STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
REKENAFDELING

Cursus Wetenschappelijk Rekenen A

Numerieke Wiskunde

deel III

door

T.J. Dekker

1967

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

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) and the Central Organization for Applied Scientific Research in the Netherlands (T.N.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

Hoofdstuk 8. Approximaties

1. Chebyshev interpolatie

1.1 Inleiding

In hoofdstuk 2 hebben we gezien, dat voor een gegeven functie f er precies één veelterm f_n^* van de graad kleiner dan n bestaat, die in n gegeven basispunten met f overeenstemt.
Deze veelterm f_n^* kan o.a. worden geschreven in de Lagrange-vorm (formule (6.1) pag. 40) of in de Newton-vorm (formule (10.5) met (10.3) pag. 51).
De rest-term $R_n(x)$ is gedefiniëerd door

1.1.1 $$f(x) = f_n^*(x) + R_n(x).$$

Als $[a,b]$ een interval is, dat de gegeven basispunten x_i, i = 0(1)n-1, en het punt x bevat en als f minstens n maal differentiëerbaar is in dit interval, dan geldt voor deze restterm (zie pag. 55 stelling 11.3):

1.1.2 $$R_n(x) = \frac{f^{(n)}(\xi)}{n!}\,\pi_n(x),$$

waarbij ξ een getal tussen a en b is en

$$\pi_n(x) = (x - x_0)(x - x_1) \ldots (x - x_{n-1}).$$

Laten nu a en b gegeven getallen zijn waarbij a < b. Wij zoeken nu een interpolatie-formule, die op het hele interval $[a,b]$ een goede benadering van f geeft. Als het interval klein is, zal $f^{(n)}(x)$ niet veel veranderen. Dan mogen we dus een goede benadering verwachten, indien de basispunten zo gekozen worden, dat $\max_{a\leq x\leq b} |\pi_n(x)|$ zo klein mogelijk wordt. Deze voorwaarde leidt tot de

1.2 Polynomen van Chebyshev

De polynomen van Chebyshev worden gedefiniëerd als volgt:

1.2.1 $$T_n(x) = \cos\,(n \arccos\,(x)).$$

Met andere woorden: $T_n(x) = \cos(n\theta)$, als $x = \cos(\theta)$. Blijkbaar geldt:

1.2.2 $$T_0(x) \equiv 1 \quad , \quad T_1(x) \equiv x.$$

Bovendien geldt de recursie-formule

1.2.3 $$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

Deze formule volgt onmiddellijk uit de goniometrische relatie

$$\cos(n+1)\theta + \cos(n-1)\theta = 2 \cos\theta \cos n\theta.$$

Met behulp van 1.2.2 en herhaald toepassen van 1.2.3 krijgen we

$$T_2(x) = 2x^2 - 1$$

$$T_3(x) = 4x^3 - 3x,$$

$$T_4(x) = 8x^4 - 8x^2 + 1.$$

Enzovoorts. Blijkbaar geldt:
$T_n(x)$ is een polynoom van de graad n. Als n even is komen er alleen de even machten van x in voor en als n oneven is alleen de oneven machten. De coëfficiënt van x^n in $T_n(x)$ is blijkbaar 2^{n-1} voor $n \geq 1$ (maar daarentegen 1 voor n = 0).
M.a.w. de polynomen

1.2.4 $$T_0(x) \ , \ \frac{1}{2^{n-1}} T_n(x) \ , \ n \geq 1,$$

hebben 1 als coëfficiënt van de hoogste x-macht.
De nulpunten van $T_n(x)$, $n \geq 1$, zijn blijkbaar

1.2.5 $$x_k = \cos\left(k + \tfrac{1}{2}\right)\frac{\pi}{n} \ , \quad k = 0(1)n-1.$$

(Voor hogere waarden van k krijgen we dezelfde waarden weer terug, bv. $x_n = x_{n-1}$, $x_{n+1} = x_{n-2}$, enz.).
Tussen de nulpunten en bovendien aan de uiteinden van het interval $[-1,+1]$ heeft $T_n(x)$ extreme waarden (maxima of minima), die afwisselend gelijk zijn aan +1 of -1. De extreme waarden worden bereikt in de punten

1.2.6 $$z_k = \cos \frac{k\pi}{n}, \quad k = 0(1)n.$$

Voor deze punten geldt immers

$$T_n(z_k) = \cos(k\pi) = (-1)^k.$$

Nu zijn we gereed om te bewijzen de

Eerste stelling van Chebyshev

Zij $P_n(x)$ een veelterm van de graad n met 1 als coëfficiënt van de hoogste x-macht. Dan bereikt

$$m(P_n) = \max_{-1 \leq x \leq 1} |P_n(x)|$$

zijn minimale waarde voor $P_n(x) = \frac{1}{2^{n-1}} T_n(x)$ en de minimale waarde van $m(P_n)$ is $\frac{1}{2^{n-1}}$.

Bewijs. Stel er is een veelterm $P_n(x)$ zodanig, dat $m(P_n) < \frac{1}{2^{n-1}}$. Dan is $P_n(x) - \frac{1}{2^{n-1}} T_n(x)$ een veelterm van de graad kleiner dan n, omdat de x^n-termen tegen elkaar wegvallen. Deze veelterm is in de punten z_k (zie 1.2.6) afwisselend positief en negatief, dus tussen deze n+1 punten z_k liggen minstens n nulpunten. Dit kan alleen als deze veelterm identiek nul is, m.a.w. als $P_n(x) = \frac{1}{2^{n-1}} T_n(x)$.
Dan is $m(P_n)$ gelijk aan $\frac{1}{2^{n-1}}$, in tegenspraak met de veronderstelling.
Hiermee is de stelling bewezen.

Opgaven (betreffende rekenen in meervoudige lengte)

136) Zij gegeven

ln (2) = 0.6931 47180 55994 53094 172321
ln (3) = 1.0986 12288 66810 96913 952452

Gevraagd: ln (6) in 25 decimalen.

137) Zij gegeven

π = 3.1415 92653 58979 32385

Gevraagd in 19 cijfers: π^2, π^3, π^{-1}.

138) Zij gegeven ln (2) (zie opgave 136) en

ln (10) = 2.3025 85092 99404 56840

Gevraagd $^{10}\log 2$ in 19 decimalen.

139) Bereken in 19 decimalen $\sqrt{5}$.

Voor de antwoorden, zie pag. 2 en 3 van:
Handbook of Mathematical functions, edited by M. Abramowitz and I.A. Stegun, Nat. Bur. of Standards, AMS 55.

1.3 De interpolatie-formule van Chebyshev

Zij gegeven een functie g(y) op een interval [a,b]. We zoeken nu een interpolatie-formule, die bij (nagenoeg) constante n-de afgeleide van g de maximale absolute waarde van de restterm op het interval [a,b] zo klein mogelijk maakt (zie inleiding 1.1).
Hiertoe gaan we eerst het interval [a,b] door een lineaire transformatie van y overvoeren in het standaard-interval [-1,+1]. De lineaire transformatie luidt:

1.3.1 $$y = \frac{b-a}{2} x + \frac{b+a}{2}.$$

De functie

1.3.2 $$f(x) = g\left(\frac{b-a}{2} x + \frac{b+a}{2} \right)$$

moet dan worden benaderd in het interval [-1,+1].
Volgens de eerste stelling van Chebyshev krijgen we, afgezien van de n-de afgeleide van f,het beste resultaat als we de basispunten zó kiezen, dat

$$\pi_n(x) = (x-x_0) \ldots (x-x_{n-1}) = \frac{1}{2^{n-1}} T_n(x).$$

M.a.w. we moeten de basispunten kiezen volgens (1.2.5). Dan hebben we dus (vgl. (1.1.1) en (1.1.2)):

1.3.3 $$f(x) = f_n^*(x) + \frac{f^{(n)}(\xi)}{n!\, 2^{n-1}} T_n(x).$$

De veelterm $f_n^*(x)$ van graad kleiner dan n kan op diverse manieren geschreven worden (zie hoofdstuk 2). Thans schrijven wij haar als een lineaire combinatie van Chebyshev-polynomen en krijgen dan de

interpolatie-formule van Chebyshev:

1.3.4 $$f(x) = \sum_{k=0}^{n-1} c_k^{(n)} T_k(x) + \frac{f^{(n)}(\xi)}{n!\, 2^{n-1}} T_n(x).$$

Voor $-1 \leq x \leq 1$ is deze formule geldig, mits de n-de afgeleide van f bestaat op het interval [-1,+1].

De coëfficiënten $c_k^{(n)}$ worden verkregen door op te lossen het lineaire stelsel:

$$1.3.5 \qquad \sum_{k=0}^{n-1} c_k^{(n)} T_k(x_i) = f(x_i) \ , \quad i = 0(1)n-1,$$

waarbij x_i de nulpunten van $T_n(x)$ zijn (zie 1.2.5).
Dit volgt meteen uit (1.3.4), omdat in de basispunten x_i, $i = 0(1)n-1$, de restterm gelijk aan 0 is. (Vgl. de overeenkomstige lineaire stelsels in hoofdstuk 2, nl. (4.1) pag. 29, (5.1) pag. 35, en (10.2) pag. 49.)
Het stelsel (1.3.5) kan gemakkelijk worden opgelost, door gebruik te maken van de orthogonaliteit der Chebyshev-polynomen. Deze polynomen voldoen namelijk aan de relatie:

$$1.3.6 \qquad \sum_{r=0}^{n-1} T_j(x_r)T_k(x_r) = \underline{\text{if}}\ j \neq k\ \underline{\text{then}}\ 0\ \underline{\text{else if}}\ j \neq 0\ \underline{\text{then}}\ n/2\ \underline{\text{else}}\ n,$$

waarbij x_r, $r = 0(1)n-1$, de nulpunten van $T_n(x)$ zijn en $j,k = 0(1)n-1$.

Bewijs. Volgens (1.2.5) zijn de nulpunten van $T_n(x)$

$$x_r = \cos(r+\tfrac{1}{2})\frac{\pi}{n} = \cos\theta_r \quad \text{als} \quad \theta_r = (r+\tfrac{1}{2})\frac{\pi}{n}\ .$$

Dus volgens definitie (1.2.1) en een bekende goniometrische formule:

$$\sum_{r=0}^{n-1} T_j(x_r)\,T_k(x_r) = \sum_{r=0}^{n-1} \cos j\theta_r \cos k\theta_r =$$

$$\tfrac{1}{2}\sum_{r=0}^{n-1} \cos(j+k)\theta_r + \tfrac{1}{2}\sum_{r=0}^{n-1} \cos(j-k)\theta_r.$$

Deze sommen kunnen we herleiden tot meetkundige reeksen, als we bedenken, dat cosinus en sinus voldoen aan

$$\cos\alpha = \mathrm{Re}(e^{i\alpha}) \quad , \quad \sin\alpha = \mathrm{Im}(e^{i\alpha}).$$

Voor de eerste som krijgen we, als $j + k \neq 0$:

$$\sum_{r=0}^{n-1} \cos(j+k)\theta_r = \sum_{r=0}^{n-1} \mathrm{Re}(e^{i(j+k)\theta_r}) = \mathrm{Re}\Big(\sum_{r=0}^{n-1} e^{i(j+k)\theta_r}\Big)$$

$$= \mathrm{Re}\Big(\frac{e^{i(j+k)(\pi+\theta_0)} - e^{i(j+k)\theta_0}}{e^{2i(j+k)\theta_0} - 1}\Big) = \mathrm{Re}\Big(\frac{(-1)^{j+k} - 1}{e^{i(j+k)\theta_0} - e^{-i(j+k)\theta_0}}\Big)$$

$$= \mathrm{Re}\Big(\frac{(-1)^{j+k} - 1}{2i\sin(j+k)\theta_0}\Big) = 0.$$

De reden van de meetkundige reeks is $e^{i(j+k)\frac{\pi}{n}} = e^{2i(j+k)\theta_0}$.
Deze is ongelijk aan 1 en dus de noemer ongelijk aan 0, omdat $j+k < 2n$.
We krijgen zo het reële deel van een zuiver imaginair getal, dat dus 0 is. Als $j+k = 0$, heeft de eerste som blijkbaar de waarde n. Dus

$$\sum_{r=0}^{n-1} \cos(j+k)\theta_r = \underline{\text{if}}\ j+k > 0\ \underline{\text{then}}\ 0\ \underline{\text{else}}\ n.$$

Op volkomen analoge wijze krijgen we voor de tweede som:

$$\sum_{r=0}^{n-1} \cos(j-k)\theta_r = \underline{\text{if}}\ j \neq k\ \underline{\text{then}}\ 0\ \underline{\text{else}}\ n.$$

Hieruit volgt onmiddellijk formule (1.3.6).
Met behulp van deze formule kunnen we expliciete formules voor de coëfficiënten $c_j^{(n)}$ vinden. Hiertoe vermenigvuldigen we (1.3.5) met $T_j(x_i)$ en sommeren over de basispunten x_i. Dan hebben we voor $j = 0(1)n-1$:

1.3.7 $$\sum_{i=0}^{n-1} f(x_i)T_j(x_i) = (\underline{\text{if}}\ j = 0\ \underline{\text{then}}\ n\ \underline{\text{else}}\ n/2) \times c_j^{(n)}.$$

Hieruit volgt:

1.3.8 $$\begin{cases} c_0^{(n)} = \frac{1}{n}\sum_{i=0}^{n-1} f(x_i), \\ c_j^{(n)} = \frac{2}{n}\sum_{i=0}^{n-1} f(x_i)T_j(x_i),\ i = 1(1)n-1, \end{cases}$$

waarbij $x_i = \cos(i + \frac{1}{2})\frac{\pi}{n}$.

Voorbeeld

Gevraagd de 3-punts Chebyshev-interpolatie-formule op het interval $[-1,+1]$ voor de functie $\cos \frac{\pi}{4} x$ en de absoluut maximale fout op dit interval.

De fout is in absolute waarde kleiner dan

$$(\tfrac{\pi}{4})^3 \sin(\tfrac{\pi}{4}) \frac{1}{3!\, 2^2} \approx .014.$$

De basispunten zijn:

$$x_0 = \cos(\tfrac{\pi}{6}) \approx .8660, \quad x_1 = 0 , \quad x_2 \approx -.8660.$$

Met formule (1.3.8) vinden we

$$c_0^{(3)} = \frac{1}{3} (\cos \tfrac{\pi}{4} x_0 + \cos(0) + \cos \tfrac{\pi}{4} x_2) \approx \frac{1}{3} (1 + 2 \cos(.7854 \times .8660))$$

$$\approx .8516$$

$$c_1^{(3)} = 0$$

$$c_2^{(3)} = \frac{2}{3} (\frac{1}{2} \cos \tfrac{\pi}{4} x_0 - \cos(0) + \frac{1}{2} \cos \tfrac{\pi}{4} x_2)$$

$$\approx \frac{2}{3} (\cos(.6802) - 1) \approx -.1484.$$

Opgaven

140) a) Schrijf een Algol-procedure, die de coëfficiënten van $T_n(x)$ uitrekent.

b) Schrijf een Algol-procedure, die in een gegeven interval $[a,b]$ voor een gegeven functie f de coëfficiënten $c_k^{(n)}$, k = 0(1)n-1, van de n-punts Chebyshev interpolatie-formule berekent.

c) Schrijf een programma, dat volgens de Chebyshev interpolatie-formule $f_n^*(x)$ en de fout $f_n^*(x) - f(x)$ berekent voor $x = a(\frac{b-a}{5n})b$, als gegeven zijn:
$f(x) = 2^x$, a = 0 , b = 1 , n = 4(1)8.
Laat het programma printen de waarden x, $f_n^*(x)$, $f_n^*(x) - f(x)$ en de absoluut maximale fout.

141) Beschouw de n-punts Chebyshev-interpolatie op het interval $[-1,+1]$ van de volgende functies:

$\cos \frac{\pi}{2} x$, $\cos \frac{\pi}{4} x$, e^x , $2^{0.5(x+1)}$.

Hoeveel moet n minstens zijn, om op het interval $[-1,+1]$ een precisie te bereiken van 5, 10 of 12 decimalen?
Vergelijk hiermee het aantal voor dezelfde precisie benodigde termen van de Taylor-reeks.

142) Bereken in 5 decimalen de coëfficiënten van de 4-punts Chebyshev interpolatie-formule op het interval $[-1,+1]$ voor de functies

$$\cos \frac{\pi}{4} x \quad \text{en} \quad 2^{0.5(x+1)}.$$

1.4 De Chebyshev-reeks

Nemen we in de n-de orde interpolatie-formule van Chebyshev (1.3.4) de limiet voor $n \to \infty$, dan krijgen we, als f oneindig vaak differentiëerbaar is en zijn afgeleiden niet te snel toenemen:

1.4.1 $$f(x) = \sum_{k=0}^{\infty} c_k^{(\infty)} T_k(x).$$

Deze reeks van Chebyshev convergeert voor een dergelijke functie f op het interval $[-1,+1]$ veel sneller naar f(x), dan de Taylor-reeks van f. Schrijven we $x = \cos\theta$ en $F(\theta) = f(\cos\theta)$, dan ontstaat de Fourier-cosinus-reeks

1.4.2 $$F(\theta) = \sum_{k=0}^{\infty} c_k^{(\infty)} \cos k\theta.$$

Deze reeks is speciaal bestemd voor periodieke even functies met periode 2π.
Om de coëfficiënten $c_k^{(\infty)}$ te vinden hebben we een orthogonaliteits-relatie nodig, die uit (1.3.6) ontstaat door n naar oneindig te laten naderen. Doet men dit zorgvuldig, dan krijgt men voor $j,k \geq 0$:

1.4.3 $$\int_{-1}^{+1} \frac{T_j(x)T_k(x)}{\sqrt{1-x^2}}\,dx = \int_0^{\pi} \cos j\theta \cos k\theta\, d\theta = \underline{\text{if}}\ j \neq k\ \underline{\text{then}}\ 0\ \underline{\text{else}}$$

$\underline{\text{if}}\ j \neq 0\ \underline{\text{then}}\ \pi/2\ \underline{\text{else}}\ \pi.$

Deze formule is echter veel gemakkelijker direct te bewijzen als volgt: Uit de eerste integraal volgt direct de tweede door de substitutie $x = \cos\theta$. Voor de tweede integraal vinden we, mits $j \neq k$:

$$\int_0^{\pi} \cos j\theta \cos k\theta\, d\theta = \frac{1}{2}\int_0^{\pi} \{\cos(j+k)\theta + \cos(j-k)\theta\}d\theta =$$

$$\frac{1}{2(j+k)} \sin(j+k)\theta \Big|_0^{\pi} + \frac{1}{2(j-k)} \sin(j-k)\theta \Big|_0^{\pi} = 0.$$

Als $j = k \neq 0$ krijgen we blijkbaar de waarde $\pi/2$ en als $j = k = 0$ de waarde π, waarmee de formule bewezen is.

Met behulp van de orthogonaliteits-relatie (1.4.3) kunnen we een expliciete formule voor de coëfficiënten $c_j^{(\infty)}$ verkrijgen. Immers uit (1.4.1) en (1.4.3) volgt:

1.4.4 $$\int_{-1}^{+1} \frac{f(x)T_j(x)}{\sqrt{1-x^2}}\,dx = (\underline{\text{if}}\ j \neq 0\ \underline{\text{then}}\ \pi/2\ \underline{\text{else}}\ \pi) \times c_j^{(\infty)}.$$

Dus

1.4.5 $$\begin{cases} c_0^{(\infty)} = \frac{1}{\pi}\int_{-1}^{+1} \frac{f(x)}{\sqrt{1-x^2}}\,dx \\ c_j^{(\infty)} = \frac{2}{\pi}\int_{-1}^{+1} \frac{f(x)T_j(x)}{\sqrt{1-x^2}}\,dx \end{cases}$$

1.5 Economiseren van machtreeksen

De integralen voorkomend in (1.4.5) zijn vaak lastig te berekenen, tenminste als f(x) geen polynoom is. In de praktijk werken we daarom meestal met een polynoom-benadering (bijvoorbeeld een stuk Taylor-reeks van f):

1.5.1 $$f(x) \approx P_m(x) = \sum_{k=0}^{m} a_k\, x^k.$$

Het polynoom $P_m(x)$ kan namelijk gemakkelijk worden omgezet in zijn Chebyshev-reeks

1.5.2 $$P_m(x) = \sum_{k=0}^{m} b_k\, T_k(x).$$

De coëfficiënten b_k nemen meestal veel sneller af dan de coëfficiënten a_k (k = 0(1)m), zodat, om voldoende overeenstemming met f(x) te houden op het interval [-1,+1], veelal minder termen van de Chebyshev-reeks nodig zijn dan van de polynoombenadering (1.5.1).

Dan hebben we dus voor zekere $n < m$:

1.5.3 $$f(x) \approx P_m(x) \approx P_n(x) = \sum_{k=0}^{n} b_k T_k(x).$$

Vervolgens schrijven we de Chebyshev-polynomen uit en nemen de termen van dezelfde macht van x samen. Dan krijgen we $P_n(x)$ in expliciete polynoom-vorm:

1.5.4 $$P_n(x) = a_0^* + a_1^* x + \ldots + a_n^* x^n.$$

Dit proces, dat uitgaande van een machtreeks $P_m(x)$ een economischer machtreeks $P_n(x)$ levert, heet "economiseren van machtreeksen".
In de praktijk heeft men de tussenstadia (1.5.2) en (1.5.3) niet eens nodig en kan men handiger als volgt te werk gaan.
Ga na of $\frac{|a_m|}{2^{m-1}}$ verwaarloosbaar klein is. Dan is blijkbaar het polynoom

1.5.5 $$P_{m-1}(x) = P_m(x) - \frac{a_m}{2^{m-1}} T_m(x)$$

van de graad m-1 eveneens een voldoende nauwkeurige benadering van f(x). Bereken de coëfficiënten van dit polynoom en herhaal hiermee het proces. Zo bereiken we tenslotte een polynoom $P_n(x)$ van de graad n, waarvan de hoogste coëfficiënt niet klein genoeg is in absolute waarde, om het proces te mogen voortzetten.
Bovendien vinden we op deze wijze de kleinste waarde van n, waarvoor $P_n(x)$ nog een voldoende benadering van $P_m(x)$ en dus van f(x) is.

Voorbeeld

Gevraagd een polynoom-benadering van $\cos \frac{\pi}{4} x$, die op het interval $[-1,+1]$ 2 decimalen nauwkeurigheid levert.

$$\cos \tfrac{\pi}{4} x = 1 - \frac{\pi^2}{32} x^2 + \frac{\pi^4}{6144} x^4 - \ldots$$

$$\approx 1 - .30842\, x^2 + .01585\, x^4$$

$\frac{a_4}{8} \approx .002$ is kleiner dan de vereiste nauwkeurigheid .005.
Dus we mogen aftrekken

$$\frac{a_4}{8} T_4(x) = a_4(x^4 - x^2 + \frac{1}{8})$$

en vinden dan de tweede-graads polynoom-benadering:

$$\cos \frac{\pi}{4} x \approx .99802 - .29257\ x^2.$$

Deze benadering is nauwkeuriger dan het resultaat van het voorbeeld op pag. 274.

Opgaven

143) (uit C.E. Fröberg, Introduction to numerical analysis).
Bepaal de absoluut kleinste eigenwaarde en bijbehorende eigenvector van de matrix

$$\begin{pmatrix} 1 & 2 & -2 & 4 \\ 2 & 12 & 3 & 5 \\ 3 & 13 & 0 & 7 \\ 2 & 11 & 2 & 2 \end{pmatrix}.$$

Maak hierbij gebruik van het feit, dat de inverse matrix de inverse eigenwaarden en dezelfde eigenvectoren heeft.

144) Economiseer het polynoom

$$1 + \frac{1}{2} x + \frac{1}{4} x^2 + \frac{1}{8} x^3 + \frac{1}{16} x^4$$

tot een polynoom van de graad 3 en daarna tot een van de graad 2.
Wat is de maximale fout hierbij?

145) Economiseer de Taylor-reeksen van de volgende functies zodanig, dat op het interval $[-1,+1]$ een precisie van 4 decimalen bereikt wordt.

$$\sin(x)\ ,\quad \cos(\tfrac{\pi}{4} x)\ ,\quad e^{-x}\ ,\quad 2^{0.5(x+1)}.$$

146) a) Schrijf een ALGOL-procedure, die een polynoom van de graad m economiseert op het interval $[a,b]$ tot een polynoom van zo laag mogelijke graad, waarbij de fout niet groter dan een gegeven ε mag worden.
b) Schrijf hieromheen een programma, dat de resultaten van opgave (145) kan toetsen.

1.6 Chebyshev-approximatie

We gaan nu na welk polynoom van graad kleiner dan n een functie f op een interval $[a,b]$ zo goed mogelijk benadert.
M.a.w. we zoeken een polynoom $\tilde{f}_n$ van graad kleiner dan n waarvoor de maximale fout $m = \max_{a \leq x \leq b} |f(x) - \tilde{f}_n(x)|$ minimaal is. Een dergelijk polynoom heet "beste benadering van f op $[a,b]$ in de zin van Chebyshev" of ook "minimax polynoom-benadering van f op $[a,b]$".
Een belangrijke eigenschap hiervan blijkt uit de

Tweede stelling van Chebyshev.

Zij f een continue functie gedefiniëerd op het interval $[a,b]$ en zij P_n een polynoom van graad kleiner dan n met de eigenschap dat de afwijking $\alpha(x) = f(x) - P_n(x)$ zijn grootste absolute waarde L bereikt in minstens n+1 punten van het interval $[a,b]$, waar de waarde afwisselend +L resp. -L is. Dan is P_n een beste benadering van f op $[a,b]$ in de zin van Chebyshev.

Bewijs: Stel er is een ander polynoom Q_n van graad kleiner dan n, dat een betere benadering van f is. M.a.w. de afwijking $\beta(x) = f(x) - Q_n(x)$ is dan op het interval $[a,b]$ in absolute waarde kleiner dan L. Dan hebben we: $\alpha(x) - \beta(x) = Q_n(x) - P_n(x)$; dit is blijkbaar een polynoom van graad kleiner dan n, dat in de bovengenoemde n+1 punten een afwisselend teken heeft. Hiertussen moet $\alpha(x) - \beta(x)$ dus n nulpunten bezitten en moet dus, als polynoom van graad kleiner dan n, identiek nul zijn. Dus $\alpha(x) = \beta(x)$, m.a.w. $P_n(x) = Q_n(x)$, wat een tegenspraak oplevert. Hiermee is de stelling bewezen.

Voorbeelden

Voor zeer lage waarden van n is een beste benadering in de zin van Chebyshev nog gemakkelijk te vinden. Voor n = 1 is de benadering blijkbaar de constante functie c, die gelijk is aan het gemiddelde van het minimum en het maximum van f op het interval $[a,b]$. Voor n = 2 is het probleem eenvoudig, als f een convexe of concave functie is op het interval $[a,b]$.

Bijvoorbeeld: zij gevraagd die lineaire functie die $\sqrt{x}$ het beste benadert in de zin van Chebyshev op het interval $[\frac{1}{4},1]$.
De kromme $y = \sqrt{x}$ ligt op het interval $[\frac{1}{4},1]$ geheel boven de lijn door de eindpunten $(\frac{1}{4},\frac{1}{2})$ en $(1,1)$, dat is de lijn $y = \frac{2}{3}x + \frac{1}{3}$.
De beste lineaire benadering van $\sqrt{x}$ moet blijkbaar evenwijdig hieraan lopen, dus $y = \frac{2}{3}x + \frac{1}{3} + b$, waarbij b de helft is van het maximum van van $g(x) = \sqrt{x} - \frac{2}{3}x - \frac{1}{3}$. Dit maximum vinden we door de afgeleide nul te stellen: $g'(x) = \frac{1}{2\sqrt{x}} - \frac{2}{3}$, dus $x = \frac{9}{16}$ en $b = \frac{1}{2}\, g\,(\frac{9}{16}) = \frac{1}{48}$.
De beste lineaire benadering van $\sqrt{x}$ op het interval $[\frac{1}{4},1]$ is dus $\frac{2}{3}x + \frac{17}{48}$ en de fout is in absolute waarde hoogstens gelijk aan $\frac{1}{48}$.

Het Austauschverfahren van Stiefel

Een minimax-polynoombenadering kan door middel van een iteratief proces verkregen worden als volgt.
De iteratie start met een (geschikt gekozen) rij basis-punten $x_0, \ldots, x_n$ in het interval $[a,b]$. Vervolgens wordt een polynoom P_n van graad kleiner dan n bepaald zodanig, dat de afwijking $\alpha(x) = f(x) - P_n(x)$ alternerende waarden e resp. -e op deze basis-punten heeft. P_n kan worden gevonden als volgt.
Bereken het polynoom f^*_{n+1} van de graad $\leq n$ dat in de basispunten $x_0, \ldots, x_n$ met f overeenstemt. Men kan dit doen met de formule van Newton (zie pag. 51), die men daarna in de expliciete polynoom-vorm van Grünert (zie pag. 34) kan omzetten. Bereken daarna evenzo het polynoom k^*_{n+1} van de graad $\leq n$, dat in de basispunten $x_0, \ldots, x_n$ de alternerende waarden $\neq 1$ heeft. Vervolgens bepalen we het getal e zó, dat

$$P_n(x) = f^*_{n+1}(x) - e\, k^*_{n+1}(x)$$

van de graad kleiner dan n is, m.a.w. de x^n-term moet hierin wegvallen. Dit geschiedt blijkbaar voor

$$e = f[x_0, \ldots, x_n]/k[x_0, \ldots, x_n],$$

omdat het n-de differentie-quotiënt tevens de coëfficiënt van x^n is.
Na de bepaling van het polynoom P_n berekenen we

$$h = \max_{a \leq x \leq b} |f(x) - P_n(x)|$$

en een argument y, waarvoor deze maximale waarde bereikt wordt.
Vervolgens wordt y in de rij der basispunten opgenomen en een der x_i daaruit verwijderd zódanig, dat weer een rij basispunten met afwijkingen van alternerend teken ontstaat. Meestal lukt dit door het dichtst bij y gelegen basispunten met afwijking van hetzelfde teken als de afwijking in y te vervangen door y. Als echter y buiten het interval $[x_0, x_n]$ ligt, moet soms elk basispunt een plaatsje opschuiven en het basispunt aan het andere eind van de rij verdwijnen.

Voor de aldus verkregen rij basispunten kan men wederom een polynoom van graad $< n$ met alternerende afwijkingen in de basispunten bepalen, enz.

Indien dit proces convergeert, vinden we tenslotte basispunten, die samenvallen met de plaatsen waar de afwijking $f(x) - P_n(x)$ extreme waarden heeft. Volgens de tweede stelling van Chebyshev is P_n dan de gezochte beste benadering van f in de zin van Chebyshev. Men kan de iteratie beëindigen als h en e weinig (bv. 1%) verschillen. De benadering kan dan niet noemenswaard meer verbeterd worden, omdat de maximale fout minstens e zal blijven.
De bepaling van de absoluut maximale afwijking h wordt veel eenvoudiger en het proces veel efficiënter (vooral voor ingewikkelde functies f), als de functie f alleen wordt beschouwd op een aantal discrete punten, bijvoorbeeld de equidistante punten $y_i = a + i(b-a)/m$ $(i = 0(1)m)$.
In dit geval convergeert het proces altijd, en wel tot de beste benadering van f op de genoemde discrete verzameling punten.
Als start voor de basispunten $x_0, \ldots, x_n$ kan men meestal heel geschikt kiezen de punten, waar het n-de graads Chebyshev-polynoom, aangepast aan het interval $[a,b]$, zijn extreme waarden ± 1 bereikt.
Zie voor bijzonderheden het hierna volgende testprogramma van procedure SPIEGEL.

```
begin comment  R1086, programma van A. van Praag gewijzigd, TJD
    221066. Dit programma dient voor het testen van procedure
    STIEFEL;
    real pi;
    integer n;

    real procedure MAX(k, a, b, fk); value a, b; integer k, a, b;
    real fk;
    begin real r, s;
    MA: k:= a; if k ≤ b then MAX:= s:= fk; goto MC;
    MB: k:= k + 1; r:= fk; if r > s then
        begin MAX:= s:= r; a:= k end;
    MC: if k < b then goto MB; k:= a
    end MAX;


    real procedure polynoom(x, graad, A); value x; real x; integer graad;
    array A; comment  [0: graad];
    begin integer i;
        real r;
        r:= 0;
        for i:= graad step - 1 until 0 do r:= r × x + A[i];
        polynoom:= r
    end polynoom;


    procedure NEWTON(f, x, a, n); value n; integer n; array a, x, f;
    comment  [0: n];
    begin integer k, i;
        real b, d;
        array c[0:n];
        a[0]:= c[0]:= f[0];
        for i:= 1 step 1 until n do
        begin c[i]:= f[i];
            for k:= 1 step 1 until i do
            begin b:= c[i]; c[i]:= (b - c[k - 1]) / (x[i] - x[i - k]);
                d:= c[k - 1]; c[k - 1]:= b
            end;
            a[i - 1]:= d
        end;
        a[n]:= c[n]
    end NEWTON;


    procedure GRUENERT(x, a, b, n); value n; integer n; array a, b, x;
    comment [0: n - 1];
    begin integer k, i;
        array c[ - 1:n - 1];
        c[ - 1]:= 0;
```

```
    for k:= 0 step 1 until n - 1 do
    begin b[k]:= 0; c[k]:= 1;
        for i:= k step - 1 until 0 do
        begin b[i]:= b[i] + c[i] × a[k];
            c[i]:= c[i - 1] - c[i] × x[k]
        end
    end
end GRUENERT;

comment  STIEFEL bepaalt ter benadering van de gegeven functie f
op het interval[p, q] een polynoom van de graad n - 1 met
coefficienten in co[0: n - 1, zodat op minstens m + 1
equidistante punten in het interval [p, q] het maximale verschil
h tussen f en het polynoom in absolute waarde zo klein mogelijk
wordt. In x[0: n] staan de argument waarden horende bij de
gevonden maximale afwijkingen en count is het aantal iteraties
(≤ 30). STIEFEL gebruikt de procedures NEWTON, GRUENERT, MAX en
polynoom;

procedure STIEFEL(f, p, q, m, n, co, x, h, count);
value p, q, m, n; integer m, n, count; real p, q, h; array co, x;
real procedure f;
begin integer k, i;
    real hh, r, pi;
    pi:= 3.14159265359;
    if (1 - cos(pi / n)) × m < 2 then m:= 2 / (1 - cos(pi / n));
    begin integer array s[0:n];
        array a, b, fun, plusmin[0:n], y, fy, g[0:m];
        s[0]:= 0; x[0]:= p; plusmin[0]:= 1;
        for k:= 1 step 1 until n do
        begin s[k]:= m × (1 - cos(k × pi / n)) / 2;
            x[k]:= p + s[k] × (q - p) / m;
            plusmin[k]:= - plusmin[k - 1]
        end;
        for i:= 0 step 1 until m do
        begin y[i]:= p + i × (q - p) / m; fy[i]:= f(y[i]) end;
        count:= 0;
    iter: count:= count + 1;
        for k:= 0 step 1 until n do fun[k]:= fy[s[k]];
        NEWTON(fun, x, a, n); NEWTON(plusmin, x, b, n);
        hh:= - a[n] / b[n];
        for k:= 0 step 1 until n - 1 do a[k]:= a[k] + hh × b[k];
        GRUENERT(x, a, co, n);
        for i:= 0 step 1 until m do g[i]:= fy[i] - polynoom(y[i],
        n - 1, co); h:= MAX(i, 0, m, abs(g[i]));
        if abs(hh) < .99 × h ∧ count < 30 then
        begin integer c, d;
            real a, b;
            c:= sign(g[i]); if s[0] ≤ i ∧ i ≤ s[n] then
            begin a:= abs(x[n] - x[0]);
```

```
                    for k:= 0 step 1 until n do
                    begin b:= abs(y[i] - x[k]);
                        if b < a ∧ c × g[s[k]] ≥ 0 then
                        begin d:= k; a:= b end
                    end;
                    x[d]:= y[i]; s[d]:= i
                end
                else if i < s[0] then
                begin if c × g[s[0]] < 0 then
                    for k:= n step - 1 until 1 do
                    begin x[k]:= x[k - 1]; s[k]:= s[k - 1] end;
                    x[0]:= y[i]; s[0]:= i
                end
                else
                begin if c × g[s[n]] < 0 then
                    for k:= 1 step 1 until n do
                    begin x[k - 1]:= x[k]; s[k - 1]:= s[k] end;
                    x[n]:= y[i]; s[n]:= i
                end;
                goto iter
            end
        end
    end STIEFEL;


    real procedure F(x); value x; real x;
    F:= if x ≤ 0.5 then sin(pi × x) else exp(0.5 - x);


    procedure COMPUTE(b1, b2, b3); value b1, b2; real b1, b2;
    real procedure b3;
    begin integer t, telling;
        real precisie;
        array coef[0:n - 1], arg[0:n];
        STIEFEL(b3, b1, b2, 10 × n, n, coef, arg, precisie, telling);
        PRINTTEXT(‹graad, precisie en telling›); NLCR; print(n - 1);
        print(precisie); print(telling); NLCR; PRINTTEXT(
        ‹coefficienten›); NLCR;
        for t:= 0 step 1 until n - 1 do print(coef[t]); NLCR;
        PRINTTEXT(‹argumenten van maximale afwijking›); NLCR;
        for t:= 0 step 1 until n do print(arg[t]); NLCR; NLCR
    end COMPUTE;

    pi:= 3.14159265359; PRINTTEXT(‹sin›); NLCR; NLCR;
    for n:= 2 step 1 until 8 do COMPUTE( - pi / 2, pi / 2, sin); NLCR;
    PRINTTEXT(‹f›); NLCR; NLCR;
    for n:= 2 step 1 until 10 do COMPUTE(0, 1, F); NLCR; PRINTTEXT(
    ‹abs›); NLCR; NLCR;
    for n:= 2, 5, 8, 10 do COMPUTE( - 1, 1, abs)
end
```

Voor theorie betreffende het Austauschverfahren zie: E. Stiefel, Ueber diskrete und lineare Tschebyscheff-Approximationen, Num. Mat. 1 (1959), 1-28.

Opgave

147) Bepaal die lineaire functie, die op het interval $[a,b]$ de functie f zo goed mogelijk benadert in de zin van Chebyshev, als

1e) $f(x) = \sin(x)$, $a = 0$, $b = \pi/2$;

2e) $f(x) = e^{-x}$, $a = 0$, $b = 2$;

3e) $f(x) = \cos(\sqrt{x})$, $a = 0$, $b = (\pi/4)^2$.

Bepaal tevens welke precisie bereikt wordt.

2. Kleinste kwadratenbenadering

2.1. Polynoom-benadering voor het discrete geval

Zij gegeven een functie f op m verschillende basispunten x_k, $k = 0(1)m-1$ en zij gevraagd f te benaderen door een polynoom van de graad kleiner dan n, m.a.w.

2.1.1 $$f(x) \approx P(x) = a_0 + a_1 x + \ldots + a_{n-1} x^{n-1}.$$

We nemen aan dat n (veel) kleiner is dan m, zodat P niet exact met f kan overeenstemmen in alle m basispunten. Kiezen we het polynoom P zó, dat het exact met f overeenstemt in een deel der gegeven basispunten, dan krijgen we geen redelijke overeenstemming in de andere basispunten. In de praktijk zijn de gegeven functie-waarden ofwel gemeten (physische) grootheden ofwel berekende waarden van een (mathematische) functie. Men wil daarom alle gegeven waarden gelijkelijk tot hun recht laten komen in de op te stellen benadering. Hiertoe is geschikt (onder andere) de benadering in de zin der kleinste kwadraten, gedefiniëerd als volgt.

Zij voor elk basispunt het residu gedefiniëerd door

2.1.2 $$r_k = f(x_k) - P(x_k),\ k = 0(1)m-1.$$

Dan is het beste polynoom van de graad kleiner dan n in de zin der kleinste kwadraten per definitie dát polynoom P, waarvoor

2.1.3 $$R = \sum_{k=0}^{m-1} r_k^2$$

minimaal is.
Deze "som der kwadraten der residuen" is blijkbaar een functie van de coëfficiënten van P. Teneinde een minimale waarde te bereiken, is het nodig, dat de partiële afgeleiden van R naar de coëfficiënten van P nul zijn. M.a.w. de coëfficiënten van P moeten voldoen aan

2.1.4 $$\frac{\partial R}{\partial a_i} = -2 \sum_{k=0}^{m-1} \left(f(x_k) - \sum_{j=0}^{n-1} a_j x_k^j\right) x_k^i = 0,\ i = 0(1)n-1.$$

Dit zijn n lineairevergelijkingen in de n onbekenden $a_0, \ldots, a_{n-1}$.
Dit stelsel kan in matrix-vector-notatie geschreven worden als volgt.
Zij V de (m × n)-matrix

2.1.5
$$\begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^{n-1} \\ 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & x_{m-1} & x_{m-1}^2 & & x_{m-1}^{n-1} \end{pmatrix}.$$

Zij verder $\vec{f}$ de vector der functie-waarden $f_k = f(x_k)$ en $\vec{r}$ de vector der residuen r_k, k = 0(1)m-1. Zij tenslotte $\vec{a}$ de vector der coëfficiënten a_j, j = 0(1)n-1.
Dan kan (2.1.2) geschreven worden in de gedaante

2.1.6
$$\vec{r} = \vec{f} - V\vec{a}$$

en het stelsel (2.1.4) krijgt de vorm

$$V^T\vec{r} = V^T\vec{f} - V^TV\vec{a} = 0,$$

m.a.w.

2.1.7
$$V^TV\vec{a} = V^T\vec{f}.$$

Lichten we uit matrix V n rijen, dan krijgen we een vierkante matrix van Van der Monde; de determinant hiervan is ongelijk aan nul (zie pag. 30, 31). M.a.w. de matrix V heeft de rang n (d.w.z. n lineair onafhankelijke kolommen). De matrix V^TV van het stelsel (2.1.7) is dus symmetrisch en positief definiet volgens stelling (4.1.2) op pag. 140. Dit stelsel kan dus worden opgelost met behulp van de wortel-methode van Cholesky (zie pag. 141). Helaas is dit stelsel vaak slecht geconditioneerd als de orde n groot is. Hierop komen we later terug.

Opmerkingen

1) Het komt nogal eens voor, dat de nauwkeurigheid der metingen in de basispunten niet overal hetzelfde is. Dit kan men tot uiting brengen door de residuen r_k een van k afhankelijk gewicht $w_k > 0$ te geven. M.a.w. de te minimaliseren grootheid is dan

2.1.3a $$R = \sum_{k=0}^{m-1} w_k^2 r_k^2.$$

Dit leidt eveneens tot een stelsel van de gedaante (2.1.7), waarbij nu evenwel de elementen van $\vec{f}$ zijn $w_k f_k$, k = 0(1)m-1, en V de matrix is, die uit (2.1.5) ontstaat door hierin de rijen resp. te vermenigvuldigen met $w_0, \ldots, w_{m-1}$.

2) Er bestaan ook kleinste kwadraten-problemen van algemener type, waarbij niet speciaal gedacht wordt aan een polynoom-benadering van f. Hierin is V een willekeurige gegeven (m × n)-matrix van de rang n en men tracht R (zie 2.1.3) te minimaliseren, waarbij de residu-vector $\vec{r}$ wederom de gedaante (2.1.6) heeft. Dit leidt, op volkomen analoge wijze als boven, tot een stelsel van de gedaante (2.1.7).
De betekenis van een dergelijk probleem is, dat men de vector $\vec{f}$ zo goed mogelijk tracht te schrijven als een lineaire combinatie van de kolommen van V, of wat men ook zegt: dat men de vector $\vec{f}$ zo goed mogelijk tracht te "verklaren" uit de kolommen van V.
Ook bij deze problemen dreigt het spook van de slechte conditie, zodra de orde n beduidend groot wordt.
Voordat we dit verder nagaan, zullen we eerst het continue geval bekijken.

Opgaven

148) a) Schrijf een ALGOL-procedure, die, gegeven een tabel van m argumenten en bijbehorende functie-waarden, die lineaire functie bepaalt, die de getabelleerde functie in de zin van kleinste kwadraten zo goed mogelijk benadert.

b) Schrijf hieromheen een programma, dat de tabel inleest en de coëfficiënten van de lineaire kleinste-kwadratenbenadering zowel als de wortel uit de som der kwadraten der residuen uittypt.

149) Bereken de lineaire kleinste-kwadratenbenadering van de functie sin(x), gegeven op m equidistante punten tussen $-\frac{\pi}{4}$ en $+\frac{\pi}{4}$. Doe dit achtereenvolgens voor m = 3, 7, 11 en 19. Toets deze resultaten met programma (148b) of een ander programma, dat procedure (148a) gebruikt.

150) Inverteer de volgende matrix met behulp van de methode van Cholesky

$$\begin{pmatrix} 1 & 2 & 0.5 & 1 \\ 2 & 5 & 0 & -2 \\ 0.5 & 0 & 2.25 & 7.5 \\ 1 & -2 & 7.5 & 27 \end{pmatrix}$$

2.2 Polynoom-benadering voor het continue geval

Nu beschouwen we een functie f gedefinieerd op een interval $[a, b]$, waarvoor wederom een benaderend polynoom P van de graad kleiner dan n gevraagd wordt (zie 2.1.1).
In elk punt x van het interval $[a, b]$ is het residu gedefinieerd door

2.2.1 $$r(x) = f(x) - P(x) = f(x) - a_0 - a_1x - \ldots - a_{n-1}x^{k-1}.$$

Het polynoom van de graad kleiner dan n dat f op $[a, b]$ zo goed mogelijk benadert in de zin der kleinste kwadraten is nu per definitie dat polynoom P, waarvoor

2.2.2 $$R = \int_a^b (r(x))^2 dx$$

minimaal is. R is wederom een functie van de coëfficiënten van P.
Om een minimale waarde te bereiken, moeten de partiele afgeleiden van R naar de coëfficiënten van P nul zijn.
Dit leidt, evenals bij het discrete geval, tot een stelsel van n lineaire vergelijkingen in de n onbekenden $a_0, \ldots, a_{n-1}$.
Om een eenvoudiger lineair stelsel te krijgen, gaan we het polynoom P niet in de expliciete vorm (zie 2.1.1) schrijven, maar als volgt:

2.2.3 $$P(x) = b_0P_0(x) + b_1P_1(x) + \ldots + b_{n-1}P_{n-1}(x),$$

waarbij P_j, $j = 0(1)n-1$, polynomen van de graad j zijn, die orthogonaal zijn wat betreft integratie over het interval $[a, b]$, d.w.z.:

2.2.4 $$\int_a^b P_i(x)\, P_j(x)dx = 0, \quad \text{als } i \neq j.$$

De te minimaliseren grootheid R krijgt dan de gedaante:

2.2.5 $$R = \int_a^b (f(x))^2 dx - 2 \sum_{j=0}^{n-1} b_j \int_a^b f(x)\, P_j(x)dx + \sum_{j=0}^{n-1} b^2 \int_a^b (P_j(x))^2 dx.$$

Voor het bereiken van een minimum moeten de partiele afgeleiden naar de coëfficiënten b_i, $i = 0(1)n-1$, nul zijn, m.a.w.

2.2.6

$$\frac{\partial R}{\partial b_i} = -2\int_a^b f(x)\,P_i(x)dx + 2b_i\int_a^b (P_i(x))^2dx = 0, \quad i = 0(1)n-1.$$

Dit is een zeer eenvoudig stelsel lineaire vergelijkingen, want de matrix van het stelsel heeft de diagonaalvorm, zodat we de oplossing meteen kunnen neerschrijven:

2.2.7 $$b_i = \int_a^b f(x)\,P_i(x)dx \Big/ \int_a^b (P_i(x))^2dx.$$

Polynomen van Legendre

We voeren het interval $[a, b]$ over in het standaard-interval $[-1, +1]$ door middel van de lineaire transformatie (1.3.1). De orthogonaliteitsrelatie (2.2.4) gaat dan over in

2.2.8 $$\int_{-1}^{+1} P_i(x)\,P_j(x)dx = 0, \quad \text{als } i \neq j.$$

Deze relatie legt de polynomen P_j op een constante factor na vast. Om deze constante te fixeren, is nog een normerings-voorwaarde nodig. Hiervoor kiest men vaak de (vrij willekeurige) eis:

2.2.9 $$P_j(1) = 1, \quad j \geq 0.$$

De polynomen, die hieraan voldoen, heten polynomen van Legendre. Voor kleine waarden van j vinden we

(2.2.10)
$$\begin{cases} P_0(x) \equiv 1, \\ P_1(x) = x, \\ P_2(x) = \frac{1}{2}(3x^2 - 1), \\ P_3(x) = \frac{1}{2}(5x^3 - 3x), \\ P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3). \end{cases}$$

Veel gemakkelijker, dan uit (2.2.8 en 2.2.9), kunnen de polynomen van Legendre verkregen worden uit de recurrente betrekking

2.2.11 $$P_j(x) = (2 - \tfrac{1}{j})x\,P_{j-1}(x) - (1 - \tfrac{1}{j})\,P_{j-2}(x), \quad j \geq 2.$$

Verder is nog van belang de eigenschap

2.2.12 $$\int_{-1}^{+1} (P_j(x))^2 dx = 2/(2j + 1), \quad j \geq 0.$$

De formules (2.2.11 en 2.2.12) kunnen samen worden afgeleid door volledige inductie.

Keren we nu terug tot het continue kleinste-kwadratenprobleem.

Zij gevraagd een gegeven functie f op het interval $[-1, +1]$ in de zin der kleinste kwadraten te benaderen door een polynoom P van graad kleiner dan n. Als P geschreven is in de gedaante (2.2.3), dan volgen de coëfficiënten b_i, $i = 0(1)n-1$, uit (2.2.7 en 2.2.12), dus

2.2.13 $$b_i = (i + \tfrac{1}{2}) \int_{-1}^{+1} f(x)\, P_i(x) dx, \quad i = 0(1)n-1.$$

Hiermee is het polynoom P volledig vastgelegd.

Alleen moet P nog in expliciete vorm (zie 2.1.1) omgeschreven worden, wat gemakkelijk gaat met behulp van een tabel van Legendre-polynomen (zie 2.2.10).

Opmerking

Ook in het continue geval kan men een van x afhankelijke gewichts-functie $w(x) > 0$ invoeren (vgl. opmerking 1 pag. 289). De te minimaliseren grootheid luidt dan

2.2.2a $$R = \int_a^b w(x)\, (r(x))^2 dx.$$

Hierbij schrijft men weer P in de gedaante (2.2.3) waarbij de polynomen P_j nu moeten voldoen aan de orthogonaliteits-relatie

2.24a $$\int_a^b w(x)\, P_i(x)\, P_j(x) dx = 0, \quad \text{als } i \neq j.$$

Dan vindt men de coëfficiënten b_i weer uit een lineair stelsel, waarvan de matrix de diagonaalvorm heeft.

Bij elke gewichtsfunctie w(x), met grenzen a en b, hoort een stelsel orthogonale polynomen. Bijvoorbeeld, bij $w(x) = \frac{1}{\sqrt{1-x^2}}$, $a = -1$, $b = +1$ horen als orthogonale polynomen de polynomen van Chebyshev (zie (1.2) en formule (1.4.3)).

Opgaven

151) a) Zij gevraagd de functie $\sin \frac{\pi}{2} x$ in de zin der kleinste kwadraten te benaderen op het interval $[-1,+1]$ door een polynoom P van de graad 3.
Stel het hierbij horende lineaire stelsel voor de coëfficiënten van P op en bepaal hieruit P.
Bepaal vervolgens P opnieuw, nu met behulp van de polynomen van Legendre.

b) Dezelfde opgave voor de functie e^x op het interval $[0,1]$.

152) a) Bepaal de polynomen van Legendre tot en met graad 5 (vgl. 2.2.10).

b) Gevraagd de "verschoven" polynomen van Legendre, die voldoen aan de orthogonaliteits-relatie

$$\int_0^1 P_i(x)P_j(x)dx = 0 \quad , \quad \text{als } i \neq j,$$

en aan de normerings-eis (2.2.9), te bepalen tot en met graad 4. Bereken voor deze polynomen tevens

$$\int_0^1 (P_j(x))^2 dx.$$

153) a) Schrijf een Algol-procedure, die de coëfficiënten van de polynomen van Legendre tot een zekere graad uitrekent.

b) Zij gevraagd een functie f in de zin der kleinste kwadraten te benaderen op het interval $[a,b]$ door een polynoom P van de graad kleiner dan n.
Schrijf een Algol-procedure, die de coëfficiënten van P bepaalt, als gegeven zijn de "momenten"

$$M_i = \int_a^b x^i f(x)dx \quad , \quad i = 0(1)n-1.$$

c) Schrijf een Algol-programma, dat de oplossingen bepaalt van opgave 151 a en b.

2.3 Conditie van stelsels lineaire vergelijkingen

Beschouwen we het lineaire stelsel

2.3.1 $$Ax = b,$$

waarbij A een vierkante niet-singuliere matrix is.
Numeriek vindt men gewoonlijk niet de exacte oplossing x, maar een benadering $x + \delta x$.
Noemen we nu het hierbij horende rechterlid $b + \delta b$, m.a.w.

2.3.2 $$A(x + \delta x) = b + \delta b,$$

dan geldt blijkbaar $A\delta x = \delta b$, dus

2.3.3 $$\delta x = A^{-1}\delta b.$$

Als de elementen van de residu-vector δb klein zijn, zal de fout-vector δx ook kleine elementen hebben, mits de elementen van A^{-1} niet al te groot zijn. A^{-1} bepaalt dus min of meer de conditie van het stelsel.

We hebben liever een enkel getal als maat voor de conditie. Hiertoe hebben we nodig een vector-norm en een matrix-norm (die een maat zijn voor de grootte van een vector resp. matrix).
Op pagina 190 hebben we twee vector-normen vermeld, nl. de Euclidische lengte

2.3.4 $$||x||_2 = \sqrt{x_1^2 + \ldots + x_n^2}$$

en de maximum-norm

2.3.5 $$||x|| = \max_{i=1,\ldots,n} |x_i|.$$

Deze maximum-norm zullen we als uitgangspunt nemen voor het definiëren van de conditie van een stelsel.
Uit (2.3.3 & 5) volgt:

$$||\delta x|| \leq \max_i \sum_j |(A^{-1})_{ij}|\ |\delta b_j|$$

$$\leq (\max_i \sum_j |(A^{-1})_{ij}|) \cdot (\max_j |\delta b_j|).$$

De laatste factor is blijkbaar $||\delta b||$, de eerste factor hangt alleen van A^{-1} af en is gelijk aan $||A^{-1}||$, als we definiëren voor willekeurige vierkante matrices A:

2.3.6 $$||A|| =_{def} \max_i \sum_j |A_{ij}|.$$

Dit is de matrix-norm horende bij de maximum-vectornorm.
Dan hebben we dus

2.3.7 $$||\delta x|| \leq ||A^{-1}|| \cdot ||\delta b||.$$

M.a.w. $||A^{-1}||$ is een maat voor de conditie van het stelsel $Ax = b$.
We hebben hier de maximale <u>absolute fout</u> $||\delta x||$ beschouwd. Vaak kijkt men liever naar de <u>relatieve fout</u> van x, die gedefiniëerd wordt als $||\delta x||/||x||$.
Wegens $Ax = b$ hebben we

$$||b|| = ||Ax|| \leq ||A|| \cdot ||x||,$$

wat op precies dezelfde wijze wordt afgeleid als (2.3.7) uit (2.3.3).
Dus

$$\frac{||\delta x||}{||x||} \leq \frac{||A^{-1}|| \cdot ||\delta b||}{||x||} \leq \frac{||A||}{||b||} \cdot ||A^{-1}|| \cdot ||\delta b||,$$

m.a.w.

2.3.8 $$\frac{||\delta x||}{||x||} \leq ||A|| \cdot ||A^{-1}|| \cdot \frac{||\delta b||}{||b||}.$$

De hier optredende factor $||A|| \cdot ||A^{-1}||$ heet <u>conditie-getal</u> van matrix A.
Dit conditie-getal is altijd minstens 1; het is o.a. voor de eenheids-matrix, voor permutatie-matrices en voor veelvouden hiervan gelijk aan 1.
Als het conditie-getal van A erg groot is, noemen we het stelsel $Ax = b$ <u>slecht geconditioneerd</u>. Een kleine storing δb van de residu-vector kan dan een grote fout δx in de oplossing ten gevolge hebben.

<u>Voorbeeld</u> (vgl. pag. 144)

Het stelsel

$$\begin{pmatrix} 1 & -1 \\ -101 & 102 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

is slecht geconditioneerd.
Het conditie-getal van de matrix is ongeveer 20000, zodat de relatieve precisie van x ruim 4 decimalen slechter kan zijn, dan de relatieve fout van b. De exacte oplossing is $x_1 = x_2 = 1$. Voor de benaderde oplossing (1.101, 1.100) vinden we

$$\frac{||\delta x||}{||x||} = .101 \quad \text{en} \quad \frac{||\delta b||}{||b||} = .001,$$

zodat de oplossing iets beter is, dan op grond van de conditie te verwachten was.

Equilibreren

Vermenigvuldigt men de vergelijkingen met zekere factoren $\neq 0$ zodanig, dat alle rijen van A dezelfde lengte krijgen (bv. Euclidische lengte of maximum-norm), dan blijft de oplossing uiteraard dezelfde, maar het conditie-getal van de matrix wordt vaak kleiner. Een matrix met rijen (en kolommen) van dezelfde lengte heet wel "geëquilibreerde matrix". Vermenigvuldigen we in bovengenoemd voorbeeld de eerste vergelijking met 100 dan is het conditie-getal van de matrix ongeveer 400, terwijl we voor bovengenoemde benaderde oplossing hebben

$$||\delta x||/||x|| = .101 \text{ en } ||\delta b||/||b|| = 0.1.$$

Gaan we van een stelsel over op een equivalent stelsel met geëquilibreerde matrix, dan wordt hierdoor het zoeken van de pivots (zie pag. 125 en 135) beïnvloed. De keuze, die afgestemd is op de geëquilibreerde matrix, verdient een zekere voorkeur.
Dit geschiedt in feite in de procedure "DET" (AP 204, zie pag. 136 en 137), waar de candidaat-pivots eerst door de (Euclidische) lengte van de betreffende rij van de gegeven matrix worden gedeeld, alvorens degene met maximale absolute waarde hieruit wordt gekozen.

2.4 Oplossing van het discrete kleinste-kwadraten-probleem door middel van orthogonalisatie.

We beschouwen nu het algemene discrete kleinste-kwadratenprobleem (vgl. pag. 289 opm. 2), waarin gegeven is een willekeurige (m × n)-matrix A van de rang n en een m-vector b, waarvoor gevraagd wordt te minimaliseren (het kwadraat van) de Euclidische lengte (vgl. 2.3.4) van de residu-vector

2.4.1 $$r = b - Ax,$$

m.a.w. de te minimaliseren grootheid is

2.4.2 $$R = r^T r = \sum_{k=0}^{m-1} r_k^2.$$

De oplossingsvector x, bestaande uit de elementen $x_0, \ldots, x_{n-1}$, moet dus voldoen aan de eis, dat de partiele afgeleiden van R naar de elementen x_i nul zijn:

2.4.3 $$\frac{\partial R}{\partial x_i} = -2 \sum_{k=0}^{m-1} \left(b_k - \sum_{j=0}^{n-1} A_{kj}\, x_j\right) A_{ki} = 0, \quad i = 0(1)n-1.$$

Anders gezegd: x moet voldoen aan het lineaire stelsel

2.4.4 $$A^T A\, x = A^T b$$

(vgl. formules 2.1.4 & 7).
Als A de rang n heeft, is $A^T A$ niet singulier en heeft 2.4.4 dus een en slechts een oplossing. Dit stelsel is evenwel vaak slecht geconditioneerd voor (tamelijk) grote n, zoals moge blijken uit het volgende

Voorbeeld

Zij A de matrix

2.4.5 $$\begin{pmatrix} 1 & 1 & 1 & 1 \\ \varepsilon & 0 & 0 & 0 \\ 0 & \varepsilon & 0 & 0 \\ 0 & 0 & \varepsilon & 0 \\ 0 & 0 & 0 & \varepsilon \end{pmatrix}$$

dan geldt:

$$A^TA = \begin{pmatrix} 1+\varepsilon^2 & 1 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 & 1 \\ 1 & 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1 & 1+\varepsilon^2 \end{pmatrix}, \quad (A^TA)^{-1} = \frac{1}{\varepsilon^2(4+\varepsilon^2)} \begin{pmatrix} 3+\varepsilon^2 & -1 & -1 & -1 \\ -1 & 3+\varepsilon^2 & -1 & -1 \\ -1 & -1 & 3+\varepsilon^2 & -1 \\ -1 & -1 & -1 & 3+\varepsilon^2 \end{pmatrix}$$

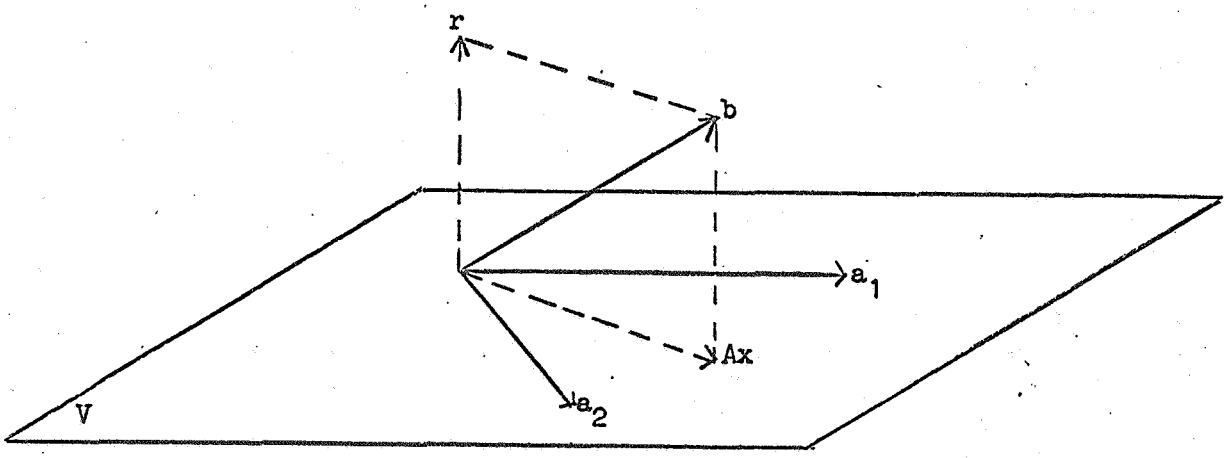
Dus $||A^TA|| = 4+\varepsilon^2$ en $||(A^TA)^{-1}|| = \frac{6+\varepsilon^2}{\varepsilon^2(4+\varepsilon^2)}$,

zodat het conditie-getal van A^TA bedraagt $\frac{6+\varepsilon^2}{\varepsilon^2}$

Voor kleine ε is A^TA dus slecht geconditioneerd.

Meetkundig beeld

De kolommen van A en vector b zijn vectoren in een m-dimensionale ruimte. Als A de rang n heeft, spannen de kolommen van A een n-dimensionale deelruimte V op. De vector Ax ligt blijkbaar in V en de Euclidische lengte van de residu-vector r is blijkbaar minimaal, als Ax gelijk is aan de projectie van b op de deelruimte V; de residu-vector r staat dan loodrecht op de deelruimte V (d.w.z. r staat loodrecht op alle in V gelegen vectoren).



Gram-Schmidt orthogonalisatie

Dit meetkundig beeld suggereert, dat het voor het oplossen van het kleinste-kwadratenprobleem nuttig is eerst de vectoren, die V opspannen, te vervangen door een orthogonale basis voor V(dat is een stelsel orthogonale vectoren van lengte een, die V opspannen).
De eenvoudigste methode hiervoor is de Gram-Schmidt orthogonalisatie, die als volgt verloopt.
Laat a_i de i-de kolom van A en q_i de i-de orthonormale basisvector aanduiden, voor i = 0(1)n-1. Dan worden de q_i achtereenvolgens berekent aldus (de hier optredende norm is de Euclidische lengte $||x|| = \sqrt{x^T x}$):
0) normeer a_0

2.4.6.0 $$q_0 = \frac{1}{||a_0||} a_0$$

1) bereken de component c_1 van a_1, die loodrecht op q_0 staat, en normeer:

2.4.6.1 $$c_1 = a_1 - (q_0^T a_1)q_0 \; ; \; q_1 = \frac{1}{||c_1||} c_1$$

en, algemeen voor i = 0(1)n-1,
bereken de component c_i van a_i, die loodrecht staat op $q_0, \ldots, q_{i-1}$ en normeer:

2.4.6.i $$c_i = a_i - (q_0^T a_i)q_0 - \ldots - (q_{i-1}^T a_i)q_{i-1}; \; q_i = \frac{1}{||c_i||} c_i.$$

De matrix Q bestaande uit de kolom-vectoren $q_0, \ldots, q_{n-1}$ is dan blijkbaar een orthogonale matrix (d.w.z. $Q^T Q = I$, maar $Q\,Q^T \neq I$, want Q is niet vierkant).
De orthogonalisatie betekent, dat A is geschreven als

2.4.7 $$A = Q\,U,$$

waarbij U gelijk is aan de bovendriehoeks-matrix

2.4.8

$$\begin{pmatrix} \|a_0\| & q_0^T a_1 & q_0^T a_2 & \cdots & q_0^T a_{n-1} \\ & \|c_1\| & q_1^T a_2 & \cdots & q_1^T a_{n-1} \\ & & \|c_2\| & & \vdots \\ & & & \ddots & q_{n-2}^T a_{n-1} \\ & & & & \|c_{n-1}\| \end{pmatrix}$$

Na het orthogonaliseren vinden we de kleinste-kwadratenoplossing als volgt. Uit (2.4.4) en (2.4.7) volgt

$$U^T Q^T Q\, U\, x = U^T Q^T b$$

Als A de rang n heeft is U^T niet-singulier en mag men dus met $(U^T)^{-1}$ voor-vermenigvuldigen. Wegens $Q^T Q = I$ levert dit dan

2.4.9 $$U\, x = Q^T b$$

Dit is een stelsel lineaire vergelijkingen met driehoeksmatrix, wat dus zonder moeite kan worden opgelost (terug-substitutie).

Voorbeeld

Gevraagd een functie f, die gegeven is op de punten ξ_k, k = 0(1)m-1, in de zin der kleinste kwadraten te benaderen door een lineaire functie. Dan hebben we

$$A = \begin{pmatrix} 1 & \xi_0 \\ 1 & \xi_1 \\ \vdots & \vdots \\ 1 & \xi_{m-1} \end{pmatrix} \qquad b = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{m-1} \end{pmatrix}$$

De Gram-Schmidt orthogonalisatie levert dan

$$Q = \begin{pmatrix} 1/\sqrt{m} & (\xi_0 - \bar{\xi})/\sigma \\ 1/\sqrt{m} & (\xi_1 - \bar{\xi})/\sigma \\ & \\ 1/\sqrt{m} & (\xi_{m-1} - \bar{\xi})/\sigma \end{pmatrix}, U = \begin{pmatrix} \sqrt{m} & \frac{1}{\sqrt{m}} \sum_{k=0}^{m-1} \xi_k \\ & \\ 0 & \sigma \end{pmatrix}$$

waarbij $\bar{\xi} = \frac{1}{m} \sum_{k=0}^{m-1} \xi_k$ en $\sigma = \sqrt{\sum_{k=0}^{m-1} (\xi_k - \bar{\xi})^2}$.

Bovendien hebben we

$$Q^T b = \begin{pmatrix} \frac{1}{\sqrt{m}} \sum_{k=0}^{m-1} f_k \\ \\ \frac{1}{\sigma} \sum_{k=0}^{m-1} (\xi_k - \bar{\xi})\, f_k \end{pmatrix}$$

Oplossing van het stelsel (2.4.9) levert dan de lineaire kleinste-kwadratenbenadering

$$y = \frac{1}{m} \sum_{k=0}^{m-1} f_k + \frac{1}{\sigma^2} \sum_{k=0}^{m-1} (\xi_k - \bar{\xi}) f_k (x - \bar{\xi})$$

(Vergelijk dit resultaat met opgave 148).

Opmerkingen

1) Men kan de Gram-Schmidt orthogonalisatie uitvoeren met verwisseling van de kolommen van A. In de i-de stap zoekt men dan uit de kolommen a_j met $j \geq i$ die kolom, die de grootste component loodrecht op $q_0, \ldots, q_{i-1}$ heeft, en verwisselt deze met de i-de kolom, waarna men c_i en q_i berekent.
Dit verwisselen is vooral van belang, als de kolommen van A (bijna) lineair afhankelijk zijn. Vindt men voor zekere i, dat alle resterende

kolommen een verwaarloosbaar kleine component loodrecht op de verkregen kolommen van Q hebben, dan is i blijkbaar de rang van A. Treedt dit evenwel voor geen der i-waarden op, dan heeft A de rang n en kan men doorgaan met het stelsel (2.4.9) op te lossen.

2) In plaats van Gram-Schmidt orthogonalisatie kan men ook Householder orthogonalisatie toepassen, waarbij Q gelijk is aan de eerste n kolommen van het product van n Householder-matrices van de orde m(zie pag. 206 formule 5.0.2).

Dit is geprogrammeerd in een ALGOL-procedure door G. Golub en P. Businger (Num. Mat. 7(1965) 269-276).

Opgaven

154a) Bewijs, dat het conditie-getal van een matrix altijd minstens een is.

b) Bereken de conditie van de matrices uit opgaven 143 en 150 en van de matrices der stelsels uit opgave 151 a en b.

155) Bereken, zowel met de hand als met het programma van opgave 148, de lineaire kleinste-kwadratenbenadering van de getabelleerde functie

x	f(x)
2	2
5	4
6	6
9	9
11	10

156) Voor de stralings-intensiteit I van een radio-actieve bron geldt de formule $I = I_0 e^{-\alpha t}$.
De logarithme van I is dus een lineaire functie van de tijd t.
Bepaal deze lineaire functie door middel van kleinste-kwadraten-benadering als de volgende tabel gegeven is.

t	0.2	0.3	0.4	0.5	0.6	0.7	0.8
I	3.16	2.38	1.75	1.34	1.00	0.74	0.56

Met name worden gevraagd de bijbehorende α en I_0.
Doe de berekening met de hand en met behulp van het programma uit opgave 148.

```
begin comment R 1086, program to test the procedure 'least squares
    solution' of G.Golub and Peter Businger, Num. Mat. 7 (1965)
    269-276. On the data tape one should give the number m of
    equations, the number n of unknowns, the number p of right
    hand sides and subsequently for each equation the
    coefficients of left and right hand side;
    integer m, n, p, i, k;

    procedure leastsquaressolution(a, x, b, m, n, p, eta, singular);
    value m, n, p, eta; array a, x, b; integer m, n, p; real eta;
    label singular;
    comment The array a[1:m,1:n] contains the given matrix of an
    overdetermined system of m linear equations in  n unknowns (m
    ≥ n). For the p right hand sides given as the columns of the
    array b[1:m,1:p], the  least squares solutions are computed
    and stored as the columns of the array x[1:n,1:p]. If rank(a)<n
    then the problem is left unsolved and the emergency exit
    singular is used.  In either case a and b are left intact. Eta
    is the relative machine precision;
    begin

      real procedure innerproduct(i, m, n, a, b, c); value m, n, c;
      real a, b, c; integer i, m, n;
      begin for i:= m step 1 until n do c:= c + a × b;
          innerproduct:= c
      end innerproduct;


      procedure decompose(m, n, qr, alpha, pivot, singular);
      value m, n; integer m, n; array qr, alpha; integer array pivot;
      label singular;
      begin integer i, j, jbar, k;
          real beta, sigma, alphak, qrkk;
          array y, sum[1:n];
          for j:= 1 step 1 until n do
          begin comment j-th column sum;
              sum[j]:= innerproduct(i, 1, m, qr[i,j], qr[i,j], 0);
              pivot[j]:= j
          end j-th column sum;
          for k:= 1 step 1 until n do
          begin sigma:= sum[k]; jbar:= k;
              for j:= k + 1 step 1 until n do if sigma < sum[j]
              then
              begin sigma:= sum[j]; jbar:= j end;
              if jbar ≠ k then
              begin comment column interchange;
                  i:= pivot[k]; pivot[k]:= pivot[jbar];
                  pivot[jbar]:= i; sum[jbar]:= sum[k];
                  sum[k]:= sigma;
                  for i:= 1 step 1 until m do
                  begin sigma:= qr[i,k]; qr[i,k]:= qr[i,jbar];
                      qr[i,jbar]:= sigma
                  end
              end;
```

```
        sigma:= innerproduct(i, k, m, qr[i,k], qr[i,k], 0);
        if sigma = 0 then goto singular; qrkk:= qr[k,k];
        alphak:= alpha[k]:= if qrkk < 0 then sqrt(sigma)
        else - sqrt(sigma);
        beta:= 1 / (sigma - qrkk × alphak);
        qr[k,k]:= qrkk - alphak;
        for j:= k + 1 step 1 until n do y[j]:= beta ×
        innerproduct(i, k, m, qr[i,k], qr[i,j], 0);
        for j:= k + 1 step 1 until n do
        begin for i:= k step 1 until m do qr[i,j]:= qr[i,j]
            - qr[i,k] × y[j]; sum[j]:= sum[j] - qr[k,j] ↑ 2
        end
    end
end;


procedure solve(m, n, qr, alpha, pivot, r, y); value m, n;
integer m, n; array qr, alpha, r, y; integer array pivot;
begin integer i, j;
    real gamma;
    array z[1:n];
    for j:= 1 step 1 until n do
    begin gamma:= innerproduct(i, j, m, qr[i,j], r[i], 0) /
        (alpha[j] × qr[j,j]);
        for i:= j step 1 until m do r[i]:= r[i] + gamma ×
        qr[i,j]
    end j;
    z[n]:= r[n] / alpha[n];
    for i:= n - 1 step - 1 until 1 do z[i]:= -
    innerproduct(j, i + 1, n, qr[i,j], z[j], - r[i]) /
    alpha[i];
    for i:= 1 step 1 until n do y[pivot[i]]:= z[i]
end solve;

integer i, j, k;
real normy0, norme0, norme1, eta2;
array qr[1:m,1:n], alpha, e, y[1:n], r[1:m];
integer array pivot[1:n];
for j:= 1 step 1 until n do
for i:= 1 step 1 until m do qr[i,j]:= a[i,j];
decompose(m, n, qr, alpha, pivot, singular); eta2:= eta ↑ 2;
for k:= 1 step 1 until p do
begin for i:= 1 step 1 until m do r[i]:= b[i,k];
    solve(m, n, qr, alpha, pivot, r, y);
    for i:= 1 step 1 until m do r[i]:= - innerproduct(j, 1,
    n, a[i,j], y[j], - b[i,k]);
    solve(m, n, qr, alpha, pivot, r, e); normy0:= norme1:= 0;
    for i:= 1 step 1 until n do
    begin normy0:= normy0 + y[i] ↑ 2;
        norme1:= norme1 + e[i] ↑ 2
    end i;
    if norme1 > 0.0625 × normy0 then goto singular;
improve: for i:= 1 step 1 until n do y[i]:= y[i] + e[i];
    if norme1 < eta2 × normy0 then goto store;
    for i:= 1 step 1 until m do r[i]:= - innerproduct(j, 1,
```

```
          , n, a[i,j], y[j], - b[i,k]);
            solve(m, n, qr, alpha, pivot, r, e); norme0:= norme1;
            norme1:= 0;
            for i:= 1 step 1 until n do norme1:= norme1 + e[i] ↑ 2;
            if norme1 ≤ 0.0625 × norme0 then goto improve;
        store: for i:= 1 step 1 until n do x[i,k]:= y[i]
        end
      end least squares solution;
      m:= read; n:= read; p:= read;
      begin array a[1:m,1:n], b[1:m,1:p], x[1:n,1:p];
        for i:= 1 step 1 until m do
        begin for k:= 1 step 1 until n do a[i,k]:= read;
            for k:= 1 step 1 until p do b[i,k]:= read;
        end;
        leastsquaressolution(a, x, b, m, n, p, 10 - 12, singular);
        PRINTTEXT(‡ least squares solution vectors ‡); NLCR;
        for k:= 1 step 1 until n do
        begin NLCR;
            for i:= 1 step 1 until p do print(x[k,i]); NLCR
        end;
        goto EINDE;
      singular: PRINTTEXT(‡singular‡);
      EINDE:
      end
end
```

Data for program least squares solution

```
6     5       2
    36      -630       3360      -7560       7560       463       2766
  -630     14700     -88200     211680    -220500    -13860     -82950
  3360    -88200     564480   -1411200    1512000     97020     580440
 -7560    211680   -1411200    3628800   -3969000   -258720   -1547280
  7560   -220500    1512000   -3969000    4410000    291060    1740060
 -2772     83160    -582120    1552320   -1746360   -116424    -695772
```

least squares solution vectors

```
+.1000000097023₁₀+1    +.1000000343724₁₀+1
+.5000000358341₁₀-0    +.1000000148986₁₀+1
+.3333333495129₁₀-0    +.1000000072416₁₀+1
+.2500000073055₁₀-0    +.1000000034146₁₀+1
+.2000000026530₁₀-0    +.1000000012762₁₀+1
```

2.5 Harmonische analyse (continue geval)

Harmonische analyse (ook Fourier-analyse geheten) is een kleinste-kwadratenbenadering speciaal voor periodieke functies.
Laat gegeven zijn een functie f met periode 2π, d.w.z.

2.5.0 $\qquad f(x + 2\pi) = f(x).$

We trachten f te schrijven als een som van speciale functies met periode 2π, aldus

2.5.1 $$f(x) \approx \sum_{j=0}^{n} a_j \cos jx + \sum_{j=1}^{n} b_j \sin jx.$$

Het residu in het punt x is gedefinieerd als

2.5.2 $$r(x) = f(x) - \sum_{j=0}^{n} a_j \cos jx - \sum_{j=1}^{n} b_j \sin jx.$$

We beschouwen eerst het continue geval.
Hierbij is de te minimaliseren grootheid

2.5.3 $$R = \int_0^{2\pi} (r(x))^2 \, dx$$

Nul stellen der partiële afgeleiden geeft de relaties

2.5.4
$$\begin{cases} \dfrac{\partial R}{\partial a_i} = -2 \displaystyle\int_0^{2\pi} \{f(x) - \sum_{j=0}^{n} a_j \cos jx - \sum_{j=1}^{n} b_j \sin jx\} \cos ix \, dx = 0, \quad i = 0(1)n, \\ \dfrac{\partial R}{\partial b_i} = -2 \displaystyle\int_0^{2\pi} \{f(x) - \sum_{j=0}^{n} a_j \cos jx - \sum_{j=1}^{n} b_j \sin jx\} \sin ix \, dx = 0, \quad i = 1(1)n. \end{cases}$$

Deze relaties kunnen aanzienlijk worden vereenvoudigd door de orthogonaliteits-relaties

2.5.5
$$\begin{cases} \displaystyle\int_0^{2\pi} \cos jx \cos ix \, dx = \underline{\text{if}}\ i \neq j\ \underline{\text{then}}\ 0\ \underline{\text{else}}\ \underline{\text{if}}\ i > 0\ \underline{\text{then}}\ \pi\ \underline{\text{else}}\ 2\pi, \\ \displaystyle\int_0^{2\pi} \sin jx \cos ix \, dx = 0, \\ \displaystyle\int_0^{2\pi} \sin jx \sin ix \, dx = \underline{\text{if}}\ i \neq j\ \underline{\text{then}}\ 0\ \underline{\text{else}}\ \pi\ (i, j > 0). \end{cases}$$

Deze relaties worden bewezen als volgt (vgl. pag. 276).
Eerst schrijven we de integraal als de som van twee cosinussen of sinussen, waarna de integraal gemakkelijk te vinden is, bijvoorbeeld

$$\int_0^{2\pi} \cos jx \cos ix \, dx = \tfrac{1}{2} \int_0^{2\pi} \cos (j+i)x \, dx + \tfrac{1}{2} \int_0^{2\pi} \cos (j-i)x \, dx.$$

Noemen we j+i resp. j-i even α, dan vinden we voor $\alpha \neq 0$

$$\int_0^{2\pi} \cos \alpha x \, dx = \frac{1}{\alpha} \sin \alpha x \Big|_0^{2\pi} = 0.$$

Voor $j \neq i$ zijn dus beide integralen 0, voor $j = i \neq 0$ levert de een 0, de ander π en voor $j = i = 0$ leveren beide π en is het antwoord dus 2π.
Met behulp van (2.5.5) vindt men de oplossing van (2.5.4):

2.5.6
$$\begin{cases} a_0 = \frac{1}{2} \int_0^{2\pi} f(x)dx \\ a_i = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos ix \, dx \\ b_i = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin ix \, dx \end{cases} i > 0.$$

Deze formules hangen blijkbaar niet van n af.
Laten we n naar oneindig gaan, dat convergeert het rechterlid van (2.5.1) in veel gevallen naar f(x), in formule

2.5.7
$$f(x) = \sum_{j=0}^{\infty} a_j \cos jx + \sum_{j=1}^{\infty} b_j \sin jx$$

Dit heet de Fourier-reeks van f en de coëfficiënten a_j en b_j heten de Fourier-coëfficiënten van f.
Nemen we nu een stuk van de Fourier-reeks, dan krijgen we blijkbaar een benadering van f in de zin der kleinste kwadraten.

Voorbeeld De functie f met periode 2π gedefinieerd door

$$f(x) = \text{sign}(x) \quad , -\pi < x < +\pi, \quad f(\pi) = 0.$$

In plaats van $\int_0^{2\pi}$ mogen we natuurlijk ook $\int_{-\pi}^{+\pi}$ nemen.

Dan vinden we $a_i = 0$ voor alle $i \geq 0$, want f is een oneven functie.
Verder hebben we

$$b_i = \frac{1}{\pi} \int_{-\pi}^{+\pi} f(x) \sin ix \, dx = \frac{2}{\pi} \int_0^{\pi} \sin ix \, dx = \frac{4}{\pi i}$$

Dus

$$\frac{\pi}{4} f(x) = \sum_{i=1}^{\infty} \frac{\sin ix}{i} .$$

In het bijzonder is dus voor $0 < x < \pi$ deze som gelijk aan $\frac{\pi}{4}$.
Voor $x = \frac{\pi}{2}$ ontstaat de wel-bekende reeks

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots$$

2.6 Harmonische analyse (discrete equidistante punten)

We beschouwen nu een functie f met periode 2π, die gegeven is op de equidistante punten

2.6.0 $$x_k = \frac{2\pi k}{m} ,$$

waarbij m een gegeven positief geheel getal is.
Wegens de periodiciteit van f geldt $f(x_{k+m}) = f(x_k)$, zodat we alleen de punten x_k voor $k = 0(1)m-1$ hoeven te beschouwen.
We trachten weer f(x) te schrijven in de gedaante (2.5.1), waarbij nu n veel kleiner dan m moet zijn (om precies te zijn, n mag niet groter zijn dan m/2).
Schrijven we $f_k = f(x_k)$ en $r_k = r(x_k)$ (zie 2.5.2), dan moeten we nu minimaliseren:

2.6.1 $$R = \sum_{k=0}^{m-1} r_k^2 .$$

In plaats van (2.5.4) komen nu de relaties

$$2.6.2\begin{cases}\dfrac{\partial R}{\partial a_i} = -2\sum_{k=0}^{m-1}\{f_k - \sum_{j=0}^{n} a_j \cos jx_k - \sum_{j=1}^{n} b_j \sin jx_k\}\cos ix_k = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 0(1)n, \\ \dfrac{\partial R}{\partial b_i} = -2\sum_{k=0}^{m-1}\{f_k - \sum_{j=0}^{n} a_j \cos jx_k - \sum_{j=1}^{n} b_j \sin jx_k\}\sin ix_k = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1(1)n.\end{cases}$$

Deze laten zich weer aanzienlijk vereenvoudigen door orthogonaliteits-relaties, die nu luiden:

$$2.6.3\begin{cases}\sum_{k=0}^{m-1} \cos jx_k \cos ix_k = \underline{\text{if}}\ i \neq j\ \underline{\text{then}}\ 0\ \underline{\text{else if}}\ i \neq 0\ \underline{\text{then}}\ m/2 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \underline{\text{else}}\ m. \\ \sum_{k=0}^{m-1} \sin jx_k \cos ix_k = 0 \\ \sum_{k=0}^{m-1} \sin jx_k \sin ix_k = \underline{\text{if}}\ i \neq j\ \underline{\text{then}}\ 0\ \underline{\text{else}}\ m/2\ (i, j > 0),\end{cases}$$

waarbij echter de restrictie geldt, dat i en j kleiner dan m/2 moeten zijn. De afleiding van deze relaties geschiedt ongeveer op dezelfde wijze als op pag. 272-273.

Bijvoorbeeld

$$\sum_{k=0}^{m-1} \cos jx_k \cos ix_k = \tfrac{1}{2}\sum_{k=0}^{m-1} \cos (j+i)x_k + \tfrac{1}{2}\sum_{k=0}^{m-1} \cos (j-i)x_k.$$

Noemen we j+i resp. j-i even α, dan krijgen we voor $\alpha \neq 0$:

$$\sum_{k=0}^{m-1} \cos \alpha x_k = \mathrm{Re} \sum_{k=0}^{m-1} \exp (i\alpha x_k).$$

Deze som is een meetkundige reeks met reden $r = \exp (\frac{2\pi}{m})$.

De som is dus gelijk aan $\dfrac{\exp(i\alpha x_m)-1}{r-1}$.

Dit is gelijk aan 0, want $\exp(i\alpha x_m) = \exp (i\alpha \frac{2\pi m}{m}) = 1$.

Hieruit volgt gemakkelijk de eerste formule van (2.6.3).

Op analoge wijze bewijst men de tweede en derde formule.

Met behulp van (2.6.3) vinden we nu de oplossing van (2.6.2):

$$2.6.4 \quad \left.\begin{aligned} a_0 &= \frac{1}{m}\sum_{k=0}^{m-1} f_k \\ a_i &= \frac{2}{m}\sum_{k=0}^{m-1} f_k \cos(ix_k) \\ b_i &= \frac{2}{m}\sum_{k=0}^{m-1} f_k \sin(ix_k) \end{aligned}\right\} \quad 0 < i \leq n < \frac{m}{2}$$

Deze formules hangen wederom niet van n af.
Het is duidelijk, dat het aantal coëfficiënten a_i en b_i, dat bepaald kan worden, maximaal gelijk is aan het aantal basispunten m. Het heeft dus geen zin n groter dan m/2 te nemen. Voor $n < m/2$ gelden bovenstaande formules. Voor m even en $n = m/2$ treedt echter een uitzondering op. Dan is enerzijds $\sin(nx_k) = 0$ voor alle x_k, zodat deze functie en dus ook de coëfficiënt b_n niet meedoen. Anderzijds geldt $\cos(nx_k) = \cos(\pi k) = (-1)^k$, zodat in afwijking van (2.6.3 & 4) geldt $\sum_{k=0}^{m-1} \cos^2(nx_k) = m$ en dus $a_n = \frac{1}{m}\sum_{k=0}^{m-1} (-1)^k f_k$.

Opmerking

Vervangt men in de formules (2.5.6) de integralen door hun benadering volgens de trapezium-regel op de basispunten $x_0, x_1, \ldots, x_m$, dan ontstaan de formules (2.6.4) (want wegens de periodiciteit van f geldt $f_m = f_0$). Men zou nu kunnen trachten slim te zijn en de trapeziumregel door een hogere orde formule (b.v. Simpson) kunnen vervangen. Deze zijn echter voor periodieke functies niet beter dan de trapeziumregel, zoals blijkt uit de formule van Gauss (3.3.1 pag. 119). Immers voor ondergrens x_0 en bovengrens x_m vallen wegens de periodiciteit van f alle hogere orde termen tegen elkaar weg ($\mu\delta_m = \mu\delta_0$, $\mu\delta_m^3 = \mu\delta_0^3$, enz.), zodat in de formule van Gauss alleen het trapeziumstuk $\mu\delta^{-1} f_m - \mu\delta^{-1} f_0$ overblijft.
Dit wil evenwel niet zeggen, dat de trapeziumregel voor periodieke functies exact is. Er blijft altijd nog een restterm over, die afhangt van het aantal basispunten.

Voorbeeld (zie Hildebrand [2], pag. 376).

Zij f een functie met periode 2π, waarvan de volgende tabel gegeven is (het argument θ noteren we hier voor het gemak in graden).

θ	$f(\theta)$	θ	$f(\theta)$	θ	$f(\theta)$
0	1.21	120	1.34	240	1.05
30	1.32	150	1.18	270	1.10
60	1.46	180	1.07	300	1.14
90	1.40	210	1.01	330	1.17

We gaan hieruit de numerieke Fourier-coëfficiënten a_i en b_i berekenen volgens (2.6.4), waarbij nu m = 12. Aangezien de cosinus een even functie is, is het handig eerst de waarden f_k en f_{m-k} bij elkaar te tellen. Daarnaast schrijven we de waarden cos $i\theta_k$ voor i = 1(1)n. Kiezen we n = 3, dan krijgen we aldus het volgende schema, waarin $\alpha = \frac{1}{2}\sqrt{3} \approx 0.8660$.

		cos θ_k	cos $2\theta_k$	cos $3\theta_k$
f_0	= 1.21	1	1	1
$f_1 + f_{11}$	= 2.49	α	.5	0
$f_2 + f_{10}$	= 2.60	.5	-.5	-1
$f_3 + f_9$	= 2.50	0	-1	0
$f_4 + f_8$	= 2.39	-.5	-.5	1
$f_5 + f_7$	= 2.19	$-\alpha$	.5	0
f_6	= 1.07	-1	1	-1

Om a_0 te krijgen moeten we de eerste kolom sommeren en door 12 delen; de andere a_i krijgen we door het scalair product van de eerste kolom en een der andere kolommen te delen door 6. We vinden dan

$$a_0 = 1.204 \quad , \quad a_1 = 0.084 \quad , \quad a_2 = -0.062 \quad , \quad a_3 = -0.012.$$

Voor het berekenen van de coëfficiënten b_i bepalen we eerst de waarden $f_k - f_{m-k}$ en de waarden sin $i\theta_k$, die we aldus in schema zetten.

	$\sin \theta_k$	$\sin 2\theta_k$	$\sin 3\theta_k$
$f_1 - f_{11} = .15$	.5	α	1
$f_2 - f_{10} = .32$	α	α	0
$f_3 - f_9 = .30$	1	0	-1
$f_4 - f_8 = .29$	α	$-\alpha$	0
$f_5 - f_7 = .17$	.5	$-\alpha$	1

Hieruit vinden we

$b_1 = 0.165$, $b_2 = 0.002$, $b_3 = 0.003$.

Men kan zo doorgaan tot n = 6. Om a_6 te krijgen moet men evenwel niet door 6 maar door 12 delen (zie pag. 312).

Opgaven

157) Schrijf een programma, dat met behulp van de procedure "least squares solution" (pag. 305) een kleinste-kwadraten polynoom-benadering van de graad n bepaalt van een functie f(x), die gegeven is op m equidistante basispunten x_k (k = 0(1)m-1), waarbij
a) $f(x) = \sin \frac{\pi}{2} x$, $x_0 = -1$, $x_{m-1} = 1$;
b) $f(x) = e^x$, $x_0 = 0$, $x_{m-1} = 1$.

Laat het programma uitvoeren voor n = 3 en m = 10, 20 en 40 en vergelijk de resultaten met elkaar en met opgave 151.

158) Geef een volledig bewijs van de formules (2.5.5) en (2.6.3).

159) Bepaal de Fourier-reeks van de zaagtand-functie f met periode 2π, gedefiniëerd door $f(x) = x$ voor $0 \leq x < 2\pi$.

160) (uit Hildebrand [2], pag. 415)
Een functie f met periode 2π is gemeten voor de argument-waarden $-\pi(\frac{\pi}{6})\pi$; de opeenvolgende gemeten waarden zijn

voor $\theta < 0$:

2.077	0.278	-1.014	-0.716	0.051	0.277

voor $\theta \geq 0$:

1.015	3.031	4.759	4.680	3.689	3.032	2.077

Bepaal hieruit de coëfficiënten voor de harmonische analyse van f.

161) Tabelleer $f(x) = \frac{4 - 2 \cos x}{5 - 4 \cos x}$ in 6 decimalen voor $x = 0(\frac{\pi}{4})\frac{7\pi}{4}$ en bereken hieruit de numerieke Fourier-coëfficiënten.
Doe de berekening eveneens met gehalveerd tabelinterval $\frac{\pi}{8}$.

2.7 Kettingbreuken

Een ander type approximaties kan worden verkregen door middel van kettingbreuken (in het Engels "continued fractions" geheten). Hierover bestaat een uitgebreide theorie (bv. H.S. Wall, Continued fractions, New York, 1948). We geven slechts een zeer beknopte beschouwing met enige voorbeelden en toepassingen. Allereerst de

Definitie. Een kettingbreuk heeft de gedaante

2.7.0
$$k_n = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \ddots + \cfrac{a_{n-1}}{b_{n-1} + \cfrac{a_n}{b_n}}}}$$

Om wat ruimte te sparen, gebruikt men vaak een compactere notatie, bv.

2.7.1
$$k_n = b_0 + \frac{a_1}{|b_1} + \frac{a_2}{|b_2} + \ldots + \frac{a_{n-1}}{|b_{n-1}} + \frac{a_n}{|b_n},$$

wat niets anders is, dan een andere schrijfwijze voor (2.7.0).
Werken we (2.7.0) uit, dan krijgen we tenslotte een quotiënt

2.7.2
$$k_n = \frac{A_n}{B_n},$$

dat voor enige lage waarden van n er aldus uitziet.

2.7.3 $A_0 = b_0$, $B_0 = 1$, want $k_0 = b_0$.

$$k_1 = b_0 + \frac{a_1}{b_1} = \frac{b_1 b_0 + a_1}{b_1}, \text{ dus}$$

2.7.4 $A_1 = b_1 b_0 + a_1$, $B_1 = b_1$.

$$k_2 = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2}} = b_0 + \frac{a_1 b_2}{b_1 b_2 + a_2} = \frac{(b_0 b_1 + a_1) b_2 + b_0 a_2}{b_1 b_2 + a_2}, \text{ dus}$$

$$2.7.5 \quad \begin{cases} A_2 = (b_0b_1 + a_1)b_2 + b_0a_2 = A_1b_2 + A_0a_2, \\ B_2 = b_1b_2 + a_2 = B_1b_2 + B_0a_2. \end{cases}$$

Dit suggereert de volgende recursie-formule voor A_n en B_n:

$$2.7.6 \quad \left\{ \begin{array}{l} A_n = A_{n-1}b_n + A_{n-2}a_n \\ B_n = B_{n-1}b_n + B_{n-2}a_n \end{array} \right\}, \quad n \geq 2.$$

Deze recursie-formule leiden we af door inductie naar n.
Volgens (2.7.5) geldt de formule voor n = 2. Als het voor zekere n geldt, dan geldt het ook voor n+1, wat we bewijzen aldus.
Voor k_n mogen we de inductie-veronderstelling (2.7.6) toepassen.
Welnu, k_{n+1} ontstaat uit k_n door b_n te vervangen door $b_n + \frac{a_{n+1}}{b_{n+1}}$, zodat we krijgen

$$k_{n+1} = \frac{A_{n-1}(b_n + \frac{a_{n+1}}{b_{n+1}}) + A_{n-2}a_n}{B_{n-1}(b_n + \frac{a_{n+1}}{b_{n+1}}) + B_{n-2}a_n} = \frac{A_n + A_{n-1}\frac{a_{n+1}}{b_{n+1}}}{B_n + B_{n-1}\frac{a_{n+1}}{b_{n+1}}} = \frac{A_nb_{n+1} + A_{n-1}a_{n+1}}{B_nb_{n+1} + B_{n-1}a_{n+1}}.$$

Dus $k_{n+1} = \frac{A_{n+1}}{B_{n+1}}$, waarbij
$A_{n+1} = A_nb_{n+1} + A_{n-1}a_{n+1}$, $B_{n+1} = B_nb_{n+1} + B_{n-1}a_{n+1}$.

M.a.w. (2.7.6) geldt ook voor n+1, waarmee het bewijs klaar is.
De theorie der kettingbreuken levert rationale benaderingen van functies.

Voorbeelden

$$\sqrt{1+x} = 1 + \frac{x}{1 + \sqrt{1+x}}.$$

Herhaald toepassen van deze formule levert:

$$\sqrt{1+x} = 1 + \frac{x}{2 + \frac{x}{2 + \cdots \frac{+ x}{2 + w}}},$$

waarbij $w = \sqrt{1 + x} - 1$. Voor $w = 0$ hebben we een benadering k_n, die lineair naar $\sqrt{1 + x}$ convergeert voor $x > -1$.
Voor $x = 2$ en enige lage waarden van n vinden we

$k_0 = 1$
$k_1 = 1.5$
$k_2 = 1 + 1/2.5 = 1.4$
$k_3 = 1 + \frac{1}{2.4} = 1\frac{5}{12} \approx 1.41667$
$k_4 = 1 + \frac{12}{29} \approx 1.41379.$

We vermelden nog enige kettingbreuk-ontwikkelingen.

$$e^x = 1 + \frac{x|}{|1} - \frac{x|}{|2} + \frac{x|}{|3} - \frac{x|}{|2} + \frac{x|}{|5} - \frac{x|}{|2} + \frac{x|}{|7} - \ldots ,$$

$$\arctan(x) = \frac{x|}{|1} + \frac{x^2|}{|3} + \frac{4x^2|}{|5} + \frac{9x^2|}{|7} + \frac{16x^2|}{|9} + \ldots .$$

Kettingbreuk-ontwikkeling van rationale functies

We beschouwen een rationale functie, d.w.z. een functie van de gedaante

2.7.7 $$f(x) = P_m(x)/Q_n(x),$$

waarbij teller en noemer polynomen van de graad m resp. n zijn.
Als $m \geq n$ kan door de bekende staartdeling een polynoom van de graad m-n als geheel deel worden afgesplitst, waardoor we krijgen

2.7.8 $$f(x) = C_{m-n}(x) + P_{n-1}(x)/Q_n(x).$$

Hierbij is C_{m-n} een polynoom van de graad m-n en P_{n-1} is de rest bij deling van $P_m(x)$ door $Q_n(x)$, dus een polynoom van de graad $\leq$ n-1.
Is P_{n-1} precies van de graad n-1, dan kunnen we f(x) schrijven als

$$f(x) = C_{m-n}(x) + \frac{a_1}{a_1\, Q_n(x)/P_{n-1}(x)} ,$$

waarbij a_1 zo gekozen wordt, dat de voorste coëfficiënt van $\frac{1}{a_1} Q_n(x)$ en van $P_{n-1}(x)$ gelijk zijn. De noemer behandelen we dan op dezelfde wijze als boven, waardoor we krijgen

$$f(x) = C_{m-n}(x) + \cfrac{a_1}{x + b_1 + \cfrac{Q_{n-2}(x)}{P_{n-1}(x)}} ,$$

waarbij Q_{n-2} van de graad $\leq n-2$ is.
Zo doorgaande krijgen we na eindig veel stappen:

2.7.9
$$f(x) = C_{m-n}(x) + \cfrac{a_1}{x + b_1 + \cfrac{a_2}{x + b_2 + \ddots + \cfrac{a_{n-1}}{x + b_{n-1} + \cfrac{a_n}{x + b_n}}}}$$

Is de graad van de optredende polynomen P of Q toevallig lager, dan men zou verwachten, dan krijgen we een iets afwijkende kettingbreuk. In elk geval vinden we een eindige kettingbreuk-ontwikkeling voor f(x), want in elke stap treedt een graad-verlaging op, zodat het proces na eindig veel stappen moet afbreken.

Vergelijken we het aantal bewerkingen nodig voor de berekening van f(x) volgens de verschillende formules, dan vinden we, aangezien een polynoom van de graad n kan worden berekend met behulp van n vermenigvuldigingen en n optellingen, het volgende.

formule	optellingen	vermenigvuldigingen	delingen
2.7.7	m+n	m+n	1
2.7.8	m+n	m+n-1	1
2.7.9	m+n	m-n	n

Gebruiken we dus (2.7.9) in plaats van (2.7.8), dan sparen we 2n-1 vermenigvuldigingen ten koste van n-1 extra delingen. Voor een machine, waar de deling sneller gaat dan twee vermenigvuldigingen betekent dit dus winst.

Toepassing

Men kan een functie over een zeker interval benaderen door een rationale functie f(x) van de gedaante (2.7.7), bijvoorbeeld een beste benadering in de zin van Chebyshev (voor het geval n=0, d.w.z. polynoom-benadering, hebben we dit behandeld, zie 1.6 pag. 280-286). Vervolgens werkt men (2.7.7) op de boven aangegeven wijze om tot een kettingbreuk.

Dit leidt vaak, voor een bepaalde vereiste precisie, tot een benaderings-formule die sneller is dan een beste polynoom-benadering in de zin van Chebyshev.

Voorbeelden

$$\frac{3x^3 - 20x^2 - 236x - 444}{x^2 - 10x - 43} = 3x + 10 - \frac{7x + 14}{x^2 - 10x - 43} =$$

$$= 3x + 10 - \cfrac{7}{x - 12 - \cfrac{19}{x + 2}}$$

C. Hastings [11] geeft op pag. 188 de volgende rationale benadering

$$x\, e^x \int_x^\infty \frac{e^{-t}}{t}\, dt \approx f^*(x) = \frac{0.250621 + 2.334733x + x^2}{1.681534 + 3.330657\, x + x^2},$$

die voor $1 \leq x < \infty$ een relatieve precisie levert van 4 decimalen. Hiervoor vinden we

$$f^*(x) = 1 - \frac{0.995924x + 1.430913}{x^2 + 3.330657x + 1.681534} =$$

$$= 1 - \cfrac{0.995924}{x + 1.89388773 - \cfrac{1.03954569}{x + 1.43676927}}$$

Opgave

162) Volgens [11] pag. 192 heeft de functie y = f(x) gedefinieerd door

$$x = \frac{1}{2\pi} \int_y^\infty e^{-\frac{1}{2}t^2} \, dt$$

de volgende benadering, die voor $0 < x \leq 0.5$ een precisie van 3 decimalen oplevert:

$$f^*(x) = p - \left\{ \frac{a_0 + a_1 p + a_2 p^2}{1 + b_1 p + b_2 p^2 + b_3 p^3} \right\}$$

waarbij $p = \sqrt{\ln \frac{1}{x^2}}$ en verder:

$a_0 = 2.515517$ $\qquad$ $b_1 = 1.432788$

$a_1 = 0.802853$ $\qquad$ $b_2 = 0.189269$

$a_2 = 0.010328$ $\qquad$ $b_3 = 0.001308$

Werk de rationale uitdrukking om tot een kettingbreuk.
Bereken vervolgens met behulp hiervan $f^*(x)$ voor 3 waarden van x en vergelijk de resultaten met een tabel.

Hoofdstuk 9

Quadratuur, sommeren van reeksen en bepaling van limieten.

1. Speciale Quadratuur-formules

1.1. Inleiding.

We zullen hier enige formules voor het berekenen van bepaalde integralen behandelen, die in hoofdstuk 3 (pag. 95-120) onbesproken zijn gebleven, doch in een besluit op pag. 119 zijn aangekondigd. Men gebruikt graag het woord "quadratuur" voor bepaalde integratie, ter onderscheiding van onbepaalde integratie en het integreren van differentiaal-vergelijkingen.
Voor de goede orde vermelden we hier de volgende

Definitie De orde van een quadratuur-formule is het kleinste natuurlijk getal m, waarbij een polynoom van de graad m te vinden is, waarvoor de formule niet exact is.

Deze definitie stemt overeen met de definitie van "orde" gegeven op pag. 101 voor de Newton-Cotes formules (Ga dit na).

1.2. Optimale quadratuur-formules van Gauss

Laat $w(x)$ een gegeven functie zijn, die gedefinieerd en positief is op een gegeven interval a, b.
We zoeken nu formules van de vorm

$$1.2.0 \qquad \int_a^b w(x)\, f(x)\, dx = A_0 f(x_0) + A_1 f(x_1) + \ldots + A_{n-1} f(x_{n-1}) + E_n,$$

waarbij $A_0, \ldots A_{n-1}$ bepaalde, van n afhankelijke, constanten zijn en E_n de restterm aanduidt.
We hebben reeds formules van dit type gezien, voor $w(x) \equiv 1$, n.l. de formules van Newton-Cotes en van Steffenson (zie pag. 100-112)

Voor deze formules waren de basispunten equidistant gekozen. Nu gaan we de coëfficiënten en de basispunten zodanig kiezen, dat de orde van de formule zo hoog mogelijk wordt. We spreken dan van de (optimale) quadratuur-formules van Gauss. (Niet te verwarren met de quadratuur-formule van Gauss, uitgedrukt in centrale differenties, zie 3.3.1. pag. 119).

Beschouwen we eerst, voor een willekeurig stel basispunten $x_0, \ldots, x_{n-1}$, de Lagrange-polynoombenadering (vgl. 6.3 pag. 40):

1.2.1 $$f_n^*(x) = \sum_{k=0}^{n-1} \mathcal{L}_k^n(x)\, f(x_k)$$

Vermenigvuldigen met w(x) en integreren levert:

1.2.2 $$\int_a^b w(x)\, f_n^*(x)\, dx = \sum_{k=0}^{n-1} \left[\int_a^b w(x)\, \mathcal{L}_k^n(x)\, dx \right] f(x_k).$$

Stellen we

1.2.3 $$A_k = \int_a^b w(x)\, \mathcal{L}_k^n(x)\, dx, \qquad k = 0(1)\ n-1,$$

dan gaat (1.2.2) over in

1.2.4 $$\int_a^b w(x)\, f_n^*(x)\, dx = \sum_{k=0}^{n-1} A_k f(x_k).$$

Als f een polynoom van de graad kleiner dan n is, dan is voor een willekeurig stel basispunten de Lagrange-benadering (1.2.1) exact en dus ook de integratieformule (1.2.4.). De maximale orde hiervan is dus minstens n.

Om een hogere orde te bereiken moeten we de basispunten geschikt kiezen. Zij nu f(x) een polynoom van de graad $n + t (t \geq 0)$, dan is $f(x) - f_n^*(x)$ een polynoom van de graad $n + t$, die gelijk aan nul is in de basispunten $x_0, \ldots, x_{n-1}$.

We kunnen dus schrijven

1.2.5 $$f(x) = f_n^*(x) + \pi_n(x)\, g(x),$$

waarbij (vgl. pag. 39)

1.2.6 $$\pi_n(x) = (x-x_0) \ldots (x-x_{n-1})$$

en g(x) een polynoom van de graad t is.

Vervangen we nu in (1.2.0) f(x) door $f(x)-f_n^*(x)$, dan blijft in het rechterlid alleen E_n over, omdat $f(x)- f_n^*(x) = 0$ voor alle basispunten. Willen we dat de formule exact is, dan zal ook E_n gelijk aan nul moeten zijn, zodat we krijgen

1.2.7 $$\int_a^b w(x)\,(f(x) - f_n^*(x))\,dx = \int_a^b w(x)\,\pi_n(x)\,g(x)\,dx = 0.$$

Dit lukt voor alle polynomen g(x) van de graad kleiner dan n (en dus voor alle polynomen f(x) van de graad kleiner dan 2 n), als $\pi_n(x)$ het n-de graadspolynoom (met voorste coëfficiënt 1)is, dat voldoet aan de orthogonaliteits-relatie

1.2.8 $$\int_a^b w(x)\,\pi_n(x)\,\pi_j(x)\,dx = 0, \qquad j \neq n$$

Als de graad van f gelijk is aan 2n (en dus g de graad n heeft), geldt (1.2.7) niet meer. Immers kiezen we nu $g(x) = \pi_n(x)$, dan geldt wegens w(x) > 0:

$$\int_a^b w(x)\,(\pi_n(x))^2\,dx > 0.$$

De hoogste bereikbare orde van formule (1.2.0) is dus 2n. Deze wordt bereikt, als de basispunten $x_0, \ldots, x_{n-1}$ worden gekozen gelijk aan de nulpunten van het n-de graads orthogonale polynoom $\pi_n(x)$, dat voldoet aan de orthogonaliteits-relatie (1.2.8).

(Deze relatie bepaalt de polynomen eenduidig op een constante factor na, die evenwel voor de nulpunten niet ter zake doet). De coëfficiënten $A_0, \ldots, A_{n-1}$ moeten hierbij worden bepaald volgens (1.2.3).

Eigenschappen van de optimale integratie-formules van Gauss

1) De nulpunten $x_0, \ldots, x_{n-1}$ van het polynoom $\pi_n(x)$ zijn enkelvoudig en liggen tussen a en b.

2) De coëfficiënten $A_0, \ldots, A_{n-1}$ zijn positief.

3) Als f een continue functie is, dan geldt voor de restterm E_n:

$$\lim_{n\to\infty} E_n = 0$$

(Dit geldt daarentegen niet voor de restterm van de n-punts Newton-Cotes formule).

4) Als de 2n-de afgeleide van f bestaat op het interval $[a, b]$, dan geldt:

$$E_n = \frac{f^{(2n)}(\xi)}{(2n)!} \int_a^b w(x)\,(\pi_n(x))^2\,dx,$$

waarbij ξ ergens tussen a en b ligt.

We gaan nu enige belangrijke bijzondere gevallen van de optimale Gauss-formules behandelen.

1.3 Formule van Gauss-Legendre

Door de keuzen $w(x) \equiv 1$, $a = -1$, $b = +1$, gaat formule (1.2.0) over in de integratie-formule van Gauss-Legendre:

1.3.0 $$\int_{-1}^{+1} f(x)dx \approx A_0\, f(x_0) + A_1\, f(x_1) + \ldots + A_{n-1}\, f(x_{n-1})$$

Hierbij zijn de basispunten $x_0, \ldots, x_{n-1}$ de nulpunten van het n-de graads polynoom van Legendre (zie pag. 292) en de coëfficiënten

voldoen aan (vgl. 1.2.3)

1.3.1 $$A_k = \int_{-1}^{+1} \mathcal{L}_k^n(x)dx, \quad k = 0(1)n-1$$

Bijvoorbeeld voor n = 2 vinden we dat de basispunten x_k de nulpunten zijn van $P_2(x) = \frac{1}{2}(3x^2 - 1)$, dus $x_k = \pm\frac{1}{3}\sqrt{3}$, en de coëfficiënten A_k blijken gelijk aan 1 te zijn.
Men kan de x_k en A_k ook direct vinden door te eisen dat (1.3.0) exact is (d.w.z. $E_n = 0$) voor f(x) = resp. 1, x, ..., x^{2n-1}. Voor n = 2 ontstaat dan het volgende stelsel vergelijkingen.

$$\begin{aligned} A_0 + A_1 &= 2 \\ A_0x_0 + A_1x_1 &= 0 \\ A_0x_0^2 + A_1x_1^2 &= 2/3 \\ A_0x_0^3 + A_1x_1^3 &= 0 \end{aligned}$$

met als oplossing: $x_0 = -\frac{1}{3}\sqrt{3}$, $x_1 = \frac{1}{3}\sqrt{3}$, $A_0 = A_1 = 1$. (Ga dit na).
Hier volgen de basispunten en bijbehorende coëfficiënten voor enige waarden van n.

n	x_k	A_k
2	± 0.57735 02692	1.00000 00000
3	± 0.77459 66692	0.55555 55556
	0.00000 00000	0.88888 88889
4	± 0.86113 63116	0.34785 48451
	± 0.33998 10436	0.65214 51549
5	± 0.90617 98459	0.23692 68851
	± 0.53846 93101	0.47862 86705
	0.00000 00000	0.56888 88889

Opmerking

De formule van Gauss-Legendre is ook bruikbaar, als de eindpunten van het integratie-interval andere waarden hebben, mits we eerst

een passende transformatie uitvoeren. Om precies te zijn, voor willekeurig interval $[a, b]$ en $w(x) \equiv 1$ luidt de formule van Gauss-Legendre

1.3.2 $$\int_a^b f(x)dx \approx \frac{b-a}{2} \{A_0 f(y_0) + \ldots + A_{n-1} f(y_{n-1})\},$$

waarbij $y_k = \frac{b-a}{2} x_k + \frac{b+a}{2}$, $k = 0(1)n-1$,

en x_k en A_k dezelfde als boven zijn.

1.4 Formules van Gauss-Jacobi

Voor deze formules geldt: $w(x) = (1-x)^p(1+x)^q$, $p, q > -1$, en $a = -1$, $b = +1$. Hierbij zijn p en/of q vaak niet geheel. Deze formules zijn speciaal geschikt voor het geval de integrand aan een of beide uiteinden van het integratie-interval een singulariteit heeft.

In het bijzondere geval $p = q = -\frac{1}{2}$, m.a.w. $w(x) = \frac{1}{\sqrt{1-x^2}}$, zijn de basispunten $x_0, \ldots, x_{n-1}$ gelijk aan de nulpunten van het n-de graads polynoom van Chebyshev (zie pag. 267-269, vooral formule (1.2.5) en de orthogonaliteitsrelatie (1.4.3) op pag. 276), dus

$$x_k = \cos\ (k+\tfrac{1}{2})\ \frac{\pi}{n}\ , \quad k = 0(1)n-1.$$

De coëfficiënten, die weer voldoen aan (1.2.3), blijken nu alle gelijk te zijn aan π/n.

Verwerken we dit alles in formule (1.2.0), dan ontstaat de formule van Gauss-Chebyshev

1.4.0 $$\int_{-1}^{+1} \frac{f(x)dx}{\sqrt{1-x^2}} \approx \frac{\pi}{n} \sum_{k=0}^{n-1} f(\cos\ (k+\tfrac{1}{2})\ \frac{\pi}{n}).$$

1.5 Formule van Gauss-Laguerre

Door de keuzen $w(x) = e^{-x}$, $a = 0$, $b = \infty$, gaat formule (1.2.0) over in de integratie-formule van Gauss-Laguerre

1.5.0 $$\int_{-1}^{+1} e^{-x} f(x)dx \approx A_0 f(x_0) + \ldots + A_{n-1} f(x_{n-1})$$

Hierbij zijn de basispunten $x_0, \ldots, x_{n-1}$ gelijk aan de nulpunten van het n-de graads polynoom van Laguerre $L_n(x)$. De polynomen van Laguerre voldoen aan de orthogonaliteits-voorwaarde

1.5.1 $$\int_0^{\infty} e^{-x} L_n(x)\, L_j(x)dx = \underline{\text{if}}\ n \neq j\ \underline{\text{then}}\ 0\ \underline{\text{else}}\ 1.$$

Voor enige lage waarden van n hebben we:

$$L_0(x) \equiv 1$$

$$L_1(x) = -x + 1$$

$$L_2(x) = \tfrac{1}{2}x^2 - 2x + 1$$

$$L_3(x) = -\frac{1}{6}x^3 + \frac{3}{2}x^2 - 3x + 1$$

$$L_4(x) = \frac{1}{24}x^4 - \frac{2}{3}x^3 + 3x^2 - 4x + 1$$

Hier volgen de basispunten en bijbehorende coëfficiënten voor enige n-waarden.

n	x_k	A_k
2	0.58578 64376	0.85355 33906
	3.41421 35624	0.14644 66094
3	0.41577 45568	0.71109 30099
	2.29428 03603	0.27851 77336
	6.28994 50829	0.01038 92565 0
4	0.32254 76896	0.60315 41043
	1.74576 11012	0.35741 86924
	4.53662 02969	0.03888 79085 15
	9.39507 09123	$0.53929\ 47056_{10}{-3}$

1.6 Formule van Gauss-Hermite

De keuze $w(x) = e^{-x^2}$, $a = -\infty$, $b = +\infty$ leidt tot de integratieformule van Gauss-Hermite

1.6.0 $$\int_{-\infty}^{+\infty} e^{-x^2} f(x)dx \approx A_0 f(x_0) + \ldots + A_{n-1} f(x_{n-1}).$$

De basispunten $x_0, \ldots, x_{n-1}$ zijn nu gelijk aan de nulpunten van het n-de graads polynoom van Hermite $H_n(x)$. De polynomen van Hermite voldoen aan de orthogonaliteitsrelatie

1.6.1 $$\int_{-\infty}^{+\infty} e^{-x^2} H_n(x)\, H_j(x)dx = 0 \quad \text{voor} \quad n \neq j.$$

Voor enige lage n-waarden hebben we

$$H_0(x) \equiv 1$$

$$H_1(x) = 2x$$

$$H_2(x) = 4x^2 - 2$$

$$H_3(x) = 8x^3 - 12x$$

$$H_4(x) = 16x^4 - 48x^2 + 12$$

De basispunten en bijbehorende coëfficiënten voor enige n-waarden zijn als volgt.

n	x_k	A_k
2	$\pm$ 0.70710 67812	0.88622 69255
3	0	1.18163 59006
	$\pm$ 1.22474 48714	0.29540 89752
4	$\pm$ 0.52464 76233	0.80491 40900
	$\pm$ 1.65068 01239	0.08131 28354 5

Opgaven

163) Bepaal de basispunten en coëfficiënten voor de 3-punts Gauss-Legendre formule (1.3.0) door oplossing van het stelsel, dat ontstaat als men exactheid van de formule eist voor $f(x) = 1, x, \ldots, x^5$.

164) Bereken $\int_2^8 \frac{dx}{x}$ met de 3- en 5-punts formules van Gauss-Legendre en van Newton-Cotes en bereken tevens de afwijking van de exacte waarde.

165) Bereken

$$\int_{-1}^{+1} \frac{x^{10}\,dx}{\sqrt{1-x^2}}$$

met de 5-punts formule van Gauss-Chebyshev en vergelijk het antwoord met de analytisch bepaalde waarde.

166) Bereken in 3 decimalen nauwkeurig

$$\int_0^{\infty} \frac{dx}{\sqrt{e^x + x}}.$$

167) Bereken in 4 decimalen nauwkeurig

$$\int_0^{\infty} \frac{e^{-x^2}\,dx}{1 + x^2}.$$

1.7 Varianten

Er bestaan enige varianten van de Gauss-formules, waarin sommige basispunten worden voorgeschreven en de andere optimaal worden gekozen. Bijvoorbeeld als variant op de formule van Gauss-Legendre (1.3.2) hebben we de formule van Lobatto, die ontstaat door twee basispunten aan de uiteinden van het integratie-interval en de andere optimaal te kiezen.
De maximale orde, die met n basispunten bereikt wordt, is dan 2n-2. Voor n = 2 of 3 krijgen we dan de trapezium-regel resp. de formule van Simpson.

1.8 Voordelen en toepassingen van de optimale Gauss-formules

Het voordeel van de optimale keuze der basispunten is gelegen in de hoge orde, die daardoor bereikt wordt. Dit voordeel komt echter alleen tot uiting, als men reeds een idee van de precisie heeft. Vergelijken we bijvoorbeeld de n-punts formule van Gauss-Legendre met de (2n-1)-punts Newton-Cotes formule, die beide van de orde 2n zijn. Om een idee van de precisie te krijgen moet men enerzijds zowel de n-punts als de (n-1)-punts Gauss-Legendre toepassen, waarvoor dus in totaal 2n-1 functie-waarden berekend moeten worden. Het verschil van deze formules geeft dan een idee van de precisie. Anderzijds kan men uit de 2n-1 functie-waarden op equidistante basispunten, die voor de Newton-Cotes formule nodig zijn, ook een idee van de precisie krijgen.
Bijvoorbeeld: Simpson = trapezium-regel + Richardson-correctie
en 5-punts Newton-Cotes = Simpson + Richardson-correctie (zie pag. 104 t/m 106), in welke gevallen de Richardson-correctie een idee van de precisie geeft.
De formule van Gauss-Legendre is dus pas voordelig, als men weet, dat voor zekere waarde van n voldoende nauwkeurigheid bereikt wordt.
Men kan dit bijvoorbeeld toepassen op het vaak voorkomende geval, dat een verzameling integralen wordt gevraagd, waarbij de integrand afhankelijk is van een of meer parameters. Men kan dan volstaan met steekproefgewijs voor sommige waarden van de parameters de nauwkeurigheid te controleren, mits de integrand niet te sterk van karakter verandert.

De formules van Jacobi kan men gebruiken, als de integrand een of meer singulariteiten van de gedaante $(x - s)^p$, $p > -1$ heeft. Er zijn echter ook andere middelen, die we zullen toelichten aan de hand van een voorbeeld.

Zij gevraagd $\int_0^1 \sqrt{x}\, e^{-x}\, dx$.

Past men direct Simpson toe, met $h = 0.1$, dan vindt men 0.37633.

In de buurt van $x = 0$ gedraagt de integrand zich als

$\sqrt{x}(1 - x + \frac{x^2}{2} - \ldots)$, dus als $\sqrt{x}$.

Trekken we nu $\sqrt{x}$ af, dan krijgen we

$$\int_0^1 \sqrt{x}\, e^{-x}\, dx = \int_0^1 \sqrt{x}\, dx + \int_0^1 \sqrt{x}\,(e^{-x} - 1)dx.$$

De integrand $\sqrt{x}(e^{-x} - 1)$ gedraagt zich als $x\sqrt{x}$, wat een "zwakkere" singulariteit is dan $\sqrt{x}$, en laat zich daardoor beter met Simpson integreren. Doen we dit met $h = 0.1$ en bepalen we de eerste integraal analytisch dan vinden we 0.37898. Door meer termen af te trekken kan men de singulariteit nog meer verzwakken. Geheel elimineren kan men de singulariteit door te stellen $x = y^2$. Wij vinden dan

$$\int_0^1 \sqrt{x}\, e^{-x}\, dx = 2 \int_0^1 y^2\, e^{-y^2}\, dy.$$

Bepalen we deze integraal met Simpson, wederom met $h = 0.1$, dan vinden we 0.378944.

Het correcte antwoord is ongeveer 0.37894 4692.

Ook singulariteiten van andere aard kan men vaak op een dergelijke wijze behandelen.

Bijvoorbeeld $\int_0^1 \cos x \ln x\, dx$ gaat door partiële integratie over in

$$\sin x \ln x \Big|_0^1 - \int_0^1 \frac{\sin x}{x}\, dx.$$

De eerste term is 0, de tweede term vertoont geen singulariteit meer. Passen we daarentegen de transformatie $x = e^{-t}$ toe, dan krijgen we

$$\int_0^1 \cos x \ln x \, dx = \int_0^\infty t \, e^{-t} \cos(e^{-t}) dt,$$

die met de formule van Laguerre geïntegreerd kan worden.
Bij het toepassen van de formules van Laguerre en Hermite is het van belang er voor te zorgen, dat de integrand asymptotisch goed is. Bijvoorbeeld de integrand van opgave 166 (pag. 330) gedraagt zich asymptotisch als $e^{-\frac{1}{2}x}$. Stellen we $x = 2y$, dan krijgen we

$$\int_0^\infty \frac{dx}{\sqrt{e^x + x}} = 2 \int_0^\infty e^{-y} \frac{1}{\sqrt{1 + 2y \exp(-2y)}} dy.$$

Nu is de integrand asymptotisch goed en kan Gauss-Laguerre worden toegepast.
Men kan ook proberen door een transformatie de integraal over te voeren in een integraal over een eindig interval zonder singulariteiten. Bijvoorbeeld de transformatie $x = -2 \ln t$ levert voor laatstgenoemde integraal:

$$\int_0^\infty \frac{dx}{\sqrt{e^x + x}} = 2 \int_0^1 \frac{dt}{\sqrt{1 - 2t^2 \ln t}}.$$

Voor meer bijzonderheden omtrent de optimale Gauss-formules zie Hildebrand [2], Fröberg [33] en vooral: Handbook of Mathematical Functions, uitgegeven door M. Abramowitz en I.A. Stegun, dat in hoofdstuk 25 een overzichtelijke opsomming geeft van de Gauss-integratie-formules, waaronder enige belangrijke gevallen van het Jacobi-type, en uitvoerige tabellen van basispunten en coëfficiënten bevat.

1.9 De integratie-formule van Filon

Integratie over een eindig interval zonder singulariteiten gaat bijna altijd goed met behulp van een formule van Newton-Cotes of Gauss-Legendre, door deze successievelijk op een aantal (voldoende kleine)

deelintervallen toe te passen. Er is echter een belangrijke uitzondering, nl. het geval, dat de integrand een oscillerende functie is. Dan vallen er namelijk veel cijfers weg. Voor dit geval bestaat een speciale integratie-formule van Filon. Hierin zijn de basispunten equidistant en bevatten de eindpunten (evenals bij Newton-Cotes), maar de gewichtsfunctie is sin mx. In geregen vorm (vgl. pag. 102) geschreven ziet de formule van Filon er aldus uit:

$$1.9.0 \qquad \int_{x_0}^{x_{2m}} f(x)\sin tx\,dx \approx h\Big[\alpha(f_0 \cos tx_0 - f_{2m} \cos tx_{2m}) + \beta \sum_{i=0}^{m}{}'' f_{2i} \sin(tx_{2i}) + \gamma \sum_{i=1}^{m} f_{2i-1} \sin(tx_{2i-1})\Big],$$

waarbij, als we stellen $th = \theta$,

$$\alpha = \frac{1}{\theta} + \frac{\sin 2\theta}{2\theta^2} - \frac{2\sin^2\theta}{\theta^3},$$

$$\beta = 2\left(\frac{1+\cos^2\theta}{\theta^2} - \frac{\sin 2\theta}{\theta^3}\right),$$

$$\gamma = 4\left(\frac{\sin\theta}{\theta^3} - \frac{\cos\theta}{\theta^2}\right),$$

en $\sum''$ betekent dat eerste en laatste term van een factor $\frac{1}{2}$ moeten worden voorzien. De formule is van de orde 4 d.w.z. is exact voor polynomen van graad kleiner dan 4.
De coëfficiënten α, β, γ zijn getabelleerd in Abramowitz-Stegun (zie pag. 333).

Opgaven

168) Bereken op de verschillende boven geschetste manieren

$$\int_0^1 \sqrt{x}\, e^{-x}\, dx.$$

169) Bereken in 5 cijfers

$$\int_0^1 \cos x \ln x\, dx.$$

170) Bereken in 5 cijfers

$$\int_0^\infty e^{-x} \ln x\, dx.$$

171) Bereken in 4 cijfers

$$\int_0^1 \frac{\arctan x}{x^{3/2}}\, dx.$$

172) Bereken in 4 cijfers

$$\int_0^{\pi/2} \frac{\cos x}{1 + x}\, dx.$$

Opgaven

173) Bereken in 3 cijfers (relatieve precisie)

$$\int_2^\infty \frac{e^{-x}}{x}\,dx.$$

174) De sinus-integraal Si is gedefiniëerd door

$$Si(x) = \int_0^x \frac{\sin(t)}{t}\,dt.$$

Bereken in 4 decimalen

$$\int_0^1 \frac{Si(x) - \sin(x)}{x^3}\,dx.$$

175) Zij y de functie, die voldoet aan de differentiaalvergelijking

$$y'' - xy = 0$$

en aan de randvoorwaarden

$$y(0) = 0.355028\ ,\quad y(1) = 0.135292.$$

Bereken y(0.5) in 4 decimalen.

2. Sommeren van reeksen

2.1 Inleiding

Wanneer de termen van een reeks snel genoeg naar nul convergeren, biedt het berekenen van de som niet veel moeilijkheden. Men neemt eenvoudig zoveel termen mee, als voor de vereiste precisie nodig zijn. In het algemeen is het evenwel gevaarlijk, op te houden zodra men een verwaarloosbaar kleine term ontmoet. Bijvoorbeeld de reeks

$$\sum_{k=1}^{\infty} \frac{\sin(k\pi/3)}{k!}$$

convergeert weliswaar snel, maar men mag toch niet ophouden bij de derde term, die toevallig nul is. Als wapen hiertegen kan men een ongelovigheidsparameter "tim" invoeren met de betekenis:
neem zoveel termen mee, totdat tim keer achtereen de termen verwaarloosbaar klein blijken te zijn.
Alleen als men weet dat de termen in absolute waarde monotoon afnemen, mag men gewoon tim = 1 kiezen. We gaan nu enige methodes bespreken voor het sommereen van langzaam convergerende reeksen.

2.2 De transformatie van Euler

Op langzaam convergerende alternerende reeksen kan men met succes de transformatie van Euler toepassen.
Beschouwen we de alternerende reeks

2.2.0 $$S = \sum_{k=0}^{\infty} (-1)^k u_k,$$

waarbij dus alle u_k positief zijn.
We gaan nu van elk paar opeenvolgende termen de helft bij elkaar voegen aldus:

$$\begin{aligned} S &= u_0 - u_1 + u_2 - u_3 + \ldots \\ &= \tfrac{1}{2}u_0 + \tfrac{1}{2}(u_0 - u_1) - \tfrac{1}{2}(u_1 - u_2) + \tfrac{1}{2}(u_2 - u_3) - \ldots \\ &= \tfrac{1}{2}u_0 - \tfrac{1}{2}(\Delta u_0 - \Delta u_1 + \Delta u_2 - \ldots). \end{aligned}$$

Dit levert de formule

2.2.1 $$S = \sum_{k=0}^{\infty} (-1)^k u_k = \tfrac{1}{2}u_0 - \tfrac{1}{2} \sum_{k=0}^{\infty} (-1)^k \Delta u_0.$$

Passen we op de aldus ontstane reeks dezelfde transformatie toe, dan krijgen we

$$S = \tfrac{1}{2}u_0 - \tfrac{1}{4}\Delta u_0 + \tfrac{1}{4} \sum_{k=0}^{\infty} (-1)^k \Delta^2 u_0.$$

Herhalen we dit procédé oneindig vaak dan krijgen we

$$S = \frac{1}{2} u_0 - \frac{1}{4} \Delta u_0 + \frac{1}{8} \Delta^2 u_0 - \frac{1}{16} \Delta^3 u_0 + \ldots .$$

Of in formule

2.2.2 $$\sum_{k=0}^{\infty} (-1)^k u_k = \frac{1}{2} \sum_{k=0}^{\infty} \frac{(-1)^k}{2^k} \Delta^k u_0.$$

Dit is de transformatie-formule van Euler. Hiervoor geldt de stelling, dat, als de eerste reeks convergeert, de tweede reeks eveneens convergeert en dezelfde som heeft. Voor alternerende reeksen levert de transformatie van Euler vaak een aanzienlijk sneller convergente reeks.

Voorbeeld

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{k+1} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \ldots$$

u_k	$\frac{1}{2}\Delta u_k$	$\frac{1}{4}\Delta^2 u_k$	$\frac{1}{8}\Delta^3 u_k$	$\frac{1}{16}\Delta^4 u_k$
1				
	- 1/4			
1/2		1/12		
	- 1/12		- 1/32	
1/3		1/48		1/80
	- 1/24		- 1/160	
1/4		1/120		
	- 1/40			
1/5				

De eerste 5 termen van Euler leveren 0.68854, wat reeds in 2 decimalen met de exacte waarde ln 2 $\approx$ 0.69315 overeenstemt. Tellen we daarentegen de eerste 5 termen gewoon op, dan vinden we nauwelijks een decimaal overeenstemming.

Handiger is het een schema op te stellen, niet van de u_k's en de gehalveerde differenties, maar van de termen $a_k = (-1)^k u_k$ en de gemiddelden. Hiertoe voeren we de voorwaartse gemiddelde operator M in, gedefiniëerd door:

2.2.3 $$Ma_k = \frac{1}{2}(a_k + a_{k+1})$$

of in operator-taal: $M = \frac{1}{2}(I + E)$.
(Vgl. formule (19.1) op pag. 80 voor de centrale gemiddelde operator μ).
De transformatie van Euler krijgt dan de gedaante

2.2.4 $$\sum_{k=0}^{\infty} a_k = \frac{1}{2} \sum_{k=0}^{\infty} M^k a_0.$$

Dit is slechts een andere schrijfwijze voor (2.2.2), als men stelt $a_k = (-1)^k u_k$. Men kan de formule formeel afleiden als volgt:

$$\sum_{k=0}^{\infty} a_k = \sum_{k=0}^{\infty} E^k a_0 = \frac{1}{I - E} a_0 = \frac{1}{2} \frac{1}{I - \frac{1}{2}(E + I)} a_0$$

$$= \frac{1}{2} \frac{1}{I - M} a_0 = \frac{1}{2} \sum_{k=0}^{\infty} M^k a_0.$$

Euler met Van Wijngaarden-strategie

Men kan de Euler-transformatie uitstellen, d.w.z. eerst enige termen gewoon optellen en dan de resterende reeks sommeren volgens Euler. Dit geeft vaak snellere convergentie.
Nog mooier is het om tijdens de opbouw van het gemiddelden-schema telkens te bepalen of men een Euler-stap zal zetten of de Euler-transformatie een keer zal uitstellen. Zijn op een gegeven moment r termen gewoon opgeteld en vervolgens n Euler-stappen gezet, dan luidt de benaderde som:

2.2.5 $$S_{r,n} = a_0 + a_1 + \ldots + a_{r-1} + \frac{1}{2} \sum_{k=0}^{n} M^k a_r.$$

Accepteren we de volgende Euler-term $M^{n+1}a_r$, dan wordt n opgehoogd en we hebben blijkbaar

2.2.6 $$S_{r,n+1} = S_{r,n} + \frac{1}{2} M^{n+1} a_r.$$

Stellen we daarentegen de Euler-sommatie een keer uit (de term $M^{n+1}a_r$ heet dan "verworpen"), dan wordt r opgehoogd en men kan bewijzen, dat

2.2.7 $$S_{r+1,n} = S_{r,n} + M^{n+1} a_r.$$

In de Van Wijngaarden-strategie wordt de Euler-term $M^{n+1}a_r$ geaccepteerd, als $|M^{n+1}a_r| < |M^n a_{r+1}|$, anders verworpen.
Dit proces is geprogrammeerd in de procedure "euler", gepubliceerd in P. Naur (ed.) , Revised Report on the algorithmic language ALGOL 60 (zie ook Modern Computing methods [1], pag. 124-126).

Keren we terug tot bovenstaand voorbeeld, dan krijgen we aldus (in 3 decimalen) het volgende schema

$S_{r,n}$	a_i	Ma_i	M^2a_i	M^3a_i	M^4a_i
.500	1.000				
		.250			
.625	-.500		.083		
		-.083			
.708	.333		-.021		
		.042		-.006	
.698	-.250		.008		-.002
		-.025		.002	
.695	.200		-.004		.000
		.017		-.001	
.693	-.167		.002		
		-.012			
.693	.143				

In elk stadium behoeft van het gemiddelden-schema slechts de laatste schuine rij te worden onthouden. De termen M^2a_0 en M^4a_1 zijn verworpen en kunnen, nadat ze bij de lopende som zijn opgeteld, worden vergeten.

Opgaven

176) Bereken de som van de volgende reeksen in 5 cijfers.

a) $1 - \frac{1}{\sqrt{3}} + \frac{1}{\sqrt{5}} - \frac{1}{\sqrt{7}} + \ldots$

b) $\frac{1}{\ln 2} - \frac{1}{\ln 3} + \frac{1}{\ln 4} - \frac{1}{\ln 5} + \ldots$

c) $1 - \frac{1}{9} + \frac{1}{25} - \frac{1}{49} + \ldots$

177) Schrijf een programma, dat de in opgave 176 vermelde reeksen sommeert met behulp van de bovengenoemde procedure "euler". Las hierbij (tijdelijk) in de procedure-body enige statements in, die het gemiddelden-schema en eventueel ook andere details typen.

2.3 Van Wijngaarden's transformatie

Deze transformatie is vooral geschikt voor langzaam convergerende reeksen met positieve termen.
Zij gegeven de reeks

2.3.0 $$S = \sum_{k=1}^{\infty} u_k \ .$$

We stellen nu

$$v_k = u_k + 2u_{2k} + 4u_{4k} + 8u_{8k} + \ldots \ .$$

Dan geldt $u_k = v_k - 2v_{2k}$, dus

2.3.1 $$\sum_{k=1}^{\infty} u_k = v_1 - 2v_2 + v_2 - 2v_4 + v_3 - 2v_6 + v_4 - 2v_8 + \ldots$$

$$= v_1 - v_2 + v_3 - v_4 + \ldots = \sum_{k=1}^{\infty} (-1)^{k-1} v_k .$$

Hiermee is de reeks overgevoerd in een alternerende reeks, die met behulp van Euler's transformatie gesommeerd kan worden. Hiervoor geldt de Stelling. Als de reeks $\sum_{k=1}^{\infty} u_k$ convergeert en de termen u_k positief zijn, dan convergeert de reeks $\sum_{k=1}^{\infty} (-1)^k v_k$ naar dezelfde limiet.
Omdat de indices van de termen in de v_k-reeksen met machten van twee oplopen, convergeren deze v_k-reeksen sneller dan de oorspronkelijke reeks. Hierin ligt juist het succes van deze transformatie.

Voorbeeld.

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} \approx 1.644934 .$$

De Van Wijngaarden transformatie levert:

$$v_k = \frac{1}{k^2}\left(1 + \frac{2}{4} + \frac{4}{16} + \frac{8}{64} + \ldots\right) = \frac{2}{k^2} \ ,$$

zodat de getransformeerde reeks luidt

$$2 \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^2} .$$

Deze reeks gaan we "euleren":

$2S_{r,n}$	v_k	Mv_k		
1.000	1			
		.375		
1.750	-.250			
		-.069		
1.681	.111		-.023	
		.024		-.008
1.658	-.062		.007	
		-.011		.002
1.642	.040		-.003	
		.006		
1.644	-.028			

In dit geval waren de termen v_k direct uit te rekenen, wat in het algemeen natuurlijk niet het geval is. Voor het berekenen van de v_k hoeft men alleen de oneven termen volgens de definitie $v_k = u_k + 2u_{2k} + 4u_{4k} + \ldots$ te berekenen; de even termen verkrijgt men dan uit de relatie

$$v_{2k} = (v_k - u_k)/2.$$

Weet men hoeveel termen meegenomen moeten worden, dan kan men nog handiger van achteren af beginnen en achtereenvolgens de formule

$$v_k = 2v_{2k} + u_k$$

toepassen, totdat k oneven is geworden.

2.4 De formules van Gregory, Gauss en Euler Maclaurin

We gaan uit van de formule van Gregory (3.2.3, pag. 118), waarin we $h = 1$ en $x_0 = 0$ stellen, zodat voor alle j geldt: $x_j = x_0 + jh = j$. Laten we nu k naar oneindig gaan, dan ontstaat, mits f_k, ∇_k, ∇_k^2, enz. naar nul gaan, de volgende formule:

$$2.4.0 \quad \sum_{j=i}^{\infty} f_j = \int_i^{\infty} f(x)dx + \frac{1}{2} f_i - \frac{1}{12} \Delta_i + \frac{1}{24} \Delta_i^2 - \frac{19}{720} \Delta_i^3 + \ldots .$$

Om deze formule te kunnen gebruiken, moeten we een gladde functie f zoeken, waarvan de waarden voor de gehele argumenten j gelijk zijn aan de termen f_j van de reeks en waarvan de integraal naar oneindig bestaat en liefst niet al te moeilijk te berekenen is. (Dezelfde formule ontstaat, als we in 3.1.1, pag. 117, k naar oneindig laten gaan.) Evenzo kunnen we uitgaan van de formule van Gauss (3.3.1, pag. 119), waarin we wederom $h = 1$ en $x_0 = 0$ stellen en k naar oneindig laten gaan. Als f_k en de centrale differenties naar nul gaan, dan ontstaat alzo de formule:

$$2.4.1 \quad \sum_{j=i}^{\infty} f_j = \int_i^{\infty} f(x)dx + \frac{1}{2} f_i - \frac{1}{12} \mu\delta_i + \frac{11}{720} \mu\delta_i^3 - \frac{191}{60480} \mu\delta_i^5 + \ldots .$$

Om deze formule te kunnen toepassen, moet de functie f bovendien nog gedefiniëerd zijn voor gehele argument-waarden $j < i$, want deze waarden zijn nodig voor de centrale differenties in het punt i. Dit lukt bijvoorbeeld niet voor de reeks $\sum_{j=1}^{\infty} \frac{1}{j^2}$, waarbij we definiëren $f(x) = \frac{1}{x^2}$; immers f(0) is niet gedefiniëerd. Een goede remedie is: eerst een paar termen gewoon optellen en dan de formule van Gauss toepassen. Ook dan convergeert de Gauss-formule niet (voor een hogere differentie is toch weer f(0) nodig), maar er treedt zgn. schijn-convergentie op, d.w.z. dat de differenties aanvankelijk voldoende snel afnemen om enige precisie te halen, maar later gaan toenemen. (Ook de equidistante interpolatie-formules zijn schijn-convergent, zie pag. 85.) Men moet in zo'n geval niet te weinig en niet te veel differenties meenemen. Heeft men hogere precisie nodig, dan moet men meer termen gewoon optellen voordat men Gauss' formule toepast.

Voorbeeld

Tellen we van bovengenoemde reeks slechts drie termen gewoon op, dan hebben we het differentie-schema rond i = 4 nodig:

f	δ	δ^2	δ^3	δ^4
1.0000				
	-7500			
.2500		6111		
	-1389		-5208	
.1111		903		4566
	- 486		- 642	
.0625		261		484
	- 225		- 158	
.0400		103		103
	- 122		- 55	
.0278		48		
	- 74			
.0204				

We vinden dan

$$\sum_{j=1}^{\infty} \frac{1}{j^2} \approx 1 + \frac{1}{4} + \frac{1}{9} + \int_4^{\infty} \frac{dx}{x^2} + \frac{1}{32} - \frac{1}{12}\,\mu\delta_4 + .015278\,\mu\delta_4^3$$

$$\approx 1 + .25 + .1111 + .25 + .0313 + .0030 - .0006$$

$$\approx 1.6448$$

wat in ruim 3 decimalen goed is. De $\mu\delta_4^5$-term is te groot, terwijl $\mu\delta_4^7$ oneindig is.

Tenslotte behandelen we de formule van Euler-Maclaurin, die speciaal geschikt is, als van de functie f gemakkelijk afgeleiden berekend kunnen worden. Deze formule wordt uit de formule van Gauss verkregen door de centrale differenties uit te drukken in afgeleiden. Hiervoor gebruiken we de centrale numerieke differentiatie-formules (vgl. pag. 98):

$$h\,f_i' = \mu\delta_i - \frac{1}{6}\,\mu\delta_i^3 + \frac{1}{30}\,\mu\delta_i^5 - \ldots$$

$$h^3 f_i^{(3)} = \mu\delta_i^3 - \frac{1}{4}\,\mu\delta_i^5 + \ldots$$

$$h^5 f_i^{(5)} = \mu\delta_i^5 - \ldots$$

Substitueren we deze met h = 1, in (2.4.1) dan krijgen we

$$\sum_{j=i}^{\infty} f_j = \int_i^{\infty} f(x)dx + \frac{1}{2} f_i - \frac{1}{12} f_i' + \frac{1}{720}\,\mu\delta_i^3 - \frac{23}{60480}\,\mu\delta_i^5 + \ldots$$

$$= \int_i^{\infty} f(x)dx + \frac{1}{2} f_i - \frac{1}{12} f_i' + \frac{1}{720} f_i^{(3)} - \frac{2}{60480}\,\mu\delta_i^5 + \ldots .$$

Dus tenslotte

2.4.2: $$\sum_{j=i}^{\infty} f_j = \int_i^{\infty} f(x)dx + \frac{1}{2} f_i - \frac{1}{12} f_i' + \frac{1}{720} f_i^{(3)} - \frac{1}{30240} f_i^{(5)} + \ldots .$$

Dit is de sommatie-formule van Euler-Maclaurin. Hierbij hoort natuurlijk ook een integratie-formule, die uit Gauss' formule (3.3.1, pag. 119) en bovenstaande differentiatie-formules op dezelfde wijze wordt verkregen. Deze integratie-formule van Euler-Maclaurin, die eigenlijk op pag. 119 thuis hoort (zie besluit aldaar) luidt:

$$2.4.3 \quad \frac{1}{h}\int_{x_i}^{x_k} f(x)dx = \sum_{j=i}^{k''} f_j - \frac{h}{12}(f'_k - f'_i) + \frac{h^3}{720}(f_k^{(3)} - f_i^{(3)}) - \ldots .$$

De Euler-Maclaurin coëfficiënten nemen blijkbaar sneller af, dan die van Gauss.
Keren we terug tot bovenstaand voorbeeld, waar we eerst weer drie termen gewoon optellen. Euler-Maclaurin levert dan

$$\sum_{j=1}^{\infty} \frac{1}{j^2} \approx 1 + \frac{1}{4} + \frac{1}{9} + \int_4^{\infty} \frac{dx}{x^2} + \frac{1}{32} + \frac{1}{12} \times \frac{1}{32} - \frac{1}{720} \times \frac{3}{128} + \frac{1}{30240} \times \frac{45}{1024}$$

$$\approx 1.642361 + .002604 - .000033 + .000001 = 1.644933$$

wat in bijna 6 decimalen correct is, vgl. $\pi^2/6 \approx 1.6449340668$.
Euler-Maclaurin kan soms met vrucht op de volgende wijze worden toegepast. We trachten een gegeven reeks te schrijven als de som van twee reeksen:

$$\sum_{j=i}^{\infty} v_j = \sum_{j=i}^{\infty} f_j + \sum_{j=i}^{\infty} (v_j - f_j),$$

waarbij de f-reeks zo eenvoudig is, dat hij met Euler-Maclaurin gesommeerd kan worden en waarbij de resterende reeks zo snel convergeert, dat gewoon optellen van niet al te veel termen voldoende precisie levert.

Opgaven.

178) De zeta-functie van Rieman is gedefiniëerd als

$$\zeta(x) = \sum_{k=1}^{\infty} k^{-x} \quad , \quad x > 1.$$

Schrijf een programma, dat $\zeta(x)$ in 5 decimalen berekent voor $x = 2(1)10$.

Doe, ter controle, de berekening met de hand voor $x = 3$, 5 en 10.

179) De constante C van Euler is gedefiniëerd door

$$C = \lim_{n\to\infty} (1 + \frac{1}{2} + \ldots + \frac{1}{n} - \ln n) = 1 + \sum_{k=2}^{\infty} (\frac{1}{k} + \ln \frac{k-1}{k}).$$

Bereken C in 5 decimalen.

3. Niet-lineaire formules ter bepaling van limieten

3.1 Aitken extrapolatie

Zij gegeven een iteratie-proces, waarin uitgaande van een startwaarde x_0 successievelijk iteratie-waarden x_i worden berekend voor $i = 1, 2, 3, \ldots$ (vgl. pag. 152). Stel nu dat de iteratie-waarden voldoen aan

$$x_i = s + a\lambda^i, \tag{3.1.0}$$

waarbij s, a en λ constanten zijn.
Als $|\lambda| < 1$, convergeert de rij blijkbaar naar de (gevraagde) limiet s.
We stellen nu van 3 opeenvolgende iteraties het differentie-schema op:

$$\begin{array}{lll} x_{i-1} & & \\ & \nabla x_i & \\ x_i & & \delta^2 x_i \\ & \Delta x_i & \\ x_{i+1} & & \end{array}$$

Uit (3.1.0) volgt blijkbaar

$$\nabla x_i = a\lambda^{i-1}(\lambda - 1),$$

$$\Delta x_i = a\lambda^i(\lambda - 1),$$

$$\delta^2 x_i = a\lambda^{i-1}(\lambda - 1)^2,$$

dus

$$a\lambda^{i+1} = \frac{(\Delta x_i)^2}{\delta^2 x_i}. \tag{3.1.1}$$

Hieruit vinden we de benaderingsformule voor s:

$$x_{i+1} - \frac{(\Delta x_i)^2}{\delta^2 x_i}. \tag{3.1.2}$$

Dit is de extrapolatie-formule van Aitken. Men spreekt ook wel van Aitkens delta-kwadraatproces.
Als (3.1.0) bij benadering geldt, zal (3.1.2) de limiet s beter benaderen, dan x_{i+1}.

Toepassing

Aitken extrapolatie kan o.a. worden toegepast bij de matrix-maal-vector methode voor het berekenen van een eigenwaarde en bijbehorende eigenvector (zie pag. 189 e.v.). Succes is verzekerd, als $|\lambda_1|$ iets groter dan $|\lambda_2|$ is en de andere eigenwaarden in absolute waarde veel kleiner zijn.
Dan is na enige iteratie-stappen het effect van de kleinere eigenwaarden te verwaarlozen.
Itereren we volgens formule (3.3.1 pag. 192), waarbij dus steeds $||u^{(i)}|| = 1$, dan geldt na enige iteratie-stappen

$$||v^{(i)}|| = ||Au^{(i)}|| \approx |\lambda_1| + C\left(\frac{\lambda_2}{\lambda_1}\right)^i ,$$

zodat we Aitken kunnen toepassen.
Beschouwen we het voorbeeld op pag. 193. Na drie iteratie-stappen vinden we:

$$||v^{(1)}|| = 309.44$$

$$\nabla_2 = 25.82$$

$$||v^{(2)}|| = 335.26 \qquad \delta_2^2 = -25.09$$

$$\Delta_2 = 0.73$$

$$||v^{(3)}|| = 335.99$$

De formule van Aitken (3.1.2) levert nu

$$\lambda_1 \approx 335.99 - \frac{(0.73)^2}{-25.09} \approx 335.99 + 0.02 = 336.01,$$

terwijl de exacte eigenwaarde 336 bedraagt.

Vector-iteratie

We beschouwen nu een iteratie-proces, waarin de iterates n-vectoren $u^{(i)}$ zijn, i = 0, 1, 2, ...
Ook hierop kan Aitkens δ^2-proces worden toegepast, en wel op twee manieren:

a) Men kan elk element van de vectoren afzonderlijk behandelen. De formule van Aitken krijgt dan de gedaante

$$3.1.3 \qquad u_k^{(i+1)} - \frac{(\Delta u_k^{(i)})^2}{\delta^2 u_k^{(i)}}, \quad k = 1(1)n,$$

waarbij $\Delta u_k^{(i)} = u_k^{(i+1)} - u_k^{(i)}$ en $\delta^2 u_k^{(i)} = u_k^{(i+1)} - 2u_k^{(i)} + u_k^{(i-1)}$.

Nadeel van deze formule is, dat zij onbruikbaar wordt, als voor sommige k-waarden $\delta^2 u_k^{(i)}$ dicht bij nul ligt.

b) Men kan gebruik maken van de Samelson-inverse van een vector, die voor een reële vector u gedefiniëerd is als

$$3.1.4 \qquad \frac{1}{u^T u} u^T.$$

De formule van Aitken krijgt dan de gedaante

$$3.1.5 \qquad u^{(i+1)} - \frac{\delta^2 u^{(i)T} \Delta u^{(i)}}{\delta^2 u^{(i)T} \delta^2 u^{(i)}} \Delta u^{(i)},$$

waarin teller en noemer scalaire producten zijn van de rij-vector $\delta^2 u^{(i)T}$ en de kolom-vector $\Delta u^{(i)} = u^{(i+1)} - u^{(i)}$ respectievelijk $\delta^2 u^{(i)} = u^{(i+1)} - 2u^{(i)} + u^{(i-1)}$.
Deze formule wordt pas onbruikbaar, als alle elementen van $\delta^2 u^{(i)}$ dicht bij nul liggen.

Toepassing

Passen we formule (3.1.5) toe op het voorbeeld van pag. 193 met i = 3, dan krijgen we

$$\nabla u^{(i)} = \begin{pmatrix} -.0141 \\ +.0279 \\ -.0473 \end{pmatrix}, \quad \Delta u^{(i)} = \begin{pmatrix} -.0025 \\ +.0050 \\ -.0076 \end{pmatrix}, \quad \delta^2 u^{(i)} = \begin{pmatrix} +.0116 \\ -.0229 \\ +.0397 \end{pmatrix}.$$

Dus $\dfrac{\delta^2 u^{(i)T}\ \Delta u^{(i)}}{\delta^2 u^{(i)T}\ \delta^2 u^{(i)}} \approx -\dfrac{.0004452}{.002235} \approx -.1992,$

zodat formule (3.1.5) levert

$$u^{(i+1)} + .1992\ \Delta u^{(i)} \approx \begin{pmatrix} .8729 \\ -.2182 \\ -.4364 \end{pmatrix},$$

wat in 4 decimalen met de exacte eigenvector overeenstemt.
Gewoon itererend zonder Aitken zouden we daarentegen 7 iteratie-stappen nodig hebben om 4 decimalen precisie te krijgen.

3.2 De epsilon-algorithme van Wynn

Zij gegeven een iteratie-proces waarin successievelijk iteratiewaarden x_i worden verkregen, die voldoen aan

3.2.0 $$x_i = s + a_1\lambda_1^i + \ldots + a_t\lambda_t^i \quad , i = 0, 1, 2, \ldots$$

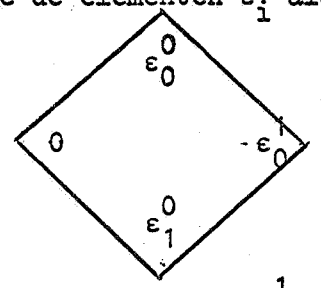
In dit geval kan men met vrucht de epsilon-algorithme toepassen. Deze wordt gedefinieerd als volgt.
Zij voor i=0,1,2,.. per definitie

3.2.1 $$\varepsilon_i^0 = x_i, \qquad \varepsilon_i^{-1} = 0,$$

dan kunnen we vervolgens voor k=0,1,2,... berekenen

3.2.2 $$\varepsilon_i^{k+1} = \varepsilon_{i+1}^{k-1} + (e_{i+1}^k - e_i^k)^{-1},$$

M.a.w. plaatsen we de elementen ε_i^k aldus in schema



$$\varepsilon_0^2 \qquad\qquad \varepsilon_0^3$$

$$0 \qquad\qquad \varepsilon_1^1$$

$$\varepsilon_2^0 \qquad\qquad \varepsilon_1^2$$

$$0 \qquad\qquad \varepsilon_2^1$$

$$\varepsilon_3^0$$

dan vormen de 4 elementen voorkomend in (3.2.2) een ruit. Voor de aldus verkregen elementen ε_i^k geldt de

Stelling Als de iteratie-waarden $\varepsilon_i^0 = x_i$ voldoen aan (3.2.0), dan geldt $\varepsilon_i^{2t} = s$.

De elementen ε_i^2 blijken gelijk te zijn aan het resultaat van Aitkens extrapolatie-formule (3.1.2). Zijn de iteratie-waarden vectoren, dan kan men in formule (3.2.2) de Samelson-inverse (zie 3.1.4) gebruiken.

Toepassingen

Evenals Aitkens δ^2-proces, kan ook de epsilon-algorithme worden toegepast op de matrix-maal-vector methode, voor het berekenen van eigenwaarden zowel als eigenvectoren. Een andere toepassingsmogelijkheid is de iteratieve oplossing van een lineair stelsel volgens Jacobi of Gauss-Seidel (zie pag. 146-148). Om dit na te gaan zullen we eerst de convergentie van deze processen onderzoeken.

Zij gevraagd de oplossing van het stelsel Ax=b. We schrijven de matrix van het stelsel in de vorm

$$A = L + D + U,$$

waarbij D een diagonaalmatrix is en L en U onder- respectievelijk bovendriehoeks-matrices met louter nullen op de hoofddiagonaal zijn. Dan kunnen we de Jacobi-iteratiestap (formule 9.1.0 pag. 147 en 9.1.1 pag. 148) schrijven in de gedaante

$$x^{(i+1)} = D^{-1}b - D^{-1}(L+U)\, x^{(i)}.$$

Voor de oplossings-vector x geldt

$$x = D^{-1}b - D^{-1}(L+U)x.$$

Aftrekking levert

$$x^{(i+1)} - x = -D^{-1}(L+U)(x^{(i)} - x) = M(x^{(i)} - x),$$

als we definiëren

$$M = -D^{-1}(L+U).$$

Hieruit volgt:

3.2.3 $$x^{(i)} - x = M^i(x^{(0)}-x)$$

Laat nu M de eigenwaarden $\lambda_1,\ldots,\lambda_n$ bezitten met lineair onafhankelijke eigenvectoren $v_1,\ldots,v_n$ (we nemen aan dat M diagonaliseerbaar is, vgl. pag. 180-184). Dan kan elke vector worden geschreven als lineaire combinatie van de eigenvectoren, dus in het bijzonder kan $x^{(0)}-x$ worden geschreven als

$$x^{(0)}-x = \alpha_1 v_1 + \ldots + \alpha_n v_n$$

Dus

$$x^{(i)}-x = M^{(i)}(x^{(0)}-x) = \alpha_1\lambda_1^i v_1 + \ldots + \alpha_n\lambda_n^i v_n$$

ofwel

3.2.4 $$x^{(i)} = x + \alpha_1\lambda_1^i v_1 + \ldots + \alpha_n\lambda_n^i v_n.$$

Hieruit volgt, dat het iteratie-proces van Jacobi convergeert, als alle eigenwaarden van matrix $M=-D^{-1}(L+U)$ in absolute waarde kleiner dan 1 zijn.

De Gauss-Seidel-iteratiestap (formules 9.2.0 & 1 pag. 148) kan worden geschreven in de gedaante

$$x^{(i+1)} = (L+D)^{-1} b - (L+D)^{-1}Ux^{(i)}.$$

Definiëren we nu

$$M = -(L+D)^{-1}U,$$

dan volgt op analoge wijze als boven, dat met deze M de formules 3.2.3 en 3.2.4 voor de Gauss-Seidel iteratie gelden. Dit proces convergeert dus, als alle eigenwaarden van $M=-(L+D)^{-1}U$ in absolute waarde kleiner dan 1 zijn.

Formule (3.2.4) stemt overeen met (3.2.0), waarbij t gelijk aan n is, s de oplossingsvector x is en $a_1,\ldots,a_n$ de vectoren $\alpha_1 v_1,\ldots,\alpha_n v_n$ zijn Hieruit volgt, dat, zowel bij Jacobi als bij Gauss-Seidel, de vectoren ε_i^{2n}, die na 2n epsilon-stappen ontstaan, gelijk zijn aan de oplossingsvector x. Aangezien hiervoor 2n iteratie-stappen nodig zijn kost dit vrij veel rekenwerk. Aantrekkelijk wordt het pas, als de meeste eigenwaarden van M dicht bij 0 liggen, zodat na enige iteratie-stappen nog maar een paar eigenwaarden meedoen en dientengevolge ε_i^{2t} voor t veel kleiner dan n reeds een goede benadering van de oplossingsvector is. Voor literatuur omtrent de epsilon-algorithme zie o.a.

P. Wynn, Math. of Comp. 16(1962) pag. 301-322 en
P. Wynn, Proc. of IFIP congres 62, pag. 149-156.

3.3 De Quotient-Differentie algorithme

Deze algorithme dient voor het bepalen van nulpunten van polynomen of transcendente functies. Voor polynomen gaat het als volgt. Beschouwen we een polynoom

$$x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n$$

en laten n startwaarden $x_0,\ldots,x_{n-1}$ gegeven zijn. Dan kunnen we voor i=n, n+1,... achtereenvolgens berekenen:

$$x_i = -a_1 x_{i-1} - \ldots - a_n x_{i-n}.$$

Als het polynoom een enkel nulpunt w_1 heeft, dat in absolute waarde alle andere overtreft, dan convergeren de quotiënten

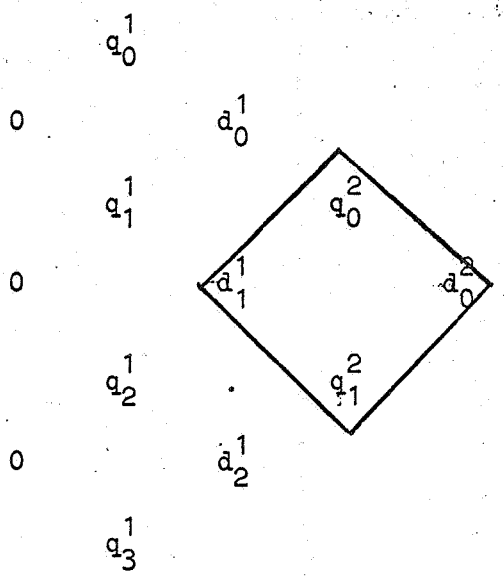
$$q_i^1 = x_{i+1}/x_i$$

lineair naar w_1 (methode van Bernoulli).
Stellen we bovendien per definitie $d_i^{-1} = 0$, dan kunnen we achtereenvolgens berekenen

$$d_i^k = d_{i+1}^{k-1} + (q_{i+1}^k - q_i^k),$$

$$q_i^{k+1} = q_{i+1}^k \times (d_{i+1}^k / d_i^k).$$

Evenals bij de epsilon-algorithme, vormen in beide formules de optredende elementen een ruit, als zij aldus in schema (het "QD-schema") worden geplaatst.



Onder bepaalde voorwaarden convergeren de rijen q_i^1, q_i^2,...,q_i^n voor toenemende i naar de nulpunten van het polynoom. Op deze wijze blijkt de methode niet erg stabiel te zijn. In de eerste ruitformule wordt het verschil van twee (op den duur nagenoeg gelijke) q's gevormd, waardoor cijfers wegvallen, in de tweede ruitformule worden de fouten in de d's opgeblazen door quotient-vorming. Het is evenwel mogelijk, na het opstellen van een geschikte start, het QD-schema van boven naar beneden op te bouwen. Hierbij berekent men telkens volgens bovenstaande ruitformules niet de rechterpunt, maar de onderpunt van de ruit uit de drie andere elementen. Op deze wijze is de berekening wel stabiel. De convergentie is langzaam. Het voordeel van de methode is, dat het schattingen voor alle nulpunten tegelijk kan leveren. Uitgaande van deze schattingen kan men daarna met bijvoorbeeld Newton-iteratie de nulpunten nauwkeuriger berekenen. Voor bijzonderheden omtrent deze QD-algorithme zie P. Henrici, Elements of Numerical Analysis (1964).

Opgaven

180) Het iteratie-proces van Jacobi convergeert, als de hoofddiagonaal van de matrix overheerst, d.w.z. als voor de matrix $A = (a_{ij})$ geldt:

$$| a_{ii} | > \sum_{\substack{j=1 \\ j \neq i}}^{n} | a_{ij} | \quad , \quad i=1(1)n.$$

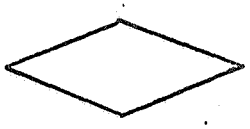
Bewijs dit.

181) Los met behulp van Gauss-Seidel iteratie, eventueel met Aitken extrapolatie of epsilon-algorithme, het volgende stelsel op in 4 decimalen

$$\begin{pmatrix} 5 & -1 & 0 & 0 & 0 \\ 1 & 7 & -2 & 0 & 0 \\ 0 & 2 & 11 & -3 & 0 \\ 0 & 0 & 3 & 9 & -2 \\ 0 & 0 & 0 & 1 & 5 \end{pmatrix} x = \begin{pmatrix} 6 \\ -8 \\ 12 \\ -8 \\ 4 \end{pmatrix}$$

182) Bereken met behulp van de matrix-maal-vector methode en Aitkens δ^2-proces de grootste eigenwaarde en bijbehorende eigenvector van de volgende matrix in 3 à 4 decimalen.

$$\begin{pmatrix} 525 & -352 & 810 & -220 \\ 5280 & -3639 & 8460 & -2310 \\ 4050 & -2820 & 6580 & -1800 \\ 7700 & -5390 & 12600 & -3450 \end{pmatrix}$$



Hoofdstuk 10. Partiële differentiaal-vergelijkingen

0. Inleiding

In dit hoofdstuk zullen we slechts enkele fragmenten uit de theorie der partiële differentiaal-vergelijkingen behandelen en een paar oplossings-methoden schetsen. We beperken ons tot vergelijkingen, waarin de onbekende $u = u(x,y)$ een functie is van twee variabelen x en y. We zullen alleen aan zgn. "quasi-lineaire" vergelijkingen van de eerste en tweede orde aandacht schenken.

1. Quasi-lineaire vergelijkingen van de eerste orde

De algemene gedaante hiervan, voor twee variabelen, is

$$a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} = c, \tag{1.1}$$

waarbij a, b en c functies zijn van x, y en u, maar niet afhangen van $\frac{\partial u}{\partial x}$ of $\frac{\partial u}{\partial y}$. Zijn a en b onafhankelijk van u (en is c lineair in u), dan heet (1.1) een lineaire vergelijking.
We gebruiken de standaard-notatie

$$\frac{\partial u}{\partial x} = p \quad , \quad \frac{\partial u}{\partial y} = q, \tag{1.2}$$

waardoor (1.1) overgaat in

$$ap + bq = c. \tag{1.3}$$

Om een eenduidige oplossing te krijgen, moet bovendien een beginvoorwaarde gegeven worden, waarin de waarde van u wordt voorgeschreven, niet in een enkel punt, maar langs een gegeven kromme O in het (x,y)-vlak.
We gaan nu na hoe de oplossing $u(x,y)$ zich zou kunnen gedragen, als het punt (x,y) een willekeurige kromme K in het (x,y)-vlak doorloopt. We kunnen K aangeven met een parameter-voorstelling

$$x = x(k) \quad , \quad y = y(k) \tag{1.4}$$

en de oplossing is dus voor deze waarden van x en y een functie van de parameter k:

$$u = u(k) = u(x(k), y(k)). \tag{1.5}$$

Hiervoor geldt

$$\frac{du(k)}{dk} = \frac{\partial u(x,y)}{\partial x}\frac{dx}{dk} + \frac{\partial u(x,y)}{\partial y}\frac{dy}{dk}.$$

Dit wordt volkomen analoog bewezen, als de stelling op pag. 216A.
Met de standaard-notatie (1.2) wordt dit:

1.6 $$\frac{du}{dk} = p\frac{dx}{dk} + q\frac{dy}{dk}.$$

Men kan beide leden nog "vermenigvuldigen" met dk, waardoor de formule overgaat in

1.7 $$du = p\,dx + q\,dy.$$

Hierin komt de parameter k niet meer voor. De betekenis van deze formule is niets anders, dan dat (1.6) geldt langs willekeurige krommen K.
We kiezen nu de kromme K zó, dat

1.8 $$\frac{dx}{dk} = a \quad , \quad \frac{dy}{dk} = b.$$

Dan gaat (1.6) wegens (1.3) over in

1.9 $$\frac{du}{dk} = pa + qb = c.$$

M.a.w. langs een kromme K die aan (1.8) voldoet, moet de functie u voldoen aan de gewone differentiaal-vergelijking $\frac{du}{dk} = c$, waarbij c een functie van x, y en u, dus van k en u is.
Als K de beginkromme O ergens snijdt, dan moet u bovendien in zo'n snijpunt de aldaar voorgeschreven waarde aannemen. Alles bij elkaar is dit dus een beginwaarde-probleem, dat onder bepaalde omstandigheden de functie u langs K eenduidig vastlegt.
Elke kromme K in het (x,y)-vlak, die aan (1.8) voldoet, heet <u>karakteristiek</u> van de differentiaal-vergelijking (1.1).
We schrijven (1.8) liever in de vorm:

1.10 $$\frac{dy}{dx} = \frac{b}{a}.$$

Deze vergelijking heet karakteristieke vergelijking. Als de differentiaal-vergelijking (1.1) lineair is, hangen a en b alleen van x en y af, zodat (1.10) dan een gewone differentiaal-vergelijking is. Door ieder punt (x_0,y_0) van de beginkromme O gaat in het algemeen één karakteristiek, die aan het beginwaarde probleem (1.10) met $y(x_0) = y_0$ moet voldoen.
Hiermee hebben we dus het oplossen van een lineaire partiële differentiaal-vergelijking herleid tot het oplossen van gewone differentiaal-vergelijkingen. Het hoeft geen betoog, dat dit vaak tot aanzienlijk rekenwerk zal leiden.
Mocht ongelukkigerwijze de beginkromme O met een karakteristiek samenvallen, dan heeft (1.1) ofwel geen enkele of oneindig veel oplossingen.

2. Quasi-lineaire vergelijkingen van de tweede orde

2.0 Inleiding

De tweede-orde vergelijkingen zijn verreweg het belangrijkste, aangezien vele physische en technische problemen hiertoe leiden. De algemene gedaante van het quasi-lineaire geval in twee variabelen is

2.0.1 $$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} = e,$$

waarbij a, b, c en e functies zijn van x, y, u, p en q, maar niet van de partiële afgeleiden van de tweede orde afhangen. Zijn a, b en c onafhankelijk van u, p en q (en is e lineair hierin), dan heet (2.0.1) een lineaire vergelijking.
We gebruiken de standaard-notatie

2.0.2 $$\frac{\partial^2 u}{\partial x^2} = r\ ,\quad \frac{\partial^2 u}{\partial x \partial y} = s\ ,\quad \frac{\partial^2 u}{\partial y^2} = t,$$

waardoor (2.0.1) de gedaante krijgt

2.0.3 $$ar + bs + ct = e.$$

We gaan weer langs een kromme K met parametervoorstelling (1.4) wandelen. Hierlangs moet de oplossing $u = u(k) = u(x(k), y(k))$, weer voldoen aan (1.6) of de equivalente formule (1.7):

2.0.4 $$du = \frac{\partial u}{\partial x}\,dx + \frac{\partial u}{\partial y}\,dy = p\,dx + q\,dy.$$

Vervangen we hierin u door p of q, dan krijgen we

$$dp = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right)dx + \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial x}\right)dy = r\,dx + s\,dy,$$

$$dq = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial y}\right)dx + \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y}\right)dy = s\,dx + t\,dy.$$

Voegen we hierbij de gegeven differentiaal-vergelijking (2.0.3) dan vinden we, dat r, s en t moeten voldoen aan het volgende stelsel lineaire vergelijkingen:

2.0.5 $$\begin{cases} dx \cdot r + dy \cdot s = dp \\ dx \cdot s + dy \cdot t = dq \\ a\,r + b\,s + c\,t = e. \end{cases}$$

Hierdoor zijn r, s en t eenduidig bepaald, tenzij voor de determinant D geldt:

$$D = \begin{vmatrix} dx & dy & 0 \\ 0 & dx & dy \\ a & b & c \end{vmatrix} = a(dy)^2 - b\,dxdy + c(dx)^2 = 0.$$

Delen we door $(dx)^2$, dan krijgen we

2.0.6 $$a\left(\frac{dy}{dx}\right)^2 - b\left(\frac{dy}{dx}\right) + c = 0.$$

Deze vergelijking heet <u>karakteristieke vergelijking</u> van de partiële differentiaal-vergelijking (2.0.1) en een kromme K in het (x,y)-vlak die hieraan voldoet heet <u>karakteristiek</u> van de partiële differentiaal-vergelijking. Voor het al of niet reëel zijn van de karakteristieken is het teken van de discriminant beslissend. We onderscheiden drie gevallen, alnaargelang voor zekere oplossing u van de differentiaal-vergelijking (2.0.1) en voor zeker gebied G van het (x,y)-vlak geldt:

1) $b^2 - 4ac > 0$.

In dit geval heet de differentiaal-vergelijking hyperbolisch in G. Door elk punt van G gaan twee karakteristieken.

2) $b^2 - 4ac = 0$.

In dit geval heet de differentiaal-vergelijking parabolisch. Door elk punt van G gaat één karakteristiek (of eigenlijk twee samenvallende karakteristieken).

3) $b^2 - 4ac < 0$.

In dit geval heet de differentiaal-vergelijking elliptisch. Er zijn nu in G geen reële karakteristieken.

Vaak zal een differentiaal-vergelijking in één gebied elliptisch en in een ander gebied parabolisch of hyperbolisch zijn. Is de vergelijking niet-lineair, dan zal het karakter meestal ook nog van de oplossing u afhangen.

Bijvoorbeeld: stationaire stroming in samendrukbaar gas kan worden beschreven met een differentiaal-vergelijking, die elliptisch bij subsonische en hyperbolisch bij supersonische snelheden is (zie Modern Computing methods, pag. 109).

Onnodig te zeggen, dat dergelijke differentiaal-vergelijkingen vaak zeer lastig op te lossen zijn. Wij zullen nu enkele lineaire differentiaal-vergelijkingen met constante coëfficiënten bekijken, die dus in het hele (x,y)-vlak hyperbolisch, parabolisch of elliptisch blijven.

2.1 Hyperbolische differentiaal-vergelijkingen

In het hyperbolische geval gaan er, zoals we gezien hebben, door elk punt twee karakteristieken. Langs de karakteristieken heeft stelsel 2.0.5, wegens het nul zijn van de determinant, dan en slechts dan een oplossing, als de rang van de matrix

2.1.1
$$\begin{pmatrix} dx & dy & 0 & dp \\ 0 & dx & dy & dq \\ a & b & c & e \end{pmatrix}$$

gelijk aan 2 is. Hieraan is bijvoorbeeld voldaan als

$$\begin{vmatrix} dx & 0 & dp \\ 0 & dy & dq \\ a & c & e \end{vmatrix} = 0,$$

wat na deling door dxdy levert:

2.1.2
$$a \frac{dp}{dx} + c \frac{dq}{dy} = e.$$

Deze voorwaarde moet dus gelden langs de karakteristieken.

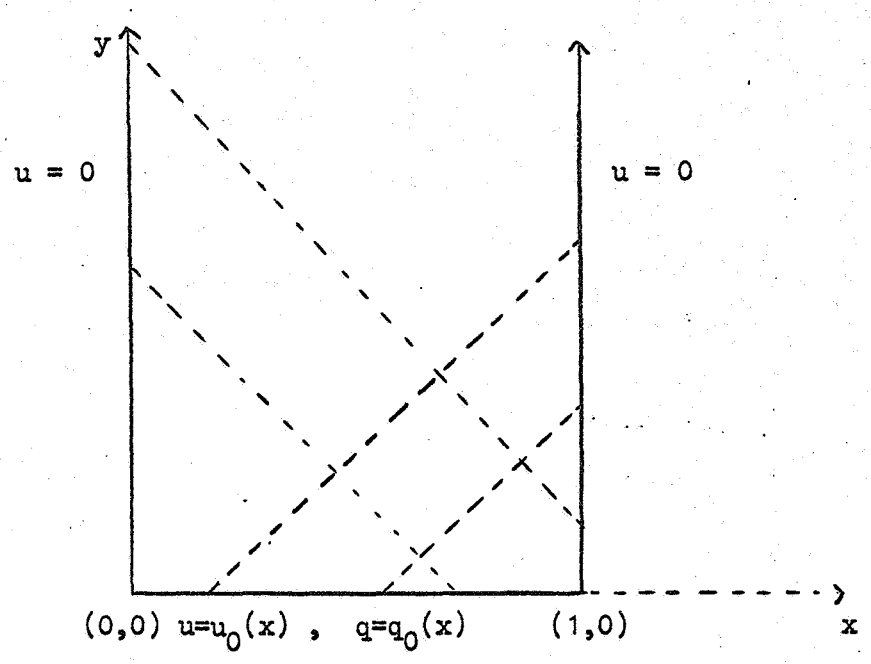
Voorbeeld

De beweging van een trillende snaar kan, als we y gelijk aan de voortplantingssnelheid maal de tijd stellen, worden beschreven door de differentiaal-vergelijking

2.1.3
$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial y^2} .$$

Hierin is x de plaats-coördinaat op de snaar en u de uitwijking van de snaar in x op het tijdstip y.
Nemen we aan, dat de snaar een lengte 1 heeft en aan beide uiteinden is vastgeklemd en dat op het begintijdstip y = 0 uitwijking en snelheid van de snaar gegeven zijn, dan luiden dus de randvoorwaarden:

2.1.4
$$\begin{cases} u(0,y) = u(1,y) = 0 , \quad y \geq 0, \\ u(x,0) = u_0(x) , \quad q(x,0) = q_0(x) , \quad 0 \leq x \leq 1. \end{cases}$$



De karakteristieke vergelijking (2.0.6) luidt in dit geval

$$(\frac{dy}{dx})^2 - 1 = 0,$$

ofwel $\frac{dy}{dx} = \pm 1$, zodat de karakteristieken voldoen aan

$$x \pm y = \text{constant}.$$

Langs deze lijnen moet voorwaarde (2.1.2) gelden, die nu de gedaante krijgt:

$$\frac{dp}{dx} - \frac{dq}{dy} = 0,$$

ofwel $\frac{dq}{dp} = \frac{dy}{dx} = \pm 1$, zodat we langs de karakteristieken de voorwaarde krijgen

$$p \pm q = \text{constant}.$$

Met behulp hiervan kan men, uitgaande van de randvoorwaarden, de functies p en q langs de karakteristieken bepalen en vervolgens de oplossing u verkrijgen (zie Modern Computing methods, hoofdstuk 11). We gaan hierop niet nader in.

Andere oplossings-methoden worden verkregen, door de partiële afgeleiden te benaderen door middel van differenties. Hoe dit gaat, zullen we alleen in de volgende gevallen toelichten.

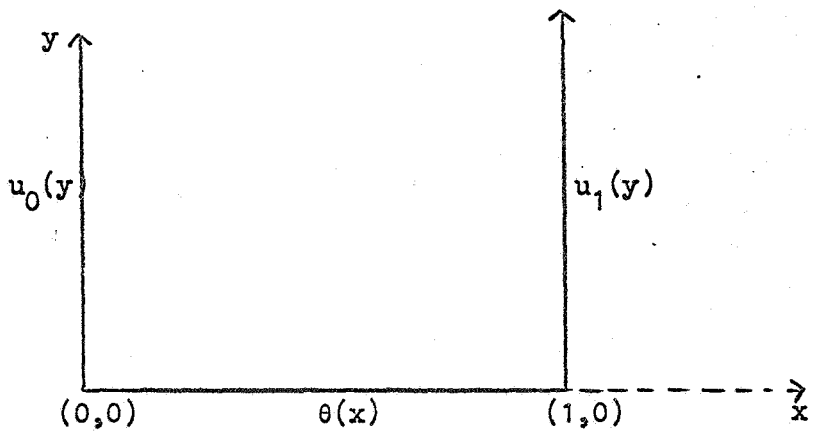
2.2 Parabolische differentiaal-vergelijkingen

Het klassieke voorbeeld is de een-dimensionale warmte-vergelijking

2.2.1 $$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial y} ,$$

waarin u de temperatuur, y (een zekere dimensieloze maat) de tijd en x de plaatscoördinaat is. Aan deze vergelijking voldoet de warmte-geleiding in een dunne homogene staaf. Nemen we aan dat aan beide einden van een staaf ter lengte 1 het temperatuurverloop gegeven is en dat bovendien de begintemperatuur langs de hele staaf gegeven is, dan zien de rand-voorwaarden er blijkbaar aldus uit:

2.2.2 $$\begin{cases} u(0,y) = u_0(y) , \quad u(1,y) = u_1(y) , \quad y \geq 0, \\ u(x,0) = \theta(x) , \quad 0 \leq x \leq 1. \end{cases}$$



Het verschil met het hyperbolische geval is,dat hier voor y = 0 alleen de functie u gegeven is.

We gaan nu de partiële afgeleiden benaderen door differenties. Hiertoe definiëren we eerst een rooster van punten (x_i,y_j) als volgt:

2.2.3
$$\begin{cases} x_i = ih\ , & i = 0(1)n, \\ y_j = jk\ , & j = 0,\ 1,\ 2,\ \ldots, \end{cases}$$

waarbij $h = 1/n$, zodat $x_0 = 0$ en $x_n = 1$ op de rand liggen. Voorlopig kiezen we n en k willekeurig. Voor de functie-waarden gebruiken we de afkorting

2.2.4
$$u(x_i,y_j) = u_{ij}.$$

Voor het benaderen van de partiële afgeleiden gebruiken we alleen de kopterm van de formules op pag. 97 & 98, dus

2.2.5
$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{ij} \approx \frac{1}{h^2}\delta_x^2 u_{ij} = \frac{1}{h^2}\left(u_{i-1,j} - 2u_{ij} + u_{i+1,j}\right),$$

2.2.6
$$\left(\frac{\partial u}{\partial y}\right)_{ij} \approx \frac{1}{k}\Delta_y u_{ij} = \frac{1}{k}\left(u_{i,j+1} - u_{i,j}\right),$$

of de centrale formule

2.2.7
$$\left(\frac{\partial u}{\partial y}\right)_{ij} \approx \frac{1}{k}\mu\delta_y u_{ij} = \frac{1}{2k}\left(u_{i,j+1} - u_{i,j-1}\right).$$

Substitutie van de benaderingen (2.2.5 & 6) in (2.2.1) levert

2.2.8
$$u_{i,j+1} = u_{ij} + \frac{k}{h^2}\left(u_{i-1,j} - 2u_{ij} + u_{i+1,j}\right),\ i = 1(1)n-1.$$

Met behulp hiervan kunnen we, als voor het tijdstip $y = jk$ benaderde functie-waarden u_{ij}, $j = 0(1)n$, bekend zijn, voor het tijdstip $y = (j+1)k$ de functie-waarden $u_{i,j+1}$ benaderen voor $i = 1(1)n-1$. De waarden $u_{0,j+1}$ en $u_{n,j+1}$ zijn gegeven door de randvoorwaarden, zodat we hierna evenzo de volgende stap van j+1 naar j+2 kunnen zetten. De eerste keer starten we met $j = 0$, waarvoor de u-waarden door de randvoorwaarde langs de x-as gegeven zijn, nl. $u_{i0} = \theta(ih)$, $i = 0(1)n$. Deze methode heet de <u>expliciete methode</u>.

De methode is slechts dan stabiel en convergeert voor fijner wordend rooster slechts dan naar de oplossing van de differentiaal-vergelijking, als $k \leq \frac{1}{2} h^2$. Dit betekent een ernstige beperking voor de stapgrootte k, zodat men in de y-richting veel stappen moet zetten, wat uiteraard veel tijd kost.

Men zou kunnen proberen, in plaats van (2.2.6) de centrale formule (2.2.7) te gebruiken. Dit leidt echter tot een formule, die altijd instabiel en dus onbruikbaar is (vgl. pag. 245 waar een analoge formule voor gewone differentiaalvergelijkingen instabiel blijkt).

We krijgen wel een stabiele formule als volgt. Formule (2.2.6) geeft eigenlijk een benadering van $(\frac{\partial u}{\partial y})_{i,j+\frac{1}{2}}$. Het is dus redelijk aan de andere kant ook een benadering van $(\frac{\partial^2 u}{\partial x^2})_{i,j+\frac{1}{2}}$ te gebruiken. Hiervoor kiezen we het gemiddelde van $(\frac{\partial^2 u}{\partial x^2})_{ij}$ en $(\frac{\partial^2 u}{\partial x^2})_{i,j+1}$ elk benaderd volgens (2.2.5). Dan ontstaat de volgende impliciete formule van Crank-Nicolson

$$2.2.9 \quad -u_{i-1,j+1} + 2(1 + \frac{h^2}{k})u_{i,j+1} - u_{i+1,j+1} =$$

$$u_{i-1,j} - 2(1 - \frac{h^2}{k})u_{i,j} + u_{i+1,j}, \quad i = 1(1)n-1.$$

Als voor zekere j de waarden u_{ij}, $i = 0(1)n$ bekend zijn en bovendien de randwaarden $u_{0,j+1}$ en $u_{n,j+1}$ gegeven zijn, dan is deze formule een stelsel van n-1 lineaire vergelijkingen in de n-1 ombekenden $u_{i,j+1}$, $i = 1(1)n-1$. De matrix van dit stelsel is de tridiagonale matrix

$$\begin{pmatrix} 2(1 + h^2/k) & -1 & & & \\ -1 & 2(1 + h^2/k) & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2(1 + h^2/k) \end{pmatrix}$$

Elke stap van j naar j+1 bestaat dus nu uit het oplossen van dit tridiagonale stelsel. Dit vergt weliswaar meer rekenwerk, dan een stap volgens de expliciete formule (2.2.8), het grote voordeel is echter, dat de methode van Crank-Nicolson stabiel is voor elke verhouding van h en k. We zijn nu dus niet gedwongen tot zeer kleine stappen in de y-richting, wat een aanzienlijke tijdwinst betekent. Een eveneens stabiele formule wordt verkregen, indien we als benadering van $\frac{\partial^2 u}{\partial x^2}$ kiezen $(\frac{\partial^2 u}{\partial x^2})_{i,j+1}$, weer benaderd volgens (2.2.5); de formule die dan ontstaat heet formule van Laasonen.

2.3 Elliptische differentiaal-vergelijkingen

In het elliptische geval worden de randvoorwaarden, in tegenstelling tot de andere gevallen, gegeven op een gesloten kromme in het (x,y)-vlak. Op deze randkromme wordt meestal gegeven de waarde van u, of van de afgeleide van u in de richting loodrecht op de randkromme. De toepassingen betreffen stationaire toestanden, in het bijzonder potentiaal-problemen. Het klassieke voorbeeld van een elliptische vergelijking is de vergelijking van Poisson

2.3.0 $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x,y)$$

met belangrijk bijzonder geval de vergelijking van Laplace

2.3.1 $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.$$

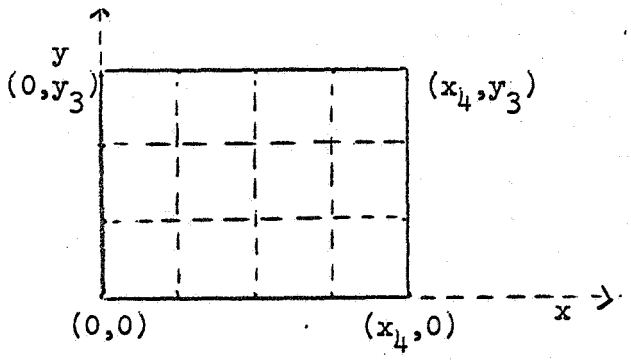
We gaan weer de partiële afgeleiden benaderen door differenties. We nemen eenvoudigheidshalve aan, dat de functie u gegeven is op een rechthoekige rand. We leggen een zijde langs de x-as en een langs de y-as en verdelen de zijden in m respectievelijk n gelijke stukken.

We stellen

2.3.2
$$\begin{cases} x_i = ih \ , & i = 0(1)m, \\ y_j = jk \ , & j = 0(1)n, \end{cases}$$

waarbij h = lengte x-zijde gedeeld door m en
k = lengte y-zijde gedeeld door n.

Voor m = 4, n = 3 ziet het rooster er uit als volgt:



De partiële afgeleiden worden benaderd volgens formule (2.2.5) en de analoge formule voor de y-richting:

2.3.3
$$\left(\frac{\partial^2 u}{\partial y^2}\right)_{ij} \approx \frac{1}{k^2}\,\delta_y^2\,u_{ij} = \frac{1}{k^2}\,(u_{i,j-1} - 2u_{ij} + u_{i,j+1}).$$

Nemen we voor het gemak aan dat h en k aan elkaar gelijk zijn (vierkant rooster), dan krijgen we voor de differentiaalvergelijking (2.3.0) de volgende benadering:

2.3.4
$$u_{i-1,j} + u_{i+1,j} - 4u_{ij} + u_{i,j-1} + u_{i,j+1} = f_{ij},$$
$$i = 1(1)m-1,\ j = 1(1)n-1,$$

waarbij $f_{ij} = f(x_i, y_j)$.
Met m = 4, n = 3 leidt dit voor de onbekende u_{ij}-waarden tot het volgende lineaire stelsel

$$\begin{pmatrix} -4 & 1 & 1 & 0 & 0 & 0 \\ 1 & -4 & 0 & 1 & 0 & 0 \\ 1 & 0 & -4 & 1 & 1 & 0 \\ 0 & 1 & 1 & -4 & 0 & 1 \\ 0 & 0 & 1 & 0 & -4 & 1 \\ 0 & 0 & 0 & 1 & 1 & -4 \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ u_{21} \\ u_{22} \\ u_{31} \\ u_{32} \end{pmatrix} = \begin{pmatrix} f_{11} - u_{10} - u_{01} \\ f_{12} - u_{13} - u_{02} \\ f_{21} - u_{20} \\ f_{22} - u_{23} \\ f_{31} - u_{41} - u_{30} \\ f_{32} - u_{42} - u_{33} \end{pmatrix}.$$

Voor willekeurige waarden van m en n is de matrix A van het stelsel een blok-tridiagonale matrix van de gedaante

2.3.5 $$A = \begin{pmatrix} B & I & & & \\ I & B & I & & \\ & I & \ddots & \ddots & \\ & & \ddots & \ddots & I \\ & & & I & B \end{pmatrix}, \text{ waarbij } B = \begin{pmatrix} -4 & 1 & & & \\ 1 & -4 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -4 \end{pmatrix}$$

en I een eenheidsmatrix is. B en I zijn van de orde n-1, A heeft m-1 matrices B langs zijn hoofddiagonaal en is dus van de orde (m-1)(n-1). Men kan een dergelijk stelsel oplossen voor verschillende waarden van m en n en kijken of in de gemeenschappelijke roosterpunten voldoende overeenstemming ontstaat. Ook kan men eventueel hogere orde differentie-correcties aan het rechterlid toevoegen en itereren (vgl. pag. 261-263). Als de rand geen rechthoek is, zal men in de buurt van de rand meestal een interpolatie-formule moeten gebruiken, om de randvoorwaarden in rekening te brengen. Ook kan het wenselijk zijn in bepaalde gedeelten een fijner rooster aan te brengen.

Eigenwaarde-problemen

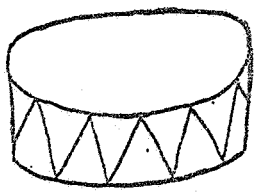
De trilling van een membraan kan worden beschreven door de differentiaal-vergelijking

2.3.6 $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \lambda u = 0,$$

waarbij de randvoorwaarde luidt, dat de functie u op de rand nul moet zijn. Alleen voor bepaalde waarden van λ, de <u>eigenwaarden</u>, heeft de vergelijking een oplossing u, die niet overal nul is. Is de rand een rechthoek, dan kunnen we, evenals boven, een rooster aanbrengen.
Het probleem leidt dan tot oplossing van het matrix-eigenwaarde-probleem

2.3.7 $$Au - \lambda u = 0,$$

waarbij A de matrix (2.3.5) is.



Nawoord

Ter aanvulling van de literatuur-lijsten op pag. 1 en pag. 21-22 wordt hieronder een lijst gegeven van publicaties, die in de loop van de cursus ter sprake zijn gekomen, van enige series ALGOL 60-procedures en van enige syllabi.

Voor de samenstelling van deze syllabus heb ik vooral gebruik gemaakt van J.A. Zonneveld [51], National Physical Laboratories [1], F.B. Hildebrand [2], C.E. Fröberg [33], E. van Spiegel [54], A. van Wijngaarden [55] en J.H. Wilkinson [29] en [29a].

Ik betuig mijn dank aan de stafleden en vroegere stafleden en gastmedewerkers van de Rekenafdeling voor hun waardevolle suggesties of critische opmerkingen. Met name ben ik hiervoor dank verschuldigd aan Prof.Dr.Ir. A. van Wijngaarden, aan Dr. J.A. Zonneveld, aan Prof.Dr. R.J. De Vogelaere en aan de heer B.J. Mailloux.
Bovendien breng ik dank aan Mevr. H. Roqué-de Hoyer voor de verzorging van het typewerk en aan de heer D. Zwarst voor de reproductie.

Juni 1967

T.J. Dekker

Mathematisch Centrum
2e Boerhaavestraat 49
Amsterdam.

Aanvullende literatuurlijst

22. A.C. Aitken, Determinants and matrices.

23. A.S. Householder, The theory of matrices in numerical analysis (1964).

24. R. Zurmühl, Matrizen.

25. C. Lanczos, Applied Analysis (1957).

26. E. Stiefel, Einführung in die numerische Mathematik (1961).

27. V.N. Faddeeva, Computational methods of linear algebra (1959).

28. D.K. Faddeev & V.N. Faddeeva, Computational methods of linear algebra (1963).

29. J.H. Wilkinson, The algebraic eigenvalue problem (Oxford 1965).

29a. J.H. Wilkinson, Rounding errors in algebraic processes.

30. W.E. Milne, Numerical solution of differential equations (1953).

31. P. Henrici, Discrete variable methods in ordinary differential equations (1962).

32. P. Henrici, Elements of numerical analysis (1964).

33. C.E. Fröberg, Introduction to Numerical Analysis (1965).

34. A. Ralston, A first course in Numerical analysis (1965).

35. R.W. Soutworth & S.L. de Leeuw, Digital Computation and Numerical methods (1965).

36. J.A. Zonneveld, Automatic Numerical Integration (Math. Centrum, Tract nr. 8, 1964).

37. M. Abramowitz & I.A. Stegun (ed.), Handbook of Mathematical Functions, Nat. Bur. Stand. AMS 55 (1964).

38. E. Stiefel, Ueber diskrete und lineare Tschebyscheff-Approximationen, Num. Math. 1 (1959), 1-28.

39. G. Golub & P. Businger, Linear least squares solutions by Householder Transformations, Num. Mat. 7 (1965), 206-216 & 269-276.

40. T.J. Dekker, Evaluation of determinants, solutions of systems of linear equations and matrix inversion (Math. Centrum, MR 63, 1963).

41. H.S. Wall, Continued fractions (New York 1948).

42. P. Wynn, Acceleration Techniques for iterated Vector and Matrix Problems, Math. of Comp. 16 (1962), 301-322 & Proc. of IFIP congres 1962, 149-156.

43. P. Wynn, General purpose Vector Epsilon Algorithm Procedures, Numer. Mat. 6 (1964), 22-36.

Series ALGOL 60 procedures

44. Algorithms, Communications of the ACM, vanaf vol. 4 (1960). Index hiervan in vol. 7 (1964), 146-149 en in de december-nummers vanaf 1964. Collected Algorithms from Comm. ACM,uitgegeven in 1967.

45. ALGOL Programming, BIT vanaf vol. 1 (1961).

46. Handbook Series linear algebra, Series numerical integration and Series special functions, Numerische Mathematik vanaf vol. 4 (1962).

47. Serie AP 200 t/m 260, Rekenafdeling Math. Centrum (1962-1965).

48. Numerieke Procedures, Wiskundige Dienst, T.H. Delft, vanaf 1963.

49. Algorithms, Computer Bulletin & Computer Journal, vanaf 1964.

50. RC informatie, Rekencentrum, T.H. Eindhoven, vanaf 1966.
Zie ook:
Index van Algorithmes, ALGOL-programma's en ALGOL-procedures, (informatieblad nr. 4, Rekencentrum R.U. Groningen 1967).

Syllabi van cursussen Wetenschappelijk Rekenen

51. J.A. Zonneveld, Cursus WRA, Numerieke Wiskunde (2 delen) 1957/1958 (niet meer voorradig).

52. F.J.M. Barning, Cursus WRA, Lineaire Algebra en Meetkunde (2 delen) 1961.

53. R. Timman, Cursus WRB, Partiële differentiaalvergelijkingen (3 delen), 1963.

54. E. van Spiegel, Cursus WRB, Numerieke Analyse (3 delen), 1964.

55. A. van Wijngaarden, Cursus WRB, Proces Analyse (1 deel), 1965.

Inhoud

Proefwerken	aantal pagina's
dd. 2-3-'65	3
practicum I en II	4
ALGOL practicum & Numerieke analyse, 15-6-'65	4

Proefwerken

	aantal pagina's
Numerieke analyse & Programmeren, 29-10-'66	2
dd. 9-2-'67, deel I en II	2
Programmeren	1
Programmeren, 27-4-'67	1
Numerieke Wiskunde, 8-6-'67	2
Numerieke Wiskunde & Programmeren, 22-6-'67	2

Errata

Kaft, deel I, "Rekenaar" moet zijn "Rekenen"

pag. 140, regel 3 "4.0" moet zijn "4.1"

pag. 158, regel 6 "positive" moet zijn "non-negative"

pag. 273, regel 2 v.o. "i = 1(1)n-1" moet zijn "j = 1(1)n-1"

pag. 281, regel 6 v.o. "waarden $\neq 1$" moet zijn "waarden ± 1"

pag. 291, regel 3 v.o. "$\sum_{j=0}^{n-1} b^2$" moet zijn "$\sum_{j=0}^{n-1} b_j^2$"

pag. 309, regel 12 "$a_0 = \frac{1}{2}\int$" moet zijn "$a_0 = \frac{1}{2\pi}\int$"

pag. 316, regel 2 "2.7" moet zijn "3."

(dezelfde wijziging geldt voor de formule-nummering op pag. 316 - 320).

pag. 318, regel 3 v.o. "$\frac{1}{a_1} Q_n(x)$" moet zijn "$a_1 Q_n(x)$"

pag. 322, regel 7 v.o. "a,b" moet zijn "[a,b]"

pag. 328, regel 5 "$\int_{-1}^{+1}$" moet zijn "$\int_0^{\infty}$"

pag. 338, regel 3 "Δu_0" moet zijn "Δu_k"

regel 6 "$\Delta^2 u_0$" moet zijn "$\Delta^2 u_k$"

pag. 342, regel 10 v.o. "positief" moet zijn "positief en monotoon niet-toenemend"

pag. 343, regel 4 ""euleren":" moet zijn ""euleren"; we stellen $a_k = (-1)^{k-1}/k^2$ en krijgen dan:"

Cursus WRA
T.J. Dekker

29-10-1966
13.00-15.00 uur

Proefwerk deel I

Uit examen voor het diploma A voor Wetenschappelijk Rekenen.

Numerieke Analyse,

dd. 6 september 1966

---

1. De functie y(x) voldoet aan de differentiaalvergelijking

$$\frac{dy}{dx} + xy = \phi(x),$$

waarbij $\phi(x)$ door de volgende tabel is gegeven

x	$\phi(x)$	x	$\phi(x)$
0.0	1.00000	0.6	1.16412
0.1	1.00499	0.7	1.21579
0.2	1.01980	0.8	1.27059
0.3	1.04399	0.9	1.32660
0.4	1.07683	1.0	1.38177
0.5	1.11730		

Voorts is gegeven y(0.45) = 1.33867.
Bepaal y(0.55) in vijf decimalen.

2. Bereken in twee decimalen een reële oplossing (x, y) van het stelsel

$$\begin{cases} x = 2\ell n(x+y) \\ y = \exp(y-x). \end{cases}$$

Cursus WRA
T.J. Dekker

29-10-1966
15.15-17.30 uur

Proefwerk deel II.

Uit examen voor het diploma A voor Wetenschappelijk Rekenen.
Programmeren,

dd. 7 september 1966.

Maak de opgaven 1 en 2 en bovendien, naar keuze, 3 of 4.

1. Maak een procedure die op grond van een gegeven integer array A[1:m, 1:n] vijf integers h, i, j, k en ℓ aflevert, zodanig dat h, gedefinieerd door

$$h = A[i,j] - A[k,\ell],$$

maximaal is.

2. Schrijf een programma, dat een getal a van een band leest en daarna de reële wortel van de vergelijking

$$x^3 - x^2 + x - a = 0$$

in 6 decimalen nauwkeurig berekent en daarna uitponst. Geef eerst aan, welke methode U gebruikt.

3. Maak een procedure die een stelsel van n lineaire vergelijkingen met n onbekenden

$$\sum_{j=1}^{n} A_{ij}\, x_j = b_i, \quad i = 1, \ldots, n,$$

oplost.
Verondersteld mag worden dat Gauss-eliminatie zonder verwisselingen geen aanleiding geeft tot pivots die nul of klein zijn.

4. Maak een procedure die met behulp van de regel van Simpson de integraal

$$\int_a^b f(x)\, dx$$

berekent.
De procedure moet de te gebruiken stapgrootte h zodanig bepalen, dat het verschil tussen integratie met stap h en integratie met stap 2h in absolute waarde kleiner is dan een voorgeschreven tolerantie eps. Gebruik bij voorkeur Jensen's device om een expressie voor de integrand als actuele paremeter mee te kunnen geven.

Cursus W.R.-A
T.J. Dekker

9-2-1967
13.00 - 15.00 uur

Proefwerk deel I

1) Zij ax + b de lineaire functie, die op het interval $\left[0, \frac{\pi}{4}\right]$ de functie sin(x) zo goed mogelijk benadert in de zin van Chebyshev. Gegeven

$$\pi \approx 3.14159265$$

wordt gevraagd a en b te bepalen in 4 decimalen nauwkeurig. Hoeveel bedraagt de maximale afwijking tussen sin(x) en de lineaire benadering?

2) Bereken van het stelsel

$$\begin{aligned}
3.7821\,x + 1.9635\,y - 0.4867\,z &= 4.2366\\
6.0543\,x - 3.1081\,y + 1.7429\,z &= 6.1121\\
-2.4708\,x + 0.0431\,y + 3.8297\,z &= -1.8345\\
1.5134\,x + 7.0346\,y - 2.7130\,z &= 2.3607\\
-1.3611\,x - 2.9787\,y + 13.2364\,z &= 0.6126
\end{aligned}$$

de beste oplossing in de zin van de kleinste kwadraten.

Cursus W.R.-A
T.J. Dekker

9-2-1967
15.15 - 17.30 uur

Proefwerk deel II

1) a. Hoe luidt de interpolatie-formule van Aitken (herhaalde lineaire interpolatie)?

b. Bereken met behulp van deze formule f(5), als f een derdegraadspolynoom is, waarvan de volgende tabel gegeven is:

x	0	2	4	6	8	10
f(x)	-5	-3	47	193	483	965

2) a. Hoe luidt de interpolatie-formule van Stirling (in centrale differenties, symmetrisch rond een basispunt x_0)?

b. Schets hoe deze formule kan worden afgeleid en laat gedetailleerd zien hoe de eerste vier termen worden verkregen.

3) a. Bepaal de waarde van

$$\int_{-1}^{+1} \frac{T_j(x)T_k(x)}{\sqrt{1-x^2}}\,dx, \qquad \text{voor } j,k=0,1,2,---,$$

als Tj het j-de graads Chebyshev-polynoom is.

b. Hoe luidt de (oneindige) reeks van Chebyshev en hoe luidt de formule voor de coëfficiënten van deze reeks?

c. Schets hoe het economiseren van machtreeksen geschiedt.

4) Van de functie f(x) zijn de volgende waarden gegeven:

$$f(-2h),\ f(-h),\ f(0),\ f(h),\ f(2h).$$

Zij het tweedegraadspolynoom $g(x) = ax^2 + bx + c$ de beste benadering in de zin der kleinste kwadraten van f(x) op de punten

$$x=-2h,\ x=-h,\ x=0,\ x=h,\ x=2h.$$

a. Druk de coëfficiënten a, b en c uit in de gegeven functiewaarden van f(x).

b. Bepaal de benaderingsformule voor de afgeleide van f(x) in het punt x=0, die ontstaat wanneer men f(x) vervangt door de benadering g(x).
Welke is de afwijking die deze benaderingsformule vertoont van $f'(0)$, indien f(x) een polynoom is van de vierde graad?

Cursus WR-A
T.J. Dekker

13.00-15.00 uur

Proefwerk

PROGRAMMEREN

1) De integratie-formule van Gauss van de orde 6 luidt:

$$\int_{x_i}^{x_k} f(x)dx \approx h \left\{ \mu\delta_k^{-1} - \mu\delta_i^{-1} - \frac{1}{12}(\mu\delta_k - \mu\delta_i) + \frac{11}{720}(\mu\delta_k^3 - \mu\delta_1^3) \right\}.$$

Schrijf een programma, dat volgens deze formule bepaalde integralen berekent, als (op een ponsband of op ponskaarten) gegeven zijn:

1) het eerste tabel-argument a en het tabel-interval h,
2) een natuurlijk getal n, gevolgd door n waarden van de integrand in de punten $x_i = a + ih$, $i = 0(1)n-1$,
3) een natuurlijk getal q, gevolgd door q paren i en k, die de indices van begin- en eindpunt van het integratie-interval zijn.

Het programma hoeft alleen een correct antwoord af te leveren, als de indices i en k zodanig gekozen zijn, dat de formule zonder meer toepasbaar is. In het andere geval geve het een alarm-indicatie.

2) a) Zij f een functie met periode 2π, gegeven als expressie of als real procedure.
Schrijf een procedure, die bij gegeven natuurlijk getal m, de coëfficiënten voor de discrete m-punts harmonische analyse van f berekent.
(Deze coëfficiënten zijn, op een constante factor na, gelijk aan een som van f-waarden maal zekere cosinus- of sinus-waarden).

b) Schrijf hieromheen een programma, dat deze procedure voor m = 8 toepast op de functie

$$f(x) = \frac{4 - 2\cos x}{5 - 4\cos x}.$$

Cursus WRA
T.J. Dekker

Proefwerk

27-4-'67
13.00 - 15.00 uur

PROGRAMMEREN

1) Schrijf een procedure, die voor gegeven a en c berekent

$$\int_c^{\infty} f(x,a)dx$$

met behulp van de 4-punts Gauss-Legendre formule, integrerend van c tot 2c, en de 3-punts Gauss-Laguerre formule, integrerend van 2c naar oneindig, als nog gegeven is, dat

$$\lim_{x\to\infty} e^{ax} f(x,a)$$

voor alle a bestaat en een eindige waarde ongelijk aan nul heeft. De functie f(x,a) is hierbij beschikbaar als real procedure.

2) Schrijf een programma, dat een matrix inleest en vervolgens een eigenwaarde en eigenvector ervan berekent met behulp van de matrix-maal-vector methode. Het programma moet redelijk efficiënt en zuinig in geheugen-gebruik zijn, voor het geval de matrix zeer groot is (orde ca. 1000) en de meeste elementen (90% of meer) nul zijn.
Bedenk hierbij tevens, hoe de matrix geschikt op een band kan worden gegeven, en pas het programma dienovereenkomstig aan.
Eventueel kan in het programma Aitken extrapolatie toegepast worden.

Cursus WRA
T.J. Dekker

8-6-1967
15.00-17.15 uur

Proefwerk Numerieke Wiskunde

(Alleen de syllabus mag worden geraadpleegd.)

1) Van de functie $Z(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$ is de volgende tabel gegeven.

x	Z(x)
.00	.39894228
.02	.39886250
.04	.39862325
.06	.39822483
.08	.39766771
.10	.39695255
.12	.39608021
.14	.39505174
.16	.39386836
.18	.39253148
.20	.39104269
.22	.38940376
.24	.38761662
.26	.38568337
.28	.38360629
.30	.38138782

Bereken, zo nauwkeurig als de gegeven tabel toelaat, door middel van numerieke integratie

$$\int_{.00}^{.20} Z(x)dx \quad \text{en} \quad \int_{.00}^{.30} Z(x)dx$$

en door middel van numerieke differentiatie

$$Z'(.20) \quad \text{en} \quad Z''(.20).$$

2) Zij gegeven matrix A en een benadering x van de bij $\lambda \approx 630.09$ horende eigenvector:

$$A = \begin{pmatrix} 420 & 210 & 140 & 105 \\ 210 & 140 & 105 & 84 \\ 140 & 105 & 84 & 70 \\ 105 & 84 & 70 & 60 \end{pmatrix} \quad , \quad x = \begin{pmatrix} 1.0000000 \\ 0.5701721 \\ 0.4067790 \\ 0.3181410 \end{pmatrix} .$$

Bereken het Rayleigh-quotiënt $\frac{x^T Ax}{x^T x}$ in 15 cijfers nauwkeurig.

Cursus WRA
T.J. Dekker

22-6-67
13.00 - 15.00

Proefwerk Numerieke Wiskunde

1. Van een functie y, die voldoet aan de differentiaalvergelijking

$$y'' + (1 + \frac{1}{4x^2})y = 0$$

is het volgende tabelletje gegeven

x	y
2.0	.316630
2.1	.241436
2.2	.163694
2.3	.084230

Bereken, zo nauwkeurig als deze gegeven tabel toelaat, de waarden $y(2.4)$, $y(2.5)$ en $y(2.6)$

2. Bereken in 6 decimalen nauwkeurig de wortels van de vergelijking

$$x^4 - 20x^3 + 101x^2 - 20x + 101 = 0.$$

Cursus WRA
T.J. Dekker

22-6-67
15.15 - 17.15

Proefwerk Programmeren

1) Schrijf een ALGOL 60 procedure, die de oplossing van een stelsel lineaire vergelijkingen berekent met behulp van Gauss-Seidel iteratie. Vermijd hierin deling door nul en laat een beperkt aantal iteraties toe.
Snelheid en zuinig geheugen-gebruik, met name voor het geval de meeste elementen van de matrix nul zijn, wordt op prijs gesteld.

2) a) Schrijf een procedure, die een reeks sommeert met behulp van van Wijngaardens transformatie

$$\sum_{k=1}^{\infty} u_k = \sum_{k=1}^{\infty} (-1)^{k-1} v_k ,$$

waarbij $v_k = u_k + 2u_{2k} + 4u_{4k} + \ldots .$
De procedure dient gebruik te maken van de procedure "euler" (Revised Report on ALGOL 60, Example 1).

b) Schrijf een programma, dat met behulp van de procedure (a) de constante van Euler berekent, die gelijk is aan

$$1 + \sum_{k=2}^{\infty} \left(\frac{1}{k} + \ln \frac{k-1}{k}\right)$$