zal in eerste instantie plaats bieden aan integratiemethoden voor parabolische, hyperbolische en 2e orde vergelijkingen. Indien de semi-discretisatie d.m.v. de eindige differentiemethode plaatsvindt, dan kunnen in principe alle expliciete TDPDE's met dit pakket numeriek opgelost worden.

4.4. Semi-discretisatie

De semi-discretisatie van de TDPDE kan plaatsvinden met behulp van de eindige differentiemethode, dan wel met de eindige elementenmethode (STRANG [14]). Bovendien kunnen we, afhankelijk van de implementatie, generatieve en interpretatieve methoden onderscheiden. Een generatieve methode (zoals TEDDY2 [11]) zal, uitgaande van een TDPDE, een stelsel ODE's genereren die vervolgens geïntegreerd kunnen worden. De "overhead" bij een dergelijke methode is in het algemeen gering (namelijk onafhankelijk van de tijdsduur van de integratie), de implementatie daarentegen gecompliceerd. Bij de interpretatieve methoden (zoals die in POST [12] en in PDEONE [13]) zal de TDPDE voortdurend als een stelsel ODE's geïnterpreteerd moeten worden. De relatief eenvoudige implementatie heeft daarbij als schaduwzijde een aanzienlijke overhead, die bij enkele experimenten met PDEONE een factor 5 bedroeg.

In de volgende subsecties gaan we nader in op de bovengenoemde semi-discretisatiemethoden.

a. De eindige differentiemethode

Bij de eindige differentiemethode wordt de onbekende functie u niet als continue functie op het definitiegebied D van de TDPDE beschouwd, maar als vector met als elementen de functiewaarden in een eindig aantal punten van D. Door ook de TDPDE slechts in deze punten toe te passen, en alle afgeleiden naar de ruimte-variabelen te vervangen door geschikte differentie-quotiënten, ontstaat een stelsel ODE's.

VOORBEELD. Beschouw de één-dimensionale diffusie-vergelijking

$$(4.4.1) \qquad u_t = u_{xx}, \qquad 0 \le x \le 1, \quad u(0,t) = 1,\ u(1,t) = 0,\ u(x,0) = 1-x.$$

Stellen we $v_j(t) = u(jh,t)$, $j = 0,\ldots,100$, $h = \frac{1}{100}$, en benaderen we $\partial^2 v_j/\partial x^2$ door

$$\frac{\partial^2 v_j}{\partial x^2} = \frac{v_{j+1} - 2v_j + v_{j-1}}{h^2}, \qquad j = 1,\ldots,99,$$

dan gaat (4.4.1) over in het stelsel ODE's

$$(4.4.2) \qquad \frac{dv_0}{dt} = \frac{dv_{100}}{dt} = 0,$$

$$\frac{dv_j}{dt} = \frac{v_{j+1}-2v_j+v_{j-1}}{h^2}, \qquad j = 1,\ldots,99,$$

met de beginwaarden $v_j(0) = 1 - jh$, $j = 0,\ldots,100$.

Een voordeel van de eindige differentiemethode is het feit dat hij gemakkelijk toegepast kan worden, ook bij meerdere ruimte-variabelen. Om deze reden wordt in het testrapport van WOLKENFELT [16] de eindige differentiemethode gebruikt voor de semi-discretisatie.

Nadelen van de eindige differentiemethode zijn: problemen en onnauwkeurigheden bij het representeren van randvoorwaarden m.b.v. eenzijdige differenties en het ontbreken van een algemeen toepasbare theorie omtrent convergentie en stabiliteit.

b. De eindige elementenmethode

Bij de eindige elementenmethode wordt de onbekende functie u benaderd door een functie v uit de eindig-dimensionale deelruimte van functies die te schrijven zijn als som van basisfuncties $B_j(\vec{x})$;

$$(4.4.3) \qquad v(t,\vec{x}) = \sum_{j=1}^{N} v_j(t)B_j(\vec{x}).$$

Schrijven we de TDPDE in operator-vorm

$$(4.4.4) \qquad Lu = 0,$$

dan vormen de vergelijkingen

$$(4.4.5) \qquad (Lv,B_j(\vec{x})) = 0, \qquad j = 1,\ldots,N,$$

een stelsel van N ODE's met als onbekende functies $v_j(t)$. Dit stelsel is in het algemeen impliciet (zoals bij POST [12]), zodat toepassing van deze methode op impliciete TDPDE's voor de hand lijkt te liggen. BAKKER [1] heeft echter aangetoond dat bij een geschikte keuze van basisfuncties en inprodukt een één-dimensionale expliciete TDPDE leidt tot een stelsel expliciete ODE's, zodat ook bij expliciete TDPDE's de eindige elementenmethode goed toepasbaar zou kunnen zijn.

Wordt de eindige elementenmethode voor één-dimensionale vergelijkingen tegenwoordig theoretisch stevig ondersteund, voor meer-dimensionale vergelijkingen is zij nog volop in ontwikkeling. Bruikbare implementaties zijn ons momenteel nog onbekend.

Tenslotte merken we op dat de zelf-geadjungeerde vorm (4.3.2) voordelen biedt voor de eindige elementenmethode, omdat door partiële integratie de tweede afgeleide naar x geëlimineerd kan worden, wat bij (4.3.3) niet het geval zou zijn.

4.5. Integratiemethoden

Beschouw een stelsel ODE's geschreven in de vorm

(4.5.1) $\quad y' = f(t,y), \qquad y(0) = y_0.$

Voor de integratie van dit stelsel gebruiken de twee reeds genoemde pakketten POST [12] en [13] impliciete methoden, en wel een éénstaps-differentiemethode gevolgd door extrapolatie (zie BULIRSCH [2]), en een lineaire meerstapsmethode (GEAR [4]).

Het onderzoek op het Mathematisch Centrum concentreert zich tot dusverre vrijwel uitsluitend op de constructie van efficiënte expliciete integratiemethoden. Impliciete methoden komen gezien de grootte van de stelsels immers niet in aanmerking, tenzij zij voorzien zijn van een efficiënte lineaire stelsel oplosser. Bij één-dimensionale vergelijkingen zijn dergelijke oplossers beschikbaar, het stelsel heeft dan namelijk een bandstructuur; voor meer-dimensionale TDPDE's, die leiden tot blok-band structuren, zijn ons echter nog geen efficiënte methoden bekend.

Van de expliciete methoden zijn de Runge-Kutta methoden de belangrijkste vertegenwoordigers. Er bestaan Runge-Kutta methoden van hoge orde, die geschikt zijn om een zeer grote nauwkeurigheid te bereiken (bijv. ZONNEVELD [17]). Bij het integreren van grote stelsels ODE's afkomstig van een TDPDE zal men in het algemeen met slechts enkele cijfers tevreden zijn, zodat men een grote tijdstap τ zou willen toelaten. Expliciete methoden worden echter instabiel indien de tijdstap groter wordt dan een bepaalde probleem en methode-afhankelijke grootheid τ_{stab}. Door deze stabiliteitsvoorwaarde wordt de efficiëntie van de methode dus negatief beïnvloed. Methoden waarbij de maximale stabiele tijdstap τ_{stab} is geoptimaliseerd, worden besproken in de volgende subsecties.

a. Methoden voor parabolische ODE's.

Als voorbeeld van een parabolische ODE kunnen we vergelijking (4.4.2) beschouwen. De Jacobiaan van dit stelsel is een tridiagonale matrix, die reële eigenwaarden liggend in het interval $[-4/h^2,0]$ bezit. Omdat de componenten behorende bij de grote negatieve eigenwaarden analytisch gezien snel uitgedempt worden, zijn we geïnteresseerd in methoden die de componenten bij de kleine eigenwaarden nauwkeurig, die bij de grote slechts stabiel representeren.

Tot methoden met gunstige stabiliteitseigenschappen met betrekking tot negatieve eigenwaarden behoren de adaptieve Runge-Kutta (ARK) methoden [6]. Laat y_n de numerieke oplossing op het tijdstip t_n voorstellen. Dan worden deze methoden gerepresenteerd door het schema

$$
\begin{aligned}
y_{n+1}^{(0)} &= y_n, \\
y_{n+1}^{(j)} &= y_n + \lambda_j h_n f(t_n+\lambda_{j-1}h_n, y_{n+1}^{(j-1)}), \qquad j = 1,\ldots,m-1, \\
y_{n+1} &= y_n + h_n f(t_n+h_n, y_{n+1}^{(m-1)}),
\end{aligned}
\tag{4.5.2}
$$

waarbij

$$h_n = t_{n+1} - t_n.$$

De variatie-vergelijking van dit schema heeft de vorm

$$y_{n+1} = R(h_n J_n) y_n,$$

waarbij $R(z) = \sum_{j=0}^{m} \beta_j z^j$ het zogenaamde stabiliteitspolynoom is, met coëfficiënten die afhangen van de parameters λ_j, $j = 1,\ldots,m$. Dit stabiliteitspolynoom en de in modulus grootste eigenwaarde σ van de Jacobiaan J_n bepalen de maximaal toegelaten tijdstap τ_{stab}. In tabel 4.5.1 wordt voor een aantal ARK formules een overzicht gegeven van deze maximale tijdstap.

Onlangs is door VERWER [15] een klasse van driestaps Runge-Kutta formules beschouwd (in het vervolg M3RK genoemd). Deze formules kunnen als volgt gerepresenteerd worden

$$y_{n+1}^{(0)} = y_n,$$

$$(4.5.3)\qquad y_{n+1}^{(j)} = (1-b_j)y_n + b_j y_{n-1} + c_j h_n f(t_{n-1}, y_{n-1}) + \lambda_j h_n f(t_n + \mu_{j-1} h_n, y_{n+1}^{(j-1)}),$$

$$j = 1,\ldots,m,$$

$$y_{n+1} = dy_{n+1}^{(m)} + (1-d)y_{n-2}, \qquad 2 \le m \le 12,$$

waarbij y_n en h_n dezelfde betekenis hebben als in schema (4.5.2). De stabiliteitseigenschappen van dit schema worden bepaald door de lineaire recursie-relatie

$$(4.5.4)\qquad y_{n+1} = dS(z)y_n + dP(z)y_{n-1} + (1-d)y_{n-2},$$

welke verkregen wordt door schema (4.5.3) toe te passen op de modelvergelijking $y' = \delta y$, en $z = h_n\delta$ te stellen. Voor bepaalde keuzen van de polynomen $S(z) = \sum_{j=0}^{m} s_j z^j$, en, $P(z) = \sum_{j=0}^{m} p_j z^j$ en van d zijn de stabiliteitseigenschappen optimaal. In tabel 4.5.1 geven we een overzicht van de maximaal toegelaten staplengte τ_{stab} voor deze methoden.

Tabel 4.5.1. Maximaal toegelaten tijdstap τ_{stab} voor integratie van parabolische vergelijkingen.

Methode	orde	m	$\sigma\cdot\tau_{stab}$	$\sigma\cdot\tau_{stab}/m$
RK1 n[17]	5	6	3.2	.53
Klassieke Runge-Kutta	4	4	2.8	.70
ARK	1	2	7.7	3.85
ARK	1	k	$\sim 1.9k^2$	1.9k
ARK	2	3	6.14	2.05
ARK	2	k	$\sim .75k^2$	.75k
M3RK	1	k	$5.15k^2$	5.15k
M3RK	2	k	$2.29k^2$	2.29k

b. Methoden voor hyperbolische ODE's.

Hyperbolische ODE's kunnen o.a. ontstaan door op de golfvergelijking

$$(4.5.5)\qquad u_t = v_x, \quad u(0,t) = u(1,t) = v(0,t) = 0,\ v(1,t) = g(t)$$
$$v_t = u_x, \quad u(x,0) = u_0(x),\ v(x,0) = v_0(x),$$

semi-discretisatie m.b.v. de eindige differentiemethode toe te passen. Stellen we $y_j(t) = u(jh,t)$, $j = 0,\ldots,100$, en $y_j(t) = v((j-100)h,t)$, $j = 101,\ldots,200$, dan krijgen we het stelsel

$$
\begin{aligned}
& y_0' = y_{100}' = 0, \\
& y_j' = \frac{y_{j+101} - y_{j+99}}{2h}, \qquad j = 1,\ldots,99 \\
& y_j' = \frac{y_{j-99} - y_{j-101}}{2h}, \qquad j = 101,\ldots,200 \\
& y_{200}' = g'(t).
\end{aligned}
\tag{4.5.6}
$$

De Jacobiaan van dit stelsel is anti-symmetrisch, en de eigenwaarden liggen in het interval $[-i/h,+i/h]$. Hoewel iedere component behorende bij een imaginaire eigenwaarde een niet gedempte trilling uitvoert, en dus altijd in de analytische oplossing aanwezig zal blijven, zijn er veel problemen waarbij we slechts in de lage frequenties geïnteresseerd zijn, zoals bijvoorbeeld bij de ondiep-water-vergelijkingen (VAN DER HOUWEN [7]). Dientengevolge zijn we ook bij hyperbolische vergelijkingen geïnteresseerd in methoden die een grote tijdstap τ_{stab} toelaten.

In VAN DER HOUWEN [6] is aangetoond dat de ARK formules gegeven door schema (4.5.3) optimale stabiliteitseigenschappen langs de imaginaire as bezitten, indien het stabiliteitspolynoom R(z) voldoet aan

$$
R(z) = T_k\left(1+\frac{z^2}{2k^2}\right) + \frac{z}{k}\left(1+\frac{z^2}{4k^2}\right) U_{k-1}\left(1+\frac{z^2}{2k^2}\right), \quad m = 2k+1,
$$

waarbij T_k en U_k Chebyshev-polynomen van de eerste en tweede soort voorstellen. Een overzicht van de bijbehorende τ_{stab} wordt gegeven in tabel 4.5.2.

Recentelijk is door DEKKER [3] een klasse van gegeneraliseerde Runge-Kutta methoden (ook wel aangeduid met CRK) onderzocht. Hierbij worden in het autonoom geschreven Runge-Kutta schema

$$
\begin{aligned}
y_{n+1}^{(0)} &= y_n, \\
y_{n+1}^{(j)} &= y_n + \sum_{l=0}^{j-1} h_n \Lambda_{jl} f(y_{n+1}^{(l)}), \qquad j = 1,\ldots,m, \\
y_{n+1} &= y_{n+1}^{(m)},
\end{aligned}
\tag{4.5.7}
$$

de parameters Λ_{jl}, $j = 1,\ldots,m$, $l = 1,\ldots,j-1$, verondersteld matrices te zijn en wel matrices van de vorm

$$
\Lambda_{jl} = \begin{pmatrix} \mu_{jl} I & 0 \\ 0 & \beta_{jl} I \end{pmatrix}.
$$

Indien schema (4.5.7) wordt toegepast op een stelsel lineaire ODE's als (4.5.6) dan ontstaat de recursie-relatie

$$
y_{n+1} = R_m(h_n J_n) y_n,
$$

waarbij $R_m(hJ_n)$ geschreven kan worden als

$$
R_m(hJ_n) = T \begin{pmatrix} P_m(h_n \Lambda) & \overline{Q_m(h_n \Lambda)} \\ Q_m(h_n \Lambda) & \overline{P_m(h_n \Lambda)} \end{pmatrix} T^{-1},
$$

T is hierin een transformatiematrix, Λ een matrix die eigenwaarden van J_n met positief imaginair deel bevat, en P en Q zijn polynomen van de graad m met coëfficiënten die van μ_{jl} en β_{jl} afhangen.

Indien m oneven gekozen wordt, en P_m en Q_m als volgt

$$
P_m(z) = \frac{1}{2}\{(1+z+\tfrac{1}{2}z^2) T_k(1 + \frac{z^2}{2k^2}) - 1 - \frac{1}{8} z^4\}
$$

en

$$
Q_m(z) = P_m(z) - 1 - z - \tfrac{1}{2}z^2,
$$

dan heeft schema (4.5.7) gunstige stabiliteitseigenschappen langs de imaginaire as. In tabel 4.5.2 geven we weer de bijbehorende τ_{stab}. Omdat de bewerkelijkheid van deze formules vaak kleiner is dan m suggereert, geven we tevens de effectieve tijdstap τ_{eff}, dat is het quotiënt van τ_{stab} en het aantal benodigde rechterlid evaluaties.

Tabel 4.5.2. Maximaal toegelaten tijdstap τ_{stab} voor integratie van hyperbolische vergelijkingen.

Methode	orde	m	$\sigma\cdot\tau_{stab}$	$\sigma\cdot\tau_{stab}/m$	$\sigma\cdot\tau_{eff}$
Klassieke Runge-Kutta	4	4	$2\sqrt{2}$	.71	.71
ARK	2	3	2	.67	.67
ARK	2	k	k-1	(k-1)/k	(k-1)/k
CRK	1	3	2	.67	1.33
CRK	1	2k+1	2k	2k/(2k+1)	4k/(2k+1)
CRK	2	3	2	.67	1.00
		5	4	.80	1.33
		7	6	.86	1.67

Tenslotte merken we op dat, bovengenoemde formules zwak stabiel zijn, d.w.z. de norm van de amplificatie functies $R(h_n\lambda)$ is voor alle eigenwaarden λ gelijk aan 1. Er zijn echter ook sterk stabiele formules ontwikkeld die de eigenschap hebben dat de hoge frequenties uitgedempt worden. Overigens gaat deze sterke stabiliteit wel ten koste van de maximale tijdstap. Voor nadere bijzonderheden verwijzen we naar [16] en [3].

c. Methoden voor tweede orde ODE's.

Hyperbolische TDPDE's leiden vaak tot tweede orde ODE's. Semi-discretisatie van de vergelijking

$$u_{tt} = u_{xx}, \tag{4.5.8}$$

leidt bijvoorbeeld tot het stelsel

$$\begin{aligned} y_0'' &= 0, \\ y_j'' &= \frac{y_{j+1}-2y_j+y_{j-1}}{h^2}, \qquad j = 1,\ldots,99, \\ y_{100}'' &= 0. \end{aligned} \tag{4.5.9}$$

Uiteraard kan dit stelsel tot de eerste orde vorm (4.5.6) teruggebracht worden. Indien de tweede orde vergelijkingen geen eerste afgeleiden bevatten, dan zijn Nyström-Runge-Kutta methoden vaak efficiënter [8].

Deze methoden kunnen beschreven worden door het schema

$$y_{n+1}^{(1)} = y_n + \mu_1 h_n y_n' ,$$

$$(4.5.10) \quad y_{n+1}^{(j)} = y_n + \mu_j h_n y_n' + \lambda_j h_n^2 f(t_n + \mu_{j-1} h_n, y_{n+1}^{(j-1)}), \qquad j = 2,\ldots,m+1,$$

$$y_{n+1} = y_{n+1}^{(m+1)}, \quad y_{n+1}' = y_n' + h_n f(x_n + \tfrac{1}{2} h_n, y_{n+1}^{(m)}) .$$

Hierin stelt f het rechterlid van de tweede orde vergelijking voor, en y_n en y_n' de numerieke benadering van de oplossing en zijn afgeleide op het tijdstip t_n.

Door VAN DER HOUWEN [9] worden zwak en sterk stabiele formules (SNRK genoemd) gegeven met optimale stabiliteitseigenschappen voor negatief reële eigenwaarden van de Jacobiaan van f. De maximaal stabiele tijdstap van deze formules vermelden we in tabel 4.5.3.

Tabel 4.5.3. Maximaal stabiele tijdstap voor integratie van 2e orde hyperbolische vergelijkingen.

Methode	orde	m	$\sqrt{\sigma \cdot \tau}_{stab}$	$\sqrt{\sigma \cdot \tau}_{stab}/(m-1)$
SNRK (zwak stabiel)	1	2	2	2
SNRK (zwak stabiel)	2	3	4	2

LITERATUUR

[1] BAKKER, M., [1977], *On the numerical solution of parabolic equations in a single space variable by the continuous time Galerkin method*, Report NW 37/77, Mathematisch Centrum, Amsterdam.

[2] BULIRSCH, R. & J. STOER, [1966], *Numerical treatment of ordinary differential equations by extrapolation methods*, Numer. Math. 8, 1-13.

[3] DEKKER, K., [1977], *Generalized Runge-Kutta methods for a class of hyperbolic differential equations*, Report NW 41/77, Mathematisch Centrum, Amsterdam.

[4] GEAR, C.W., [1971], *The automatic integration of ordinary differential equations,* Comm. ACM 14, 176-179.

[5] HENRICI, P., [1962], *Discrete variable methods in ordinary differential equations,* John Wiley & Sons, New York.

[6] HOUWEN VAN DER, P.J., [1977], *Construction of integration formulas for initial value problems,* North-Holland Publ. Comp., Amsterdam.

[7] HOUWEN VAN DER, P.J., [1975], *Two level difference schemes with varying mesh sizes for the shallow water equations,* Report NW 22/75, Mathematisch Centrum, Amsterdam.

[8] HOUWEN VAN DER, P.J.,[1975], *Stabilized Runge-Kutta methods for second order differential equations without first derivatives,* Report NW 26/75, Mathematisch Centrum, Amsterdam.

[9] HOUWEN VAN DER, P.J., [1977], *Runge-Kutta methods for the integration of hyperbolic differential equations,* Report NW 40/77, Mathematisch Centrum, Amsterdam.

[10] LAMBERT, J.D., [1973], *Computational methods in ordinary differential equations,* John Wiley & Sons, London.

[11] POLAK, S.J. & J. SCHROOTEN, [1975], *Preliminary TEDDY2 user manual,* Philips-ISA-UDV-DSA-SCA/SP/75/024/mw, Eindhoven.

[12] SCHRIJER, N.L., [1975], *Numerical Solution of Time-Varying Partial Differentail Equations in One Space Variable,* Comp. Science Techn. Rep. No. 53, Bell Laboratories, Murray Hill.

[13] SINCOVEC, R.F. & N.K. MADSEN, [1975], *Software for nonlinear partial differential equations,* ACM Transactions on Mathematical Software 1, 232-260.

[14] STRANG, G. & G. FIX, [1973], *An analysis of the finite element method,* Prentice Hall, New York.

[15] VERWER, J., [1977], *A class of stabilized three-step Runge-Kutta methods for the numerical integration of parabolic equations,* Report NW 39/77, Mathematisch Centrum, Amsterdam.

[16] WOLKENFELT, P.H.M. et al., [1977], *Comparing stabilized Runge-Kutta methods for semi-discretized parabolic and hyperbolic equation.* Report NW 45/77, Mathematisch Centrum, Amsterdam.

[17] ZONNEVELD, J.A., [1964], *Automatic numerical integration*, MC Tract 8, Mathematisch Centrum, Amsterdam.