TECHNISCHE HOGESCHOOL EINDHOVEN
Afdeling Algemene Wetenschappen
Onderafdeling der Wiskunde

NUMERIEKE WISKUNDE I

Prof. Dr. G.W. Veltkamp

Voorjaarssemester 1961

Inhoudsbeschrijving

NUMERIEKE WISKUNDE I

Voorjaarssemester 1961

Paragrafen	blz
0. INLEIDING	1
1. HET OPLOSSEN VAN VERGELIJKINGEN	5
1.2. Successieve substitutie	5
1.3. De vergelijking $F(x) = 0$	13
1.4. Regula Falsi	15
1.5. Stelsels vergelijkingen	16
1.6. Algebraïsche vergelijkingen	19
2. LINEAIRE VERGELIJKINGEN	37
2.1. Inleiding	37
2.2. Directe methoden	41
2.3. Iteratieve methoden	49
3. HET BEPALEN VAN EIGENWAARDEN VAN EEN MATRIX	61
3.1. Inleiding	61
3.2. Een proces voor de bepaling van een grootste eigenwaarde	65
3.3. De overige eigenwaarden	68
3.4. Triagonaalmatrices	71
4. INTERPOLATIE	78
4.1. Interpolatie volgens Lagrange	78
4.2. Interpolatie bij gelijke intervallen met behulp van differenties	81
4.3. Inverse interpolatie	88
4.4. Algemenere interpolatiemethoden	89
5. NUMERIEKE INTEGRATIE	93
5.1. Inleiding	93
5.2. Hogere orde integratieformules	97
5.3. Andere integratieformules met equidistante abscissen	100
5.4. Integratieformules van Gauss	101

JdG, 28 Juli 2005

Jan de Graaf.

1965

AFDELING ALGEMENE WETENSCHAPPEN

Onderafdeling Wiskunde

N U M E R I E K E W I S K U N D E I

Prof. dr. G.W. Veltkamp

Voorjaarssemester 1961

T e c h n i s c h e H o g e s c h o o l E i n d h o v e n

Lichtmans Proefschrift moest Rondschrijven

TECHNISCHE HOGESCHOOL EINDHOVEN

Afdeling Algemene Wetenschappen

Onderafdeling Wiskunde

NUMERIEKE WISKUNDE I

Voorjaarssemester 1961

Prof.dr.G.W.Veltkamp

0. Inleiding

0.1. De numerieke wiskunde houdt zich bezig met het onderzoek van efficiënte methoden om numerieke oplossingen van wiskundige problemen te vinden.
Een oplossing van een wiskundig probleem van practisch belang in de vorm van een formule kan nuttig zijn indien hierdoor het inzicht in het probleem verhelderd wordt, bv. omdat uit de formule de afhankelijkheid van bepaalde parameters duidelijk blijkt. Vaak is echter alleen de oplossing behorend bij één stel waarden van de parameters nodig en het is niet altijd zo dat substitutie van deze waarden in de algemene formule de beste methode is om deze speciale oplossing te verkrijgen. Van andere problemen is de oplossing in de vorm van formules zo gecompliceerd, dat deze niet overzichtelijk is. En ook komt het voor dat in het algemeen alleen bewezen kan worden dàt een oplossing bestaat (bv. uit het ongerijmde). In deze gevallen is het voor de practicus van veel belang om de oplossing numeriek te benaderen.

0.2. Ruw gesproken kunnen de numerieke problemen in twee klassen verdeeld worden :
0.2.1. Problemen waarbij alleen rationale getallen optreden en waarbij een exacte oplossing door oneindig vaak optellen, aftrekken, vermenigvuldigen en delen gevonden kan worden (Vb. : lineaire vergelijkingen met gehele coëfficiënten). Dat ook in deze gevallen nog sprake kan zijn van "the art of computing" (in tegenstelling tot "the science of mathematics") heeft verschillende redenen :

a) Het werken met breuken is vaak tijdrovend en kan tot zeer grote getallen aanleiding geven. Werkt men daarentegen met afgeronde decimale breuken dan ontstaat het probleem welke invloed de afrondingsfouten die bij de afzonderlijke bewerkingen geïntroduceerd worden, op het eindresultaat hebben.

b) Voor sommige problemen bestaan meerdere oplossingsmethoden, welke van wiskundig standpunt equivalent zijn. In numeriek opzicht kunnen deze echter verschillend zijn in verband met

aantal uit te voeren operaties
overzichtelijkheid en eenvoud van de berekening
mogelijkheid van eenvoudige controleberekeningen
invloed van afrondingsfouten.

Natuurlijk zal men vaak een compromis tussen deze factoren moeten zoeken.

Voorbeeld

Moet men de waarde van een polynoom $P(x) = a_o x^n + a_1 x^{n-1} + \ldots + a_n$ (met numeriek gegeven coëfficiënten) berekenen voor een gegeven waarde van x, dan kan men achtereenvolgens bepalen $x^2, x^3, \ldots, x^n$, $a_o x^n, a_1 x^{n-1}, \ldots, a_{n-1} x$ en tenslotte p(x). Dit kost 2n-1 vermenigvuldigingen en n optellingen.

Ook kan men bepalen

$$\begin{aligned} b_o &= a_o \\ b_1 &= b_o x + a_1 \\ b_2 &= b_1 x + a_2 \\ &\ldots\ldots\ldots\ldots \\ b_n &= b_{n-1} x + a_n \end{aligned}$$

en dan is $p(x) = b_n$, daar dit rekenschema correspondeert met de schrijfwijze $p(x) = \ldots((((a_o x + a_1)x + a_2)x + a_3)\ldots$. Nu zijn slechts n vermenigvuldigingen en n optellingen nodig. Bovendien is de procedure veel systematischer.

Door het kleinere aantal bewerkingen zal de invloed van afrondingsfouten in het algemeen kleiner zijn.

0.2.2. Problemen waarbij een exacte oplossing slechts met behulp van infinite processen te geven is.

Hier is het de taak van de numerieke wiskunde, eindige (in verband met de beperkte levensduur van mensen, machines, etc) processen aan te geven waarmee de oplossing met voorgeschreven nauwkeurigheid bepaald kan worden.

Voorbeelden zijn : sommeren van reeksen, uitrekenen van integralen, oplossen van differentiaalvergelijkingen.

0.3. Foutenbronnen en fouten

Men kan vijf soorten van fouten naar hun bronnen onderscheiden.

a) Modelfouten. Deze ontstaan doordat het wiskundige probleem een vereenvoudigde beschrijving van het fysische probleem is. Deze zijn voor de numericus slechts in zoverre van belang dat bij een grof model erg nauwkeurige berekening weinig zin heeft.

b) Beginfouten (initial errors). Vele problemen bevatten parameters die door metingen bepaald moeten worden en dus slechts met beperkte nauwkeurigheid bekend zijn. Voor de numericus gelden dezelfde consequenties als ad a).

c) Afbreekfouten (truncation errors). Deze ontstaan als men een infiniet proces door een finiet proces vervangt. Vervangt men $\log(1 + x)$ door $x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4}$, dan maakt men een afbreekfout (die bv. voor $0 \leqslant x < \frac{1}{10}$ kleiner is dan $0{,}5.10^{-5}$).

Ook als men $\int_{-h}^{h} f(x)dx$ vervangt door $\frac{1}{3}[f(-h) + 4f(o) + f(h)]$ (regel van Simpson). Tenslotte ook bij het afbreken van iteratie processen na eindig veel stappen : zij $a > 0$ en zij de rij getallen $x_0, x_1, \ldots$ bepaald door

$$x_0 = 1,\ x_{n+1} = \tfrac{1}{2}\left[x_n + \frac{a}{x_n}\right],\quad n = 0,1,\ldots \qquad (1)$$

Dan is $\lim_{n\to\infty} x_n = \sqrt{a}$. In de practijk is men gedwongen te stoppen na eindig veel stappen en een x_N als benadering voor $\sqrt{a}$ te beschouwen.

d) Afrondingsfouten. Deze ontstaan doordat men meestal met een vast aantal decimale cijfers werkt. Het resultaat van een vermenigvuldiging wordt dan als regel afgerond, is dus niet exact.

e) Rekenfouten (vergissingen, machinefouten e.d.). Om deze te onderdrukken zijn controleberekeningen zeer gewenst. Bij veel iteratieve processen (bv. het proces (1)) betekent het maken van een niet te ernstige rekenfout alleen dat het langer duurt voordat de gewenste nauwkeurigheid bereikt wordt. Uit dit oogpunt zijn deze processen dus aanbevelenswaardig.

Het is duidelijk dat fouten, die in een bepaald stadium van een rekenproces geïntroduceerd worden, zich in het algemeen in het verdere verloop zullen voortplanten en hun invloed op het eindresultaat zullen hebben. Het is van belang enige indruk van deze invloed te hebben. In het algemeen is dit echter een zeer moeilijk probleem.

0.4. Afgeronde decimale getallen

Het is vaak zinvol een afgerond positief decimaal getal a te schrijven als $a = q.10^p$, waarin $0,1 \leqslant q < 1$. q is dan een echte decimale breuk (met eindig veel cijfers). Schrijft men q met k cijfers achter de komma, dan bedoelt men dat $|10^{-p}a - q| \leqslant \frac{1}{2}10^{-k}$.

Bv. $p = 0$, $q = 0.12$ betekent $0.115 \leqslant a \leqslant 0.125$
$p = 2$, $q = 0.120$ betekent $11.95 \leqslant a \leqslant 12.05$.

Men noemt k het aantal significante cijfers waarin a bekend is.
Deze afspraak heeft het nadeel dat als $1201 \leqslant a \leqslant 1207$, men zou moeten schrijven $a = 0.12.10^3$, waardoor informatie verloren gaat. Derhalve schrijft men ook vaak in zo'n geval $a = 0.120.10^3$ of zelfs $a = 0.1204.10^3$ met de opmerking dat de onnauwkeurigheid van a van de orde van één, resp. enkele "eenheden van de laatste decimaal" is.

Opmerking. Men rond natuurlijk steeds af naar het meest nabijgelegen getal. Bv. geeft 0.405149, afgerond in 5,4,3,2, resp. 1 cijfer(s) 0.40515; 0.4051 ; 0.405 ; 0.41 ; 0.4. Eindigt het af te ronden getal op een 5, dan rondt men zo af dat het laatste cijfer even wordt (dit is beter dan bv. steeds de 5 naar beneden af te ronden daar hierdoor systematische fouten geïntroduceerd worden).
Zo wordt 0.40515 afgerond in 4 cijfers : 0.4052, terwijl 0.405, afgerond in 2 cijfers, wordt : 0.40.

0.5. Literatuur

Er bestaan zeer veel boeken en tijdschriftartikelen over numerieke wiskunde.
Enkele leerboeken zijn :

F.B.Hildebrand, Introduction to numerical analysis, 1956.
K.S.Kunz, Numerical analysis, 1957.
Z.Kopal, Numerical analysis, 1955.
W.F.Milne, Numerical calculus, 1954.
D.R.Hartree, Numerical analysis, 1955.

Meer gericht op speciale onderwerpen en aspecten zijn :
S.A.Householder, Principles of numerical analysis, 1953.
A.Ralston and H.S.Wilf, Mathematical methods for digital computers, 1960.
W.E.Milne, Numerical solution of differential equations, 1953.
E.Bodewig, Matrix Calculus, 1956.
V.N.Faddeeva, Computational methods of linear algebra, 1959.
P.S.Dwyer, Linear computations, 1951.
E.Durand, Solutions numériques des équations algébriques, 1960.

Enkele tijdschriften zijn :
Mathematics of Computation (heette voor Math.Tables and other aids to computation = MTAC).
Journ. Soc. Industrial Appl. Math. (SIAM).
Journ. Assoc. Computing Machinery.
Numerische Mathematik.
Journ. of Research Nat. Bur. of Standards.

1. Het oplossen van vergelijkingen

1.1. De meeste hier te behandelen methoden zijn iteratie-methoden : men begint met een zekere schatting voor de oplossing, voert een zekere bewerking uit waardoor een betere benadering van de oplossing verkregen wordt. Dit proces wordt herhaald tot een voldoende nauwkeurigheid bereikt is. Het succes van de methode hangt vaak af van de juistheid van de eerste schatting. Voor het vinden hiervan bestaan geen algemene methoden. Vaak kan men door een ruw analytisch onderzoek of het uitrekenen van een aantal functiewaarden bruikbare beginschattingen vinden.

1.2. Successieve substitutie

Zij de vergelijking gegeven in de vorm

$$x = f(x) \qquad (1)$$

waarin f een continue functie is die voor een willekeurig gegeven waarde van x betrekkelijk eenvoudig te berekenen is.

Zij x_o een beginschatting voor de oplossing.
Bereken achtereenvolgens

$$x_1 = f(x_0)$$
$$x_2 = f(x_1)$$
$$\cdots\cdots\cdots$$
$$x_{n+1} = f(x_n).$$

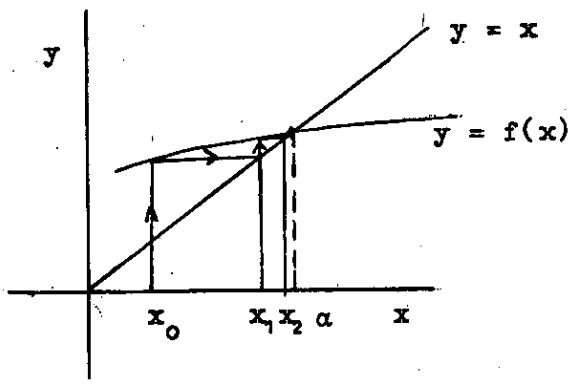
Stelling 1. Als de rij $\{x_n\}$ een limiet α heeft, dan is α een oplossing van (1).

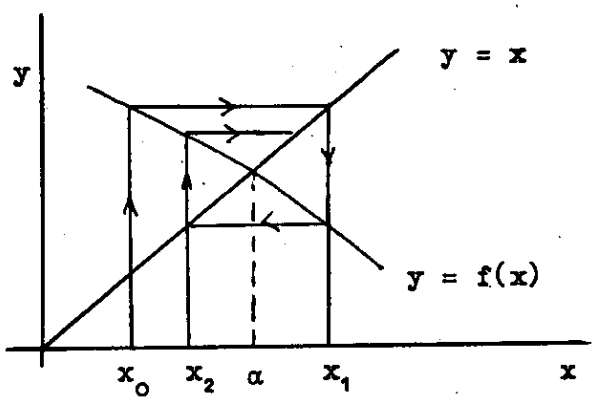
Bewijs

Daar f continu is, geldt

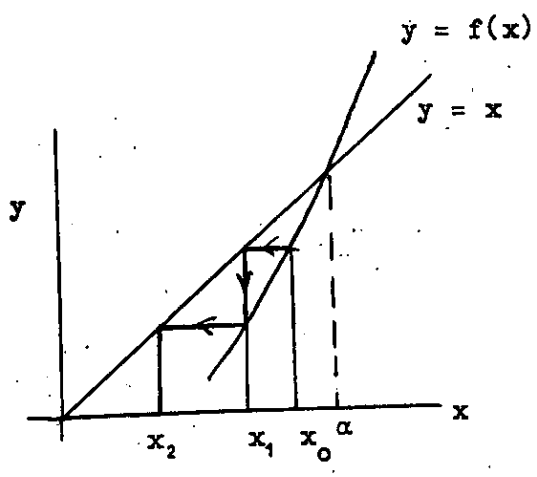
$$\alpha = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} f(x_n) = f(\lim_{n\to\infty} x_n) = f(\alpha).$$

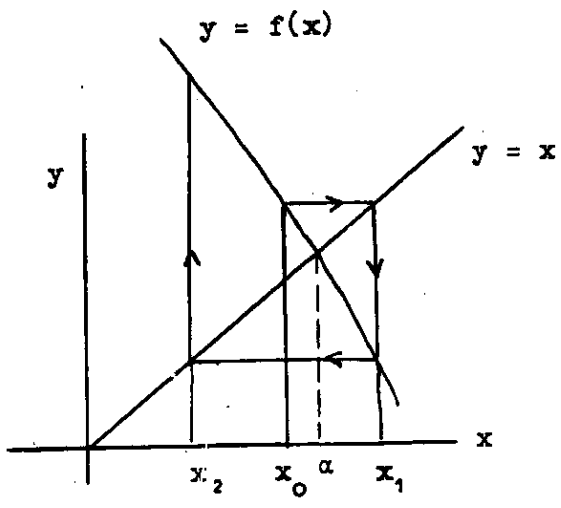
De gang van de iteratie kan eenvoudig afgelezen worden uit de plaatjes :





Vergelijking van deze plaatjes met de onderstaande suggereert dat het proces convergeert als $|f'(x)| < 1$ en divergeert als $|f'(x)| > 1$ in de omgeving van α





(1) $x = f(x)$

Verder blijkt dat de rij $\{x_n\}$ monotoon is als $f'(x) > 0$ en oscillerend om α is als $f'(x) < 0$.

Stelling 2. Zij α een oplossing van (1). Zij f differentieerbaar in een interval I [$|x - \alpha| < \rho$] en zij hier $|(f'(x)| \leqslant \lambda < 1$. Dan geldt voor iedere beginschatting x_o die in I ligt

1) alle geïtereerde x_n liggen in I
2) $|x_n - \alpha| \leqslant \lambda |x_{n-1} - \alpha|$ (2)
3) $\lim x_n = \alpha$.

Bewijs. Daar $x_o \in I$ geldt volgens de middelwaardestelling $x_1 - \alpha = f(x_o) - \alpha = f(x_o) - f(\alpha) = (x_o - \alpha) f'(\xi)$ met ξ tussen x_o en α, dus zeker in I.
Derhalve is $|x_1 - \alpha| \leqslant \lambda |x_o - \alpha|$.
Daar $\lambda < 1$ volgt hieruit 1) en 2) voor n = 1. Door inductie volgt dat 1) en 2) voor alle n gelden. En uit (2) volgt dan eenvoudig

$$|x_n - \alpha| \leqslant \lambda^n |x_o - \alpha| . \qquad (3)$$

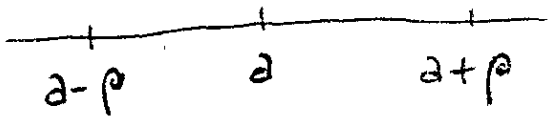
en dus $\lim x_n = \alpha$, daar $\lambda < 1$.

Opmerkingen

1. Uit (3) volgt een schatting van de fout na n iteraties. Men ziet dat deze minstens naar nul gaat als de termen van een meetkundige reeks met reden λ.
Men noemt dit een proces van de eerste orde en λ de convergentiefactor.
Men noemt

$$\lim_{n\to\infty} \frac{x_n - \alpha}{x_{n-1} - \alpha} = \lim_{n\to\infty} \frac{f(x_{n-1}) - f(\alpha)}{x_{n-1} - \alpha} = f'(\alpha) \qquad (4)$$

de asymptotische convergentiefactor.
Is f' continu in α en $|f'(\alpha)| < 1$ dan volgt hieruit direct de convergentie mits x_o dicht genoeg bij α ligt. In het algemeen is het enigzins gevaarlijk om alleen naar de asymptotische convergentiefactor te kijken : pas vlak bij α geeft deze een maat voor de snelheid van de convergentie.



2. Een fraaiere stelling waarin de existentie van een oplossing α niet a priori verondersteld wordt is :
Zij voor zekere a, ρ en λ $|f'(x)| \leqslant \lambda < 1$ als $|x - a| < \rho$ en zij bovendien $|f(a) - a| \leqslant \rho(1 - \lambda)$.
Dan heeft de vergelijking (1) in het interval $|x - a| < \rho$ precies één oplossing α en voor iedere beginschatting x_o die voldoet aan $|x_o - a| < \rho$ geldt $\lim x_n = \alpha$.
(Voor het bewijs van een analoge stelling vgl. Householder, p. 119-120).

3. Wat is de invloed van een enigzins onnauwkeurige bepaling van $f(x_n)$ (bv. t.g.v. afrondings- of afbreekfouten) ? Dan geldt $\bar{x}_{n+1} = f(\bar{x}_n) + \delta_n$, waarin $\bar{x}_n$ de reeds onnauwkeurige n^e iterand is en δ_n de fout is die gemaakt wordt bij de berekening van $f(\bar{x}_n)$. Men kan niet verwachten dat de rij $(\bar{x}_n)$ een limiet heeft. Maar wel geldt, dat als voor alle n $|\delta_n| \leqslant \delta$ en x_o dicht genoeg bij α ligt, dat

$$|\bar{x}_n - \alpha| < \lambda^n |x_o - \alpha| + \frac{\delta}{1 - \lambda}.$$

Dit is voor de practijk voldoende : wil men α tot op een afstand kleiner dan ε benaderen, dan moet men zorgen dat de fout in de berekening van $f(x_n)$ niet groter is dan $(1 - \lambda)\varepsilon$. Men noemt op grond van deze eigenschap het proces <u>stabiel</u> : ondanks een mogelijke ongunstige accumulatie van de fouten veroorzaakt door onnauwkeurige berekening van $f(x_n)$ komt men tenslotte in de buurt van α.

1.2.1. <u>Processen van hogere orde</u>

Het is duidelijk dat de convergentie van het iteratieproces snel is indien in de omgeving van α $f'(x)$ dicht bij nul is. Stel eens dat $f'(\alpha) = 0$ en dat in een omgeving van α

$$|f(x) - \alpha| = |f(x) - f(\alpha)| \leqslant A\,|x - \alpha|^p, \text{ met } p > 1. \quad (1)$$

Dan geldt hier

$$|x_{n+1} - \alpha| = |f(x_n) - \alpha| \leqslant A|x_n - \alpha|^p.$$

Men noemt in dit geval de orde van het proces p.

Zij $|x_n - \alpha| \leqslant \frac{1}{2}.10^{-m}$ (m correcte cijfers achter de komma).
Dan is $|x_{n+1} - \alpha| \leqslant 2^{-p}.A.10^{-pm}$, dus (tenzij A erg groot is) ca. pm correcte cijfers. Voor bv. p = 2 wordt het aantal correcte cijfers

dus bij iedere iteratie ongeveer verdubbeld. Hieruit blijkt de kracht van iteratiemethoden van hogere orde.
Ook is de convergentie steeds verzekerd, <u>althans</u> als x_0 dicht genoeg bij α ligt.

<u>Stelling</u>. Zij voor $|x - \alpha| < \rho$ $\quad |f(x) - \alpha| \leqslant A|x - \alpha|^2$.
Zij $|x_0 - \alpha| < \min(\rho, A^{-1})$. (2)
Dan geldt $\lim x_n = \alpha$.

<u>Bewijs</u>. We hebben

$$|x_1 - \alpha| = |f(x_0) - \alpha| \leqslant A|x_0 - \alpha|^2. \qquad (3)$$

Dus $\quad |x_1 - \alpha| < A|x_0 - \alpha| . \rho < \rho$

en ook $\quad |x_1 - \alpha| < A.A^{-2} = A^{-1}$.

Dus x_1 voldoet ook aan (2). En bovendien volgt uit (3)

$$A|x_1 - \alpha| \leqslant (A|x_0 - \alpha|)^2.$$

Door inductie leidt men af dat ook x_n aan (2) voldoet en dat $A|x_n - \alpha| \leqslant (A|x_n - \alpha|)^2 \leqslant (A|x_{n-1}|)^{2n}$, waaruit volgt dat $x_n \to \alpha$, daar $A|x_0 - \alpha| < 1$.

<u>Opmerkingen</u>

1. Voor p-de orde processen geldt natuurlijk een analoge stelling.
2. Geldt in een omgeving van α dat $f(x) = \alpha + A(x - \alpha)^p + \ldots$ *), dan blijkt dat

$$\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{(x_n - \alpha)^p} = A.$$

Deze relatie die het asymptotische gedrag van de rij $\{x_n\}$ bepaalt, komt in plaats van de relatie (4) uit 1.2.

3. In de practijk is het niet altijd eenvoudig om x_0 zo te kiezen dat aan (2) voldaan is. Vaak is er echter een groot gebied om α waar $|f'(x)| < 1$. Start men binnen dit gebied dan heeft men aanvankelijk in ieder geval 1e-orde convergentie (lineaire convergentie), op een gegeven moment (als $|x_n - \alpha| < A^{-1}$ geworden is) zet de kwadratische (2e-orde) convergentie in.

---

*) Met deze notatie wordt bedoeld : $\lim_{x\to\alpha} \frac{f(x) - \alpha}{(x - \alpha)^p} = A.$

1.2.2. Het δ^2-proces van Aitken

Zij $\{x_n\}$ een rij iteranden, verkregen uit een iteratieproces $x_{n+1} = f(x_n)$. (1)

Als x_n dicht bij een wortel α van $x = f(x)$ ligt dan geldt (als f continu differentieerbaar is)

$$x_{n+1} - \alpha = (x_n - \alpha)\,[f'(\alpha) + \varepsilon_n]\,,$$
$$x_{n+2} - \alpha = (x_{n+1} - \alpha)\,[f'(\alpha) + \varepsilon_{n+1}]\,,$$

waarin ε_n en ε_{n+1} klein zijn. Is $f'(\alpha) \neq 0$ en verwaarlozen we ε_n en ε_{n+1}, dan zou hieruit volgen

$$\frac{x_{n+1} - \alpha}{x_n - \alpha} = \frac{x_{n+2} - \alpha}{x_{n+1} - \alpha}$$

waaruit volgt

$$\alpha = \frac{x_n x_{n+2} - x_{n+1}^2}{x_{n+2} - 2x_{n+1} + x_n} =$$

$$= x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n} = x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n}$$

$$= x_{n+2} - \frac{(x_{n+2} - x_{n+1})^2}{x_{n+2} - 2x_{n+1} + x_n} = x_{n+2} - \frac{(\nabla x_{n+2})^2}{\nabla^2 x_{n+2}}\,,$$

waarin Δ en ∇ de symbolen voor voorwaartse, resp. achterwaartse differentie zijn.

Deze formules zouden exact zijn indien $\varepsilon_n = \varepsilon_{n+1} = 0$. Men kan zich afvragen of in het algemeen de uit x_n, x_{n+1} en x_{n+2} verkregen waarde

$$\bar{x} = x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n} = x_{n+2} - \frac{(\nabla x_{n+2})^2}{\nabla^2 x_{n+2}} \qquad (2)$$

een betere benadering is dan x_n, x_{n+1} en x_{n+2}. Dit blijkt inderdaad het geval te zijn. Men kan nl. het volgende resultaat bewijzen.

Zij in een omgeving van α $f(x) = \alpha + A(x - \alpha) + B(x - \alpha)^p + \ldots$, met $p > 1$, $B \neq 0$. Beschouw, uitgaande van een x_o in een omgeving van α de rij $\{x_n\}$, gedefinieerd door

$$
\begin{aligned}
&\gamma_n \\
&x'_n = f(x_n) \\
&x''_n = f(x'_n) \\
&x_{n+1} = x_n - \frac{(x'_n - x_n)^2}{x''_n - 2x'_n + x_n} = x''_n - \frac{(x''_n - x'_n)^2}{x''_n - 2x'_n + x_n}
\end{aligned} \qquad (3)
$$

Dan geldt

1) Als $A \neq 0$ of 1 : $\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{(x_n - \alpha)^p} = \frac{(A - A^p)B}{(A-1)^2}$

2) Als $A = 1$: $\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{x_n - \alpha} = 1 - \frac{1}{p}$

3) Als $A = 0$: $\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{(x_n - \alpha)^{2p-1}} = - B^2$.

De betekenis hiervan is de volgende :
x'_n en x''_n worden verkregen door, uitgaande van x_n, de oorspronkelijke iteratieformule (1) eenmaal, resp. tweemaal toe te passen. Vervolgens wordt uit het rijtje x_n, x'_n, x''_n met behulp van (2) x_{n+1} verkregen. Nu gaat men, met x_{n+1} als startwaarde, weer twee maal de iteratie (1) uitvoeren. Etc.
Het resultaat is dat, als $A \neq 0$ (zodat het proces (1) van de eerste orde is, convergent als $|A| < 1$, divergent als $|A| > 1$), het proces (3) de orde p heeft (ook als het proces (1) divergeert!), tenzij $A = 1$, in welk geval (3) toch altijd nog een convergent proces van de orde 1 is (met asymptotische convergentiefactor $1 - 1/p$).
Is $A = 0$, zodat het proces (1) de orde p heeft, dan heeft het proces (3) de orde $2p - 1$. Dit is ongunstiger, daar een stap van (3) correspondeert met twee stappen van (1) en $2p - 1 < p^2$ voor $p \neq 1$. Wel blijft in dit geval de orde van (3) groter dan 1.

Voorbeeld. Zij $f(x) = \frac{1}{10}(x^3 + 9)$, $x_o = 1.5$. Dan vindt men met het proces (1)

$$x_0 = 1.5$$
$$x_1 = 1.2375$$
$$x_2 = 1.0895$$
$$x_3 = 1.0293$$
$$x_4 = 1.0091$$

Pas op de laatste drie waarden het δ^2-proces toe.

$$\begin{array}{ll} \bar{x}_0 = 1.0895 & -602 \\ \bar{x}'_0 = 1.0293 & -202 \quad +400\,. \\ \bar{x}''_0 = 1.0091 & \end{array}$$

Dan is $\bar{x}_1 = 1.0091 - \frac{(0.0202)^2}{0.0400} = 0.9989$.

Deze waarde is inderdaad aanzienlijk beter dan x_4.
Zet nu het δ^2-proces voort, uitgaande van $\bar{x}_1$.

$$\begin{array}{ll} \bar{x}_1 = 0.9989000 & +7704 \\ \bar{x}'_1 = 0.9996704 & +2308 \quad -5396\,. \\ \bar{x}''_1 = 0.9999012 & \\ \bar{x}_2 = \underline{0.9999999} & \end{array}$$

Opmerking. Men mag niet te vroeg met het δ^2-proces beginnen. Past men het toe op x_0, x_1, x_2 dan vindt men $\bar{x}_1 = 0.8982$. Deze benadering is slechter dan x_2. Globaal gezegd : het δ^2-proces wordt pas zinvol als het lineaire proces inderdaad redelijk lineair geworden is.

Voorbeeld 2. $f(x) = x^3 - 3x + 3$, $x_0 = 1.1$.
Proces (1) levert

$$x_0 = 1.1$$
$$x_1 = 1.031$$
$$x_2 = 1.0030.$$

Toepassing van het δ^2-proces levert $\bar{x}_1 = 0.984$, dat is slechter dan x_2 (en beter dan x_1). Dit klopt : het proces (1) heeft hier de orde twee (omdat $f'(1) = 0$).

1.3. De vergelijking F(x) = 0

Zij gezocht een wortel α van de vergelijking $F(x) = 0$. We trachten deze vergelijking in de vorm $x = f(x)$ te brengen, waarbij f zo is dat $|f'(\alpha)|$ dicht bij nul is. Dit kan bv. door te nemen

$$f(x) = x - \varphi(x)\, F(x), \tag{1}$$

waarbij $\varphi(x) \neq 0$ (althans in een omgeving van α). Dan is $f'(x) = 1 - \varphi'(x)\, F(x) - \varphi(x)\, F'(x)$ en met name dus $f'(\alpha) = 1 - \varphi(\alpha)\, F'(\alpha)$ (daar $F(\alpha) = 0$). De convergentie van het iteratieproces

$$x_{n+1} = f(x_n) = x_n - \varphi(x_n)\, F(x_n) \tag{2}$$

is dus verzekerd (bij voldoende gladde φ en F en als x_0 dicht genoeg bij α ligt) indien $|1 - \varphi(\alpha)\, F'(\alpha)| < 1$. En het proces (2) is van de tweede orde indien $\varphi(\alpha)\, F'(\alpha) = 1$ (en φ en F voldoende glad zijn).

1.3.1. De koorde-methode

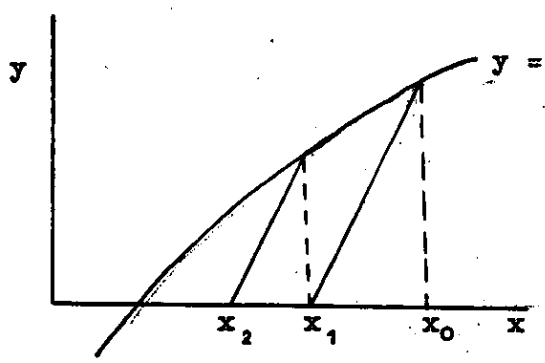
Kies $\varphi(x) = \frac{1}{m}$ met m zo dat $\left|1 - \frac{F'(\alpha)}{m}\right| < 1$.

Is $F'(\alpha) > 0$ dan betekent dit dat $\frac{1}{2} F'(\alpha) < m < \infty$,

is $F'(\alpha) < 0$ dan moet $-\infty < m < -\frac{1}{2}|F'(\alpha)|$.

Het proces
$$x_{n+1} = x_n - \frac{1}{m} F(x_n) \tag{1}$$

is van de eerste orde (tenzij $m = F'(\alpha)$ - maar $F'(\alpha)$ is meestal nog onbekend!) en de asymptotische convergentiefactor is $1 - \frac{1}{m} F'(\alpha)$.

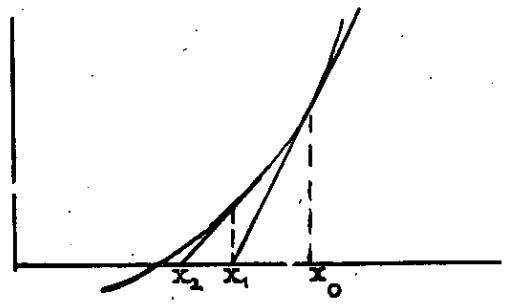


Meetkundig betekent de formule (1) dat men door het punt $(x_n, F(x_n))$ een rechte met richtingscoëfficiënt m trekt (vergelijking $y = F(x_n) + m(x - x_n)$) en het snijpunt hiervan met de x-as als x_{n+1} neemt.

1.3.2. De iteratiemethode van Newton-Raphson

Kies $\varphi(x) \triangleq \frac{1}{F'(x)}$. De iteratieformule wordt dan

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)} . \qquad (1)$$



Meetkundig betekent dit dat men in het punt $(x_n, F(x_n))$ de raaklijn aan de kromme $y = F(x)$ trekt (vergelijking : $y = F(x_n) + (x-x_n)F'(x_n)$) en het snijpunt hiervan met de x-as als x_{n+1} neemt.

Het is uit 1.3 duidelijk dat dit proces in het algemeen van de tweede orde is, althans als $F'(\alpha) \neq 0$. Om de convergentie iets nader te onderzoeken veronderstellen we dat

$$F(x) = A(x-\alpha) + B(x-\alpha)^p + \ldots$$
$$F'(x) = A + pB(x-\alpha)^{p-1} + \ldots$$

met $p > 1$ *)

Korter p.7 geschreven dictaat.

Hiermee volgt uit (1)

$$x_{n+1} - \alpha = \frac{(p-1)\,B(x_n-\alpha)^p + \ldots}{A + pB(x_n-\alpha)^{p-1} + \ldots} .$$

Is $A = f'(\alpha) \neq 0$, dan is er dus convergentie van de orde p (meestal is $p = 2$, nl. als $F''(\alpha) \neq 0$), is $f'(\alpha) = 0$, dan is het proces van de eerste orde, met asymptotische convergentiefactor $1 - \frac{1}{p}$ (dus wel steeds convergent).

<u>Voorbeelden</u>

1. $F(x) = x^k - a$, k geheel ≥ 2. Dan wordt $\alpha = a^{1/k}$ en

$$x_{n+1} = x_n - \frac{x_n^k - a}{kx_n^{k-1}} = \frac{1}{k}\left[(k-1)x_n + \frac{a}{x_n^{k-1}}\right].$$

Dit is een zeer gebruikelijke methode om $\sqrt[k]{a}$ uit te rekenen.

*) met de ... wordt bedoeld : "termen van hogere orde". Nauwkeuriger zou zijn

$$\lim_{x \to a} \frac{F(x) - A(x-\alpha)}{(x-\alpha)^p} = B, \text{ etc.}$$

Voor k = 2 wordt de formule $x_{n+1} = \frac{1}{2}[x_n + \frac{a}{x_n}]$.
Deze formule was al 100 jaar v.Chr. bekend (Heron).

Voor k = -1 krijgen we $x_{n+1} = x_n(2 - ax_n)$. Met deze formule kan men dus "delen zonder te delen". Dit proces wordt wel gebruikt bij automatische rekenmachines die geen ingebouwde deling hebben.

2. $F(x) = 1 - \frac{1}{ax^k}$; $\alpha = a^{-1/k}$; $x_{n+1} = \frac{1}{k} \cdot x_n(k + 1 - ax_n^k)$.
Deze formule is ook geschikt voor machines zonder ingebouwde deling (de deling door k wordt vervangen door vermenigvuldiging met de constante k^{-1}).

3. $F(x) = x^{k-m} - ax^{-m}$; $\alpha = a^{1/k}$.
Bewijs dat het proces de orde drie heeft indien $m = \frac{1}{2}(k - 1)$. De formule wordt dan

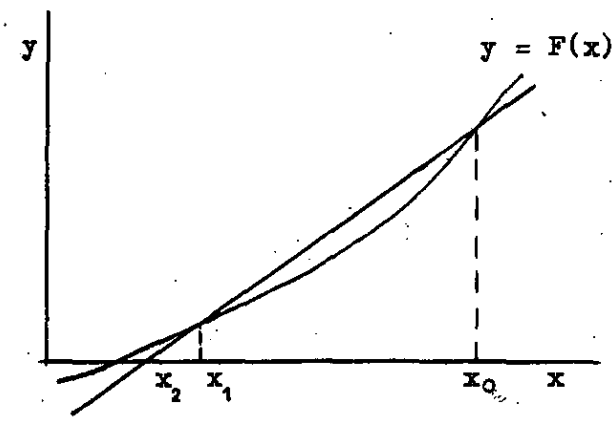
$$x_{n+1} = x_n \cdot \frac{(k - 1)x_n^k + (k + 1)a}{(k + 1)x_n^k + (k - 1)a}.$$

4. $F(x) = x^3 - 4x$. Ga na (met behulp van plaatjes) dat in dit geval het proces convergeert naar de wortel $\alpha = 2$ indien $x_o > \frac{2}{3}\sqrt{3}$ naar de wortel $\alpha = -2$ indien $x_o < -\frac{2}{3}\sqrt{3}$ en naar de wortel $\alpha = 0$ indien $|x_o| < \frac{2}{5}\sqrt{5}$.
Tracht ook de overige gevallen te overzien!

1.4. Regula falsi

Het is natuurlijk niet noodzakelijk om zich te beperken tot iteratieformules van de vorm $x_{n+1} = f(x_n)$. De zogenaamde regula falsi leidt tot een formule van de vorm $x_{n+1} = f(x_n, x_{n-1})$.
Zij x_o en x_1 twee (niet gelijke) schattingen voor een wortel α van $F(x) = 0$. Neem nu de verbindingslijn

$$\frac{y - F(x_1)}{x - x_1} = \frac{F(x_o) - F(x_1)}{x_o - x_1}$$



van de punten $(x_o, F(x_o))$ en $(x_1, F(x_1))$ en beschouw het snijpunt van deze rechte met de x-as als nieuwe benadering :

$$x_2 = x_1 - F(x_1) \cdot \frac{x_o - x_1}{F(x_o) - F(x_1)}. \quad (1)$$

Ga nu verder met x_1 en x_2. Enz. De algemene iteratie formule wordt dan

$$x_{n+1} = x_n - F(x_n) \cdot \frac{x_{n-1} - x_n}{F(x_{n-1}) - F(x_n)} .$$

Men kan bewijzen dat (als $F'(\alpha) \neq 0$ en $F''(\alpha) \neq 0$) de orde van dit proces gelijk is aan $\frac{1}{2}(1 + \sqrt{5}) = 1{,}618\ldots$. De convergentie is dus minder snel dan bij Newton. Maar daar staat tegenover dat men $F'(x_n)$ niet behoeft te berekenen, hetgeen bij gecompliceerde F een wezenlijk voordeel is.

Opmerking. Vaak past men de regula falsi iets anders toe. Men start met getallen x_0 en x_1 zodanig dat $F(x_0)$ en $F(x_1)$ verschillend teken hebben (er ligt dan dus minstens één wortel tussen x_0 en x_1). Men bepaalt nu x_2 met (1) (x_2 ligt nu stellig ook tussen x_0 en x_1). Hebben nu $F(x_1)$ en $F(x_2)$ verschillend teken dan gaat men (als boven) verder met x_1 en x_2. Hebben echter $F(x_1)$ en $F(x_2)$ hetzelfde teken (zodat $F(x_0)$ en $F(x_2)$ verschillend teken hebben) dan gaat men verder met x_0 en x_2.
Dit proces is in het algemeen slechts van de eerste orde. Maar bij willekeurige x_0 en x_1 (zodanig dat $F(x_0).F(x_1) < 0$) convergeert het proces altijd naar (een van) de wortel(s) tussen x_0 en x_1 . Bovendien kan men tegen het eind de convergentie versnellen met het δ^2-proces. Een nadeel is dat men wortels met even multipliciteit niet vinden kan (waarom niet?).

1.5. Stelsels vergelijkingen

Naast één vergelijking met één onbekende ontmoet men ook stelsels van k vergelijkingen met k onbekenden ; bv.

$$\begin{aligned} F_1(x_1, x_2, \ldots, x_k) &= 0 \\ F_2(x_1, x_2, \ldots, x_k) &= 0 \\ &\ldots\ldots\ldots\ldots \\ F_k(x_1, x_2, \ldots, x_k) &= 0 \end{aligned}$$

We spreken van lineaire vergelijkingen indien de functies F_i (eventueel inhomogeen) lineair zijn in $x_1, \ldots, x_n$, dus als $F_i(x_1, \ldots, x_k) = \sum_{j=1}^{k} A_{ij} x_j - b_i$.

Dit type vergelijkingen wordt in hoofdstuk 2 uitvoerig besproken.

We beperken ons hier verder tot k = 2 en schrijven de vergelijkingen als

$$F(x,y) = 0,\ G(x,y) = 0. \tag{1}$$

Merk op dat $z = F(x,y)$ een oppervlak in R_3 voorstelt en dat $F(x,y) = 0$ de snijkromme van dit oppervlak met het vlak $z = 0$ is. Analoog $G(x,y) = 0$. Zij $x = \alpha,\ y = \beta$ een oplossing van (1). Dan wordt de raaklijn in dit punt aan $F(x,y) = 0$ gegeven door $F_x(\alpha,\beta) + \frac{dy}{dx} F_y(\alpha,\beta) = 0$ en die aan $G(x,y) = 0$ door $G_x(\alpha,\beta) + \frac{dy}{dx} G_y(\alpha,\beta) = 0$. Indien

$$\begin{vmatrix} F_x(\alpha,\beta) & F_y(\alpha,\beta) \\ G_x(\alpha,\beta) & G_y(\alpha,\beta) \end{vmatrix} = \frac{\partial(F,G)}{\partial(x,y)}\Big|_{(\alpha,\beta)} = 0, \tag{2}$$

vallen de raaklijnen samen. Dit geval, dat moeilijker te behandelen is, sluiten we verder uit.

1.5.1. Iteratiemethode

Stel de vergelijkingen (1) geschreven als

$$x = f(x,y),\quad y = g(x,y). \tag{3}$$

Men kan dit, uitgaande van (1) bv. verkrijgen door te nemen

$$\begin{aligned} f(x,y) &= x - A(x,y)\ F(x,y) - B(x,y)\ G(x,y) \\ g(x,y) &= y - C(x,y)\ F(x,y) - D(x,y)\ G(x,y) \end{aligned} \tag{4}$$

waarin de functies A, B, C en D zo moeten zijn dat

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} \neq 0$$

(zodat uit $\alpha = f(\alpha,\beta)$, $B = g(\alpha,\beta)$ volgt dat $F(\alpha,\beta) = G(\alpha,\beta) = 0$).
Kies nu een beginschatting (x_0,y_0) en bepaal volgende benaderingen (x_n, y_n) $(n = 1,2,\ldots)$ door

$$x_{n+1} = f(x_n, y_n),\quad y_{n+1} = g(x_n, y_n). \tag{5}$$

Het is duidelijk dat als de rijen $\{x_n\}$ en $\{y_n\}$ limieten α, resp. β hebben (α,β) een oplossing van (2) is (als f en g continu zijn).
Men kan voorwaarden afleiden die de convergentie van het proces (5) verzekeren. Bij voorbeeld geldt :

Stelling. Als (α,β) een oplossing van (3) is en voor $|x - \alpha| < \rho$, $|y - \beta| < \rho$

$$|f_x| + |f_y| \leqslant \lambda, \quad |g_x| + |g_y| \leqslant \lambda,$$

met $\lambda < 1$, dan convergeert het proces (4) voor iedere beginschatting (x_o, y_o) die voldoet aan $|x_o - \alpha| < \rho$, $|y_o - \beta| < \rho$.
Het bewijs van deze stelling is analoog aan dat uit 1.2. Voorts blijkt dat $|x_n - \alpha| \leqslant \lambda^n |x_o - \alpha|$, $|y_n - \beta| \leqslant \lambda^n |y_o - \beta|$. Het proces is dus (in het algemeen) van de eerste orde.

1.5.2. De methode van Newton-Raphson

We gaan uit van de vergelijkingen (1). Zij (α,β) een oplossing en (x_o, y_o) een naburig punt. Dan kan men schrijven (Taylor-Reeks)

$$0 = F(\alpha,\beta) = F(x_o,y_o) + (\alpha - x_o)\, F_x(x_o,y_o) + (\beta - y_o)\, F_y(x_o,y_o) + \dots$$
$$0 = G(\alpha,\beta) = G(x_o,y_o) + (\alpha - x_o)\, G_x(x_o,y_o) + (\beta - y_o)\, G_y(x_o,y_o) + \dots$$

Of ook

$$\left.\begin{aligned}(\alpha - x_o)\, F_x(x_o,y_o) + (\beta - y_o)\, F_y(x_o,y_o) &= - F(x_o,y_o) + \dots \\ (\alpha - x_o)\, G_x(x_o,y_o) + (\beta - y_o)\, G_y(x_o,y_o) &= - G(x_o,y_o) + \dots\end{aligned}\right\}$$

Los hieruit α en β op :

$$\alpha = x_o - \left.\frac{F.G_y - G.F_y}{F_x G_y - G_x F_y}\right|_{x_o,y_o} + \dots$$

$$\beta = y_o - \left.\frac{F_x G - G_x F}{F_x G_y - G_x F_y}\right|_{x_o,y_o} + \dots$$

Hier wordt overal met ... bedoeld : "termen van hogere orde".
Verwaarloos nu deze termen van hogere orde. Dan gelden de laatste formules niet meer exact, doch gaan over in iteratie formules die algemeen luiden

$$\left.\begin{aligned}x_{n+1} &= x_n - \left.\frac{F\,G_y - G\,F_y}{F_x G_y - G_x F_y}\right|_{x_n,y_n} \\ y_{n+1} &= y_n - \left.\frac{F_x G - G_x F}{F_x G_y - G_x F_y}\right|_{x_n,y_n}\end{aligned}\right\} \qquad (6)$$

Deze formules corresponderen met die uit 1.4. (en gaan erin over indien men neemt $F(x,y) = F(x) - y$, $G(x,y) = y - ga\ na!$). Ook de meetkundige interpretatie is analoog.
Immers $z = F(x_o,y_o) + (x - x_o) F_x(x_o,y_o) + (y - y_o) F_y(x_o,y_o)$ is het raakvlak aan $z = F(x,y)$ in het punt $(x_o,y_o, F(x_o,y_o))$. Men neemt dus de raakvlakken aan $z = F(x,y)$, resp. $z = G(x,y)$ in de punten $(x_o,y_o, F(x_o,y_o))$, resp. $(x_o,y_o, G(x_o,y_o))$, bepaalt de snijlijn van deze raakvlakken met het vlak $z = 0$ en beschouwt het snijpunt (x_1,y_1) van deze snijlijnen als nieuwe benaderingen.
Men kan weer bewijzen dat dit proces in het algemeen de orde twee heeft.

Opmerkingen

1. Het is duidelijk dat moeilijkheden ontstaan indien $F_xG_y - G_xF_y = \frac{\partial(F,G)}{\partial(x,y)} = 0$ in een der punten (x_n,y_n). Vergelijking met formule (2) uit 1.5 leert dat dit (in de buurt van (α,β) en als F en G continu differentieerbaar zijn) niet kan gebeuren als de kromme $F(x,y) = 0$ en $G(x,y) = 0$ elkaar in (α,β) niet raken.
2. Laat zien dat de formules (6) corresponderen met die uit (5) indien men in (4) voor de matrix $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ neemt de inverse van de matrix $\begin{pmatrix} F_x & F_y \\ G_x & G_y \end{pmatrix}$.

1.6. Algebraïsche vergelijkingen

We beschouwen nu vergelijkingen van de vorm $p(x) = 0$, waarin p een polynoom is :

$$p(x) = a_o x^n + a_1 x^{n-1} + \ldots + a_{n-1}x + a_n.$$

We veronderstellen $a_o \neq 0$. Het polynoom heeft dan de graad n.
Uit de zg. hoofdstelling van de algebra (die strikt genomen niet tot de algebra maar tot de analyse behoort) volgt dat een dergelijke vergelijking steeds n oplossingen $\alpha_1,\ldots,\alpha_n$ heeft. Onder deze oplossingen (ook wortels genaamd) kunnen gelijke voorkomen. Nauwkeuriger geldt : er zijn (op de volgorde na eenduidig bepaalde) getallen $\alpha_1, \alpha_2, \ldots, \alpha_n$ zodanig dat

$$p(x) = a_o(x - \alpha_1)(x - \alpha_2)\ldots(x - \alpha_n).$$

Is $\alpha_1 = \alpha_2 = \ldots = \alpha_k \neq \alpha_j$ voor $j > k$ dan heet α_1 een k-voudige wortel. Etc. De wortels behoeven niet reeel te zijn. Zijn $a_o, a_1, \ldots, a_n$ <u>reeel</u> (wat we verder veronderstellen) dan is met α_j ook $\overline{\alpha_j}$ (de complex geconjugeerde) een wortel (dit volgt uit het feit dat $F(\alpha_j) = 0$ impliceert dat $F(\overline{\alpha_j}) = \overline{F(\alpha_j)} = 0$), zodat eventuele complexe wortels steeds als paren toegevoegd complexe wortels voorkomen.

1.6.1. <u>Rekenmethoden voor polynomen</u>

1.6.1.1. <u>Schema van Horner</u>

Daar $$p(x) = a_o x^n + a_1 x^{n-1} + \ldots + a_n =$$
$$= (\ldots(((a_o x + a_1)x + a_2)x + a_3)\ldots)x + a_n \qquad (1)$$

kan men voor een gegeven α $p(\alpha)$ voordelig als volgt berekenen.

Stel
$$b_o = a_o$$
$$b_1 = a_1 + \alpha b_o$$
$$b_2 = a_2 + \alpha b_1$$
$$\ldots\ldots\ldots\ldots$$
$$b_n = a_n + \alpha b_{n-1}$$

Dan volgt direct uit (1) dat $b_n = p(\alpha)$ *).

Men noteert de getallen $a_o, \ldots, a_n$, en $b_o, \ldots, b_n$ meestal als volgt :

a_o		
a_1	b_o	$(= a_o)$
a_2	b_1	$(= a_1 + \alpha b_o)$
.....		
.....		
a_n	b_{n-1}	$(= a_{n-1} + \alpha b_{n-2})$
	b_n	$(= a_n + \alpha b_{n-1})$.

Dit heet het <u>schema van Horner.</u>

---

*) Een nettere manier om dit te bewijzen is, aan te tonen (door volledige inductie, uitgaande van $b_o = a_o$ en gebruikmakend van $b_i = a_i + \alpha b_{i-1}$ $(1 \leq i \leq n)$) dat voor $0 \leq j \leq n$ geldt

$$b_j = \sum_{i=o}^{j} a_i x^{j-i} .$$

Dit schema levert nog meer. Zij gevraagd de deling $p(x) : (x - \alpha)$ uit te voeren. Als $p(\alpha) \neq 0$ dan gaat de deling niet op. Maar er bestaan altijd getallen $b_o, b_1, \ldots, b_n$ zodanig dat

$$\frac{p(x)}{x - \alpha} = b_o x^{n-1} + b_1 x^{n-2} + \ldots + b_{n-1} + \frac{b_n}{x - \alpha} . \qquad (2)$$

Of : $$p(x) = (x - \alpha)(b_o x^{n-1} + \ldots + b_{n-1}) + b_n . \qquad (3)$$

(Hieruit volgt al dat $b_n = p(\alpha)$ - de deling gaat dus alleen op als $p(\alpha) = 0$ - de reststelling!).
Uitwerking van het rechterlid van (3) en vergelijking met (1) leert dat

$$\left.\begin{array}{ll} b_o = a_o & \\ b_1 - \alpha b_o = a_1 & \text{of} \quad b_1 = a_1 + \alpha b_o \\ b_2 - \alpha b_1 = a_2 & \qquad b_2 = a_2 + \alpha b_1 \\ \ldots\ldots\ldots\ldots & \qquad \ldots\ldots\ldots\ldots \\ b_n - \alpha b_{n-1} = a_n & \qquad b_n = a_n + \alpha b_{n-1} \end{array}\right\}$$

Hieruit blijkt dat de $b_o, b_1, \ldots, b_n$ uit (2) precies dezelfde zijn als de $b_o, \ldots, b_n$ uit het schema van Horner!

Zij $q(x) = b_o x^{n-1} + \ldots + b_{n-1}$.
Dan is dus $p(x) = (x - \alpha)q(x) + b_n$.
Pas nu op $b_o, \ldots, b_{n-1}$ het schema van Horner toe, zodat coëfficiënten $c_o, \ldots, c_{n-1}$ ontstaan en stel $r(x) = c_o x^{n-2} + \ldots + c_{n-2}$. Dan is $q(x) = (x - \alpha)r(x) + c_{n-1}$, of $p(x) = (x - \alpha)[(x - \alpha)r(x) + c_{n-1}] + b_n$.
Zo kunnen we verder gaan. Zetten we het schema volledig voort dan krijgen we (in iets gewijzigde notatie : voor $a_o, a_1, \ldots$ is geschreven $a_o^{(o)}, a_1^{(o)}, \ldots$, voor $b_o, b_1, \ldots$ is geschreven $a_o^{(1)}, a_1^{(1)}, \ldots$, enz).

$$\begin{array}{lllll} a_o^{(o)} & & & & \\ a_1^{(o)} & a_o^{(1)} & & & \\ a_2^{(o)} & a_1^{(1)} & a_o^{(2)} & & \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & & \\ a_n^{(o)} & a_{n-1}^{(1)} & a_{n-2}^{(2)} & \ldots\ldots\ldots & a_o^{(n)} \\ & a_n^{(1)} & a_{n-1}^{(2)} & \ldots\ldots\ldots & a_1^{(n)} \qquad a_o^{(n+1)} \end{array}$$

Zij dan voor $j = 0,1,\ldots,n$

$$p_j(x) = a_o^{(j)}x^{n-j} + a_1^{(j)}x^{n-j-1} + \ldots + a_{n-j}^{(j)},$$

dan is $p_o(x) = p(x)$ en er geldt

$$\begin{aligned}
p_o(x) &= (x - \alpha)p_1(x) + a_n^{(1)} \\
p_1(x) &= (x - \alpha)p_2(x) + a_{n-1}^{(2)} \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
p_{n-1}(x) &= (x - \alpha)p_n(n) + a_1^{(n)} \\
p_n(x) &= a_o^{(n+1)}.
\end{aligned}$$

$p(\alpha + (x-\alpha)) =$ Taylor.

En hieruit volgt eenvoudig dat

$$p(x) = a_o^{(n+1)}(x - \alpha)^n + a_1^{(n)}(x - \alpha)^{n-1} + \ldots + a_{n-1}^{(2)}(x - \alpha) + a_n^{(1)}. \quad (4)$$

De coëfficiënten van de ontwikkeling van $p(x)$ naar machten van $x - \alpha$ staan dus (in omgekeerde volgorde) in de onderste rij van het schema.
Tevens zien we

$$\begin{aligned}
p(\alpha) &= a_n^{(1)} \\
p'(\alpha) &= a_{n-1}^{(2)} \\
p''(\alpha) &= 2!\, a_{n-2}^{(3)} \\
&\ldots\ldots\ldots\ldots \\
p^{(n)}(\alpha) &= n!\, a_o^{(n+1)} \qquad (= n!a_o).
\end{aligned}$$

<u>Opmerking</u> : Men kan formule (4) ook lezen als

$$p(x + \alpha) = a_o^{(n+1)}x^n + a_1^{(n)}x^{n-1} + \ldots + a_n^{(1)},$$

dus als ontwikkeling van $p(x + \alpha)$ naar machten van x.
Of nog anders : Stel dat de nulpunten van $p(x)$ zijn $\alpha_1, \alpha_2, \ldots, \alpha_n$. Dan is $a_o^{(n+1)}x^n + \ldots + a_n^{(1)}$ het polynoom waarvan de nulpunten zijn $\alpha_1 - \alpha$, $\alpha_2 - \alpha, \ldots, \alpha_n - \alpha$.

Voorbeeld. $p(x) = x^3 - 3x^2 + 2x - 1, \quad \alpha = 2.$

a	$a^{(1)}$	$a^{(2)}$	$a^{(3)}$	$a^{(4)}$
1				
-3	1			
2	-1	1		
-1	0	1	1	
	-1	2	3	1

Dus $p(2) = -1$, $p'(2) = 2$, $p''(2) = 2!.3 = 6$, $p'''(2) = 3!.1 = 3$ en
$p(x) = (x-2)^3 + 3(x-2)^2 + 2(x-2) - 1$.

1.6.1.2. Symmetrische functies

Zij weer $p(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n =$
$= a_0(x - \alpha_1)(x - \alpha_2)\ldots(x - \alpha_n).$ (1)

Door uitvermenigvuldigen van de laatste uitdrukking vinden we

$$\sum_{j=1}^{n} \alpha_j = \alpha_1 + \alpha_2 + \ldots + \alpha_n = -\frac{a_1}{a_0}$$

$$\sum_{j=1}^{n} \sum_{k=j+1}^{n} \alpha_j\alpha_k = \frac{a_2}{a_0}$$

$$\sum_{j=1}^{n} \sum_{k=j+1}^{n} \sum_{=k+1} \alpha_j\alpha_k\alpha = -\frac{a_3}{a_0}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\alpha_1\alpha_2\ldots\alpha_n = (-1)^n \frac{a_n}{a_0}.$$

De linkerleden van deze uitdrukkingen zijn zogenaamde symmetrische functies van de wortels van $p(x) = 0$, d.w.z. functies die niet van waarde veranderen indien de wortels onderling verwisseld worden. Een hier niet te bewijzen stelling zegt dat iedere rationale symmetrische functie van de wortels rationaal uitgedrukt kan worden in de coëfficiënten van $p(x)$.

We beschouwen speciaal de symmetrische functies

$$S_p = \alpha_1^p + \alpha_2^p + \ldots + \alpha_n^p, \quad p \text{ geheel.}$$

Hoe kunnen we deze berekenen ?

Voor $j = 1,2,\ldots,n$ geldt

$$a_o \alpha_j^n + a_1 \alpha_j^{n-1} + \ldots + a_n = 0 .$$

Vermenigvuldig met α_j^k en sommeer over j. Dan blijkt dat voor iedere $k \geqslant 0$

$$a_o S_{n+k} + a_1 S_{n+k-1} + \ldots + a_n S_k = 0. \qquad (2)$$

Hieruit kunnen we successief S_n, S_{n+1}, S_{n+2},... berekenen indien S_o, S_1,...,S_{n-1} bekend zijn.
Deze laatste vinden we als volgt. Uit (1) vplgt eenvoudig

$$\frac{p'(x)}{p(x)} = \sum_{j=1}^{n} \frac{1}{x - \alpha_j} .$$

Stel $|x|$ > de grootste van $|\alpha_1|$, $|\alpha_2|$,..., $|\alpha_n|$.

Dan is $\frac{1}{x - \alpha_j} = \frac{1}{x} \cdot \frac{1}{1 - \frac{\alpha_j}{x}} = \sum_{k=o}^{\infty} \frac{\alpha_j^k}{x^{k+1}} .$

En dus $p'(x) = p(x) . \sum_{k=o}^{\infty} \frac{S_k}{x^{k+1}} .$

of $na_o x^{n-1} + (n-1)\, a_o x^{n-2} + \ldots + a_{n-1} = (a_o x^n + \ldots + a_n)(\frac{S_o}{x} + \frac{S_1}{x^2} + \ldots).$

Hieruit volgt :

$$\begin{aligned}
na_o &= a_o S_o \\
(n-1)a_1 &= a_o S_1 + a_1 S_o \\
(n-2)a_2 &= a_o S_2 + a_1 S_1 + a_2 S_o \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
a_{n-1} &= a_o S_{n-1} + a_1 S_{n-2} + \ldots + a_{n-1} S_o \\
0 &= a_o S_{n+k} + a_1 S_{n+k-1} + \ldots + a_n S_k . \qquad (k \geqslant 0)
\end{aligned}$$

Uit de eerste relatie volgt $S_o = n$ (allicht!).
Gebruiken we dit verder, dan vinden we

$$\left.\begin{aligned}
a_o S_1 + a_1 &= 0 \\
a_o S_2 + a_1 S_1 + 2a_2 &= 0 \\
a_o S_3 + a_1 S_2 + a_2 S_1 + 3a_3 &= 0 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
a_o S_{n-1} + a_1 S_{n-2} + \ldots + a_{n-2} S_1 + (n-1)a_{n-1} &= 0
\end{aligned}\right\} \qquad (3)$$

terwijl voor $k \geqslant 0$ geldt

$$a_o S_{n+k} + a_1 S_{n+k-1} + \ldots + a_n S_k = 0.$$

(dit wisten we al).

Uit de relaties (3) kunnen we dus successief S_1, S_2,..., S_{n-1} berekenen waarna S_{n+k} ($k \geq 0$) uit (1) volgt.

Opmerkingen

1. Is $a_n \neq 0$ (zodat geen der wortels α_j gelijk is aan nul) dan geldt (2) voor iedere waarde van k (is een der wortels gelijk aan nul, dan heeft S_k voor $k < 0$ geen zin).
Uit (3) en (2) volgt dan dat (daar $S_0 = n$)

$$a_n S_{-1} + a_{n-1} = 0$$
$$a_n S_{-2} + a_{n-1} S_{-1} + 2a_{n-2} = 0$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$a_n S_{-n+1} + a_{n-1} S_{-n+2} + \ldots + (n-1)\, a_1 = 0.$$

Hieruit kunnen dus S_{-1}, S_{-2},..., S_{-n+1} berekend worden. S_{-n-k} ($k \geq 0$) volgt dan uit de relatie (2) geschreven in de vorm

$$a_n S_{-n-k} + a_{n-1} S_{-n-k+1} + \ldots + a_0 S_{-k} = 0.$$

2. Zij u_0, u_1,..., u_{n-1} gegeven getallen en zij voor $k \geq 0$ u_{n+k} bepaald door

$$a_0 u_{n+k} + a_1 u_{n+k-1} + \ldots + a_n u_k = 0. \qquad (4)$$

Men noemt de rij $u_0, u_1, \ldots$ dan oplossing van de lineaire recursievergelijking (4) met beginwaarden u_0, u_1,..., u_{n-1}.
Zijn α_1,..., α_n alle onderling verschillend dan is de algemene oplossing van (4)

$$u_k = C_1 \alpha_1^k + C_2 \alpha_2^k + \ldots + C_n \alpha_n^k . \qquad (5)$$

De coëfficiënten C_1,...,C_n vindt men door in (5) $k=0,1,\ldots, n-1$ te nemen.
Uit het bovenstaande blijkt dat $u_0 = S_0$, $u_1 = S_n$, $u_{n-1} = S_{n-1}$ correspondeert met $C_1 = C_2 = \ldots = C_n = 1$ (en dus $u_k = S_k$ voor alle k).

1.6.1.3. Schattingen voor de grootste wortel

Het kan van belang zijn een schatting te hebben voor de wortel met de grootste absolute waarde, d.w.z. een getal M zodanig dat $|\alpha_j| < M$ voor $j = 1,\ldots,n$.
Weet men op een of andere grond dat alle wortels reeel zijn, dan is een eenvoudige schatting

$$|\alpha_j| \leqslant \sqrt{\alpha_1^2 + \ldots + \alpha_n^2} = \sqrt{S_2} = \sqrt{\frac{a_1^2 - 2a_0a_2}{a_0^2}}.$$

In het algemeen kan men als volgt schattingen verkrijgen.
Veronderstel voor het gemak dat $a_0 = 1$. Zij z een wortel van $p(x) = 0$.
Dan is $z^n = -a_1 z^{n-1} - \ldots - a_n$.

Dus $$|z|^n \leqslant |a_1||z|^{n-1} + \ldots + |a_n|. \qquad (1)$$

Veronderstel dat $|z| \leqslant 1$. Dan volgt uit (1)

$$|z|^n \leqslant |a_1| + \ldots |a_n|.$$

Veronderstel dat $|z| \geqslant 1$. Dan volgt uit (1)

$$\begin{aligned} |z| &\leqslant |a_1| + |a_2||z|^{-1} + \ldots + |a_n||z|^{1-n} \\ &\leqslant |a_1| + \ldots + |a_n| \end{aligned}$$

Combinatie van deze resultaten levert (ga na)

Als $|a_1| + \ldots + |a_n| \leqslant 1$ dan geldt voor iedere wortel z

$$|z| \leqslant \sqrt[n]{|a_1| + \ldots + |a_n|}$$

Als $|a_1| + \ldots + |a_n| \geqslant 1$ dan geldt voor iedere wortel z

$$|z| \leqslant |a_1| + \ldots + |a_n|.$$

Een ander resultaat verkrijgt men als volgt. Stel

$$A = \text{Max}\{|a_1| + 1, |a_2| + 1, \ldots, |a_{n-1}| + 1, |a_n|\}.$$

Dan volgt uit (1)

$$|z|^n \leqslant (A - 1)[|z|^{n-1} + \ldots + |z|] + A,$$

of

$$|z|^n + |z|^{n-1} + \ldots + |z| \leqslant A[|z|^{n-1} + \ldots + 1].$$

Derhalve geldt voor iedere wortel z

$$|z| \leqslant A.$$

Deze schattingen zijn in het algemeen vrij grof (hoewel er vergelijkingen zijn waarvoor het gelijkteken geldt).
Iets scherpere schattingen krijgt men door in plaats van (1) te schrijven

$$|z|^{n-1}|z + a_1| \leqslant |a_2||z|^{n-2} + \ldots + |a_n|.$$

Men krijgt dan als resultaten de stellingen van Gershgorin :

Iedere wortel z ligt op of binnen tenminste een van de cirkels $|z| = 1$ en $|z + a_1| = |a_2| + \ldots + |a_n|$.
En ook op of binnen tenminste een van de cirkels $|z + a_1| = 1$ en $|z| = \text{Max}\{|a_2| + 1, \ldots, |a_{n-1}| + 1, |a_n|\}$.

1.6.2. Bepaling van de wortels

Er bestaan diverse methoden voor de bepaling van de wortels van een n^e graadsvergelijking. Geen der methoden is voor alle gevallen geheel bevredigend. We behandelen een drietal methoden die geschikt in combinatie gebruikt kunnen worden.

1.6.2.1. Methode van Bernoulli

In 1.6.1.2. hebben we gezien hoe op betrekkelijk eenvoudige manier de getallen $S_0, S_1, S_2, \ldots$, gedefinieerd

$$S_k = \sum_{j=1}^{n} \alpha_j^k$$

bepaald kunnen worden. Uit het gedrag van deze getallen voor grote k kunnen we benaderingen voor de grootste wortel(s) afleiden. We moeten (helaas) diverse gevallen onderscheiden .

a. $|\alpha_1| > |\alpha_2| \geqslant \ldots \geqslant |\alpha_n|$. Er is dus één reële wortel waarmee de absolute waarde groter is dan die van alle andere wortels.
Daar

$$S_k = \alpha_1^k \left[1 + \left(\frac{\alpha_2}{\alpha_1}\right)^k + \ldots + \left(\frac{\alpha_n}{\alpha_1}\right)^k\right]$$

geldt in dit geval

$$\alpha_1 = \lim_{k \to \infty} \frac{S_{k+1}}{S_k} . \qquad (1)$$

De convergentie is lineair, als $|\alpha_2| > |\alpha_1|$ dan geldt

$$\frac{S_{k+1}}{S_k} = \alpha_1 - (\alpha_1 - \alpha_2)\left(\frac{\alpha_2}{\alpha_1}\right)^k + \ldots$$

De convergentie is dus goed als $\left|\frac{\alpha_2}{\alpha_1}\right|$ flink wat kleiner is dan 1, slecht als dit getal dicht bij 1 ligt.

b. $|\alpha_1| = |\alpha_2| < |\alpha_3| \leqslant \ldots \leqslant |\alpha_n|$.
In dit geval moeten we drie deelgevallen onderscheiden

b_1. $\alpha_2 = \alpha_1$ (reeel)
b_2. $\alpha_2 = -\alpha_1$ (reeel)
b_3. $\alpha_2 = \bar{\alpha}_1$ (complex).

In geval b_1 is er in principe geen moeilijkheid, daar nu

$$S_k = \alpha_1^k \left[2 + \left(\frac{\alpha_3}{\alpha_1}\right)^k + \ldots + \left(\frac{\alpha_n}{\alpha_1}\right)^k\right],$$

zodat (1) blijft gelden. De invloed van afrondingsfouten op de convergentiesnelheid blijkt echter groter te zijn dan in geval a.

In geval b_2 geldt

$$S_{2k} = \alpha_1^{2k}\left[2 + \left(\frac{\alpha_3}{\alpha_1}\right)^{2k} + \ldots\right]$$

$$S_{2k+1} = \alpha_3^{2k+1}\left[1 + \left(\frac{\alpha_4}{\alpha_3}\right)^{2k+1} + \ldots\right].$$

De rij $\left\{\frac{S_{2k}}{S_{2k-1}}\right\}$ gaat dus naar ∞, de rij $\left\{\frac{S_{2k+1}}{S_{2k}}\right\}$ gaat naar nul. Heeft men deze situatie echter herkend, dan volgen de wortels α_1 en $\alpha_2 = -\alpha_1$ eenvoudig uit

$$\alpha_1^2 = \lim_{k\to\infty} \frac{S_{2k+2}}{S_{2k}}. \qquad (2)$$

In geval b_3 (dat in de praktijk vaker voor zal komen dan b_1 of b_2) stellen we

$$\alpha_1 = r\,e^{i\theta},\quad \alpha_2 = r\,e^{-i\theta},\quad \text{met } r > 0,\quad 0 < \theta < \pi$$

Dan geldt $r > |\alpha_3| \geqslant \ldots \geqslant |\alpha_n|$. En we kunnen schrijven

$$S_k = r^k\left[2\cos k\theta + \left(\frac{\alpha_3}{r}\right)^k + \ldots + \left(\frac{\alpha_n}{r}\right)^k\right]. \qquad (3)$$

De rij $\left\{\frac{S_{k+1}}{S_k}\right\}$ gedraagt zich dus grillig.

Na enig rekenen volgt echter uit (3) dat *)

---

*) Met $O\left(\left(\frac{\alpha_3}{r}\right)^{2k}\right)$ wordt bedoeld : er is een getal M zodanig dat voor alle k $|S_k^2 - S_{k-1}\,S_{k+1} - 4r^2\sin^2\theta| < M\left|\frac{\alpha_3}{r}\right|^{2k}$.

$$S_k^2 - S_{k-1}S_{k+1} = 4r^2 \left[\sin^2\theta + 0\left(\left(\frac{\alpha_3}{r}\right)^{2k}\right)\right]$$

$$S_kS_{k+1} - S_{k-1}S_{k+2} = 8r^{2k+1}\left[\cos\theta\,\sin^2\theta + 0\left(\left(\frac{\alpha_3}{r}\right)^{2k+1}\right)\right].$$

Hieruit volgt dat

$$\left.\begin{aligned} r^2 &= \lim_{k\to\infty} \frac{S_{k+1}^2 - S_kS_{k+2}}{S_k^2 - S_{k-1}S_{k+1}} \\ 2r\cos\theta &= \lim_{k\to\infty} \frac{S_kS_{k+1} - S_{k-1}S_{k+2}}{S_k^2 - S_{k-1}S_{k+1}} \end{aligned}\right\} \qquad (4)$$

Heeft men op deze wijze r^2 en $2r\cos\theta$ benaderd dan volgen α_1 en $\alpha_2 = \bar{\alpha}_1$ als oplossingen van de vierkantsvergelijking

$$x^2 - 2r\,x\cos\theta + r^2 = 0,$$

dus $$\alpha_{1,2} = r\cos\theta \pm i\sqrt{r^2 - r^2\cos^2\theta}.$$

c. $|\alpha_1| = |\alpha_2| = |\alpha_3| \geqslant \ldots$

Dit geval dat in de praktijk niet zo vaak voor zal komen (behalve in licht herkenbare gevallen als $x^n - a = 0$) behandelen we niet.

In de praktijk gelukt het meestal wel om na het uitrekenen van een betrekkelijk gering aantal termen van de rij $S_0, S_1, S_2, \ldots$ uit te maken in welk geval men zich bevindt en om de limieten (1), (2) of (3) met enige nauwkeurigheid te bepalen. Voor een nauwkeurige bepaling van de wortel moet men meestal echter vrij ver doorgaan. Derhalve gebruikt men de methode van Bernoulli in hoofdzaak om een ruwe benadering van de grootste wortel(s) te vinden. Een nauwkeuriger waarde kan dan zo nodig bepaald worden met behulp van een der hieronder beschreven kwadratische processen.

<u>Voorbeeld</u>. $x^3 - 15x^2 + 84x + 100 = 0$.

Men vindt

$S_0 = 3$ $\qquad$ $S_2 = 15 \times 15 - 84 \times 2 = 57$

$S_1 = 15 \times 1 = 15$ $\qquad$ $S_3 = 15 \times 57 - 84 \times 15 - 100 \times 3 = -705.$

$S_4 = -16863$

$S_5 = -199425$

$S_6 = -1504383$

$S_7 = -4127745$

$S_8 = +84394497$

$S_9 = +1763086335$

Hieruit volgt

n	$\frac{S_{n+1}^2 - S_n S_{n+2}}{S_n^2 - S_{n-1} S_{n+1}}$	$\frac{S_n S_{n+1} - S_{n-1} S_{n+2}}{S_n^2 - S_{n-1} S_{n+1}}$
3	98.6	15.95
4	100.18	16.014
5	99.986	15.9982
6	100.00046	16.00014
7	100.000065	15.999995

Daar $p(x) = (x^2 - 16x + 100)(x + 1)$ geldt hier $|\alpha_1| = |\alpha_2| = 10$; $\alpha_3 = -1$.
De convergentie factor is dus ca 100!

1.6.2.2. Methode van Newton-Raphson

Heeft men (met de methode van Bernoulli of op andere wijze) een schatting α gevonden voor een wortel α dan kan men deze schatting verbeteren met de methode van Newton-Raphson (bij toepassing op n^e graads vergelijkingen ook wel methode van Birge-Vieta genaamd) :

$$\alpha_1 = \alpha - \frac{p(\alpha)}{p'(\alpha)} = \alpha_o - \frac{b_n(\alpha)}{c_{n-1}(\alpha)} ,$$

waarin met $b_n(\alpha)$ resp. $c_{n-1}(\alpha)$ de laatste coëfficiënten van de eerste, resp. tweede kolom van het schema van Horner (zie 1.6.1.1), berekend voor de waarde α, bedoeld is. Met de gevonden α_1 bepaalt men vervolgens een α_2, enz.
Bij een enkelvoudige wortel geeft dit, indien men uitgaat van een redelijke beginschatting een snelle benadering van α. Bij een meervoudige wortel is de convergentie slechts lineair, doch men kan deze desgewenst versnellen met behlup van het δ^2-proces van Aitken.
Het proces van Newton werkt ook voor complexe wortels (het meetkundige beeld van de raaklijnen wordt dan zinloos, het analytische bewijs blijft echter precies hetzelfde). Bij vergelijkingen met reële coëfficiënten (zoals wij steeds hier beschouwen) kan men voor het geval van toegevoegd complexe wortels beter het hieronder volgende proces van Bairstow gebruiken.

1.6.2.3. Methode van Bairstow

Zij $p(x) = a_o x^n + \ldots + a_n$, $n \geqslant 2$ en zij β en γ gegeven getallen. Dan kan men getallen $b_o,\ldots,b_n$ bepalen zodanig dat

$$\frac{p(x)}{x^2 - \beta x - \gamma} = b_o x^{n-2} + \ldots + b_{n-2} + \frac{b_{n-1}(x - \beta) + b_n}{x^2 - \beta x - \gamma} \quad ;$$

of $$p(x) = (x^2 - \beta x - \gamma)(b_o x^{n-2} + \ldots + b_{n-2}) + b_{n-1}(x - \beta) + b_n.$$

Uitvermenigvuldigen levert (ga na)

$$\begin{aligned} a_o &= b_o \\ a_1 &= b_1 - \beta b_o \\ a_2 &= b_2 - \beta b_1 - \gamma b_o \\ &\ldots\ldots\ldots\ldots\ldots \\ a_{n-2} &= b_{n-2} - \beta b_{n-3} - \gamma b_{n-4} \\ a_{n-1} &= b_{n-1} - \beta b_{n-2} - \gamma b_{n-3} \\ a_n &= b_n - \beta b_{n-1} - \gamma b_{n-2} \end{aligned}$$

of

$$\begin{aligned} b_o &= a_o \\ b_1 &= a_1 + \beta b_o \\ b_2 &= a_2 + \beta b_1 + \gamma b_o \\ &\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

Dus algemeen $b_j = a_j + \beta b_{j-1} + \gamma b_{j-2}$, $j = 0,1,\ldots,n$, mits men $b_{-2} = b_{-1} = 0$ stelt. Men kan dit weer rangschikken in een schema dat op dat van Horner lijkt :

	β	γ
a_o	0	0
a_1	b_o	0
a_2	b_1	b_o
a_3	b_2	b_1
.....	.....	.....
a_n	b_{n-1}	b_{n-2}
	b_n	

Indien blijkt dat $b_{n-1} = b_n = 0$ dan is $x^2 - \beta x - \gamma$ een factor van $p(x)$ en de wortels $x = \frac{1}{2}[\beta \pm \sqrt{\beta^2 + 4\gamma}\,]$ van $x^2 - \beta x - \gamma = 0$ zijn ook wortels van $p(x) = 0$. Bovendien is in dat geval $b_o x^{n-2} + \ldots b_{n-2}$ het quotiënt van $p(x)$ en $x^2 - \beta x - \gamma$.

Men kan de door het bovenstaande schema bepaalde getallen b_{n-1} en b_n beschouwen als functies van β en γ :

$$b_{n-1} = b_{n-1}(\beta,\gamma), \quad b_n = b_n(\beta,\gamma).$$

Het zoeken van de kwadratische factor $x^2 - \beta x - \gamma$ van $p(x)$ is dus equivalent met het zoeken van een oplossing van de vergelijkingen
$b_{n-1}(\beta,\gamma) = 0$, $b_n(\beta,\gamma) = 0$.
Het iteratieproces van Bairstow zoekt kwadratische factoren van $p(x)$ door het stelsel vergelijkingen

$$F(\beta,\gamma) = 0, \quad G(\beta,\gamma) = 0, \tag{1}$$

waarin $F(\beta,\gamma) = b_{n-1}(\beta,\gamma)$, $G(\beta,\gamma) = b_n(\beta,\gamma) - \beta b_{n-1}(\beta,\gamma)$ met behulp van Newton-Raphson iteratief op te lossen.*) p. 18 verklaring
Volgens 1.5.3. vindt men uit een benadering (β,γ) voor een oplossing van (1) een betere benadering (β_1,γ_1) gegeven door

$$\beta_1 = \beta - \frac{B}{\Delta}, \qquad \gamma_1 = \gamma - \frac{C}{\Delta},$$

waarin

$$B = FG_\gamma - F_\gamma G$$

$$C = F_\beta G - FG_\beta$$

$$\Delta = F_\beta G_\gamma - F_\gamma G_\beta,$$

(alle functies genomen in het punt (β,γ)).

Om B, C en Δ te kunnen berekenen moeten we de afgeleiden van b_{n-1} en b_n naar β en γ kennen.
Nu was b_j bepaald door

$$b_{-2} = 0, \quad b_{-1} = 0, \quad b_j = a_j + \beta b_{j-1} + \gamma b_{j-2} \qquad (j \geqslant 0).$$

Hieruit volgt

$$\frac{\partial b_{-1}}{\partial \beta} = 0, \quad \frac{\partial b_0}{\partial \beta} = 0 \quad \text{(daar } b_0 = a_0\text{)} \quad \text{en}$$

*) Men kan natuurlijk in plaats van (1) ook het stelsel

$$b_{n-1}(\beta,\gamma) = 0, \quad b_n(\beta,\gamma) = 0 \tag{2}$$

oplossen. In de praktijk blijkt dat bij een niet erg nauwkeurige beginschatting het Newtonproces voor het stelsel (1) vaak beter convergeert dan voor het stelsel (2).

$$b_j = a_j + \beta b_{j-1} + \gamma b_{j-2}$$

$$\frac{\partial b_j}{\partial \beta} = b_{j-1} + \beta \frac{\partial b_{j-1}}{\partial \beta} + \gamma \frac{\partial b_{j-2}}{\partial \gamma} \qquad (j \geqslant 1).$$

Of, als we stellen $\frac{\partial b_j}{\partial \beta} = c_{j-1}$ $(j \geqslant -1)$,

$$c_{-2} = c_{-1} = 0, \quad c_j = b_j + \beta c_{j-1} + \gamma c_{j-2} \qquad (j \geqslant 0).$$

Analoog geldt

$$\frac{\partial b_0}{\partial \gamma} = 0, \quad \frac{\partial b_1}{\partial \gamma} = 0 \qquad (\text{daar } b_1 = a_1 + \beta a_0)$$

$$\frac{\partial b_j}{\partial \gamma} = b_{j-2} + \beta \frac{\partial b_{j-1}}{\partial \gamma} + \gamma \frac{\partial b_{j-2}}{\partial \gamma} \qquad (j \geqslant 2).$$

Of, als we stellen $\frac{\partial b_j}{\partial \gamma} = d_{j-2}$ $(j \geqslant 0)$,

$$d_{-2} = d_{-1} = 0, \quad d_j = b_j + \beta d_{j-1} + \gamma d_{j-2} \qquad (j \geqslant 0).$$

Maar hieruit volgt dat $d_j = c_j$ voor $j \geqslant 0$ (zelfde recurrente betrekking en zelfde beginwaarden).

We vinden dus :

Als de getallen c_j $(j \geqslant -2)$ bepaald zijn door

$$c_{-2} = c_{-1} = 0, \quad c_j = b_j + \beta c_{j-1} + \gamma c_{j-2} \qquad (j \geqslant 0)$$

dan is

$$\frac{\partial b_{n-1}}{\partial \beta} = c_{n-2} \qquad \frac{\partial b_n}{\partial \beta} = c_{n-1}$$

$$\frac{\partial b_{n-1}}{\partial \gamma} = c_{n-3} \qquad \frac{\partial b_n}{\partial \gamma} = c_{n-2}.$$

Merk op dat de getallen c_j op precies dezelfde manier uit de getallen b_j volgen als de b_j uit de a_j!

Met deze resultaten vindt men

$$B = b_{n-1}(c_{n-2} - \beta c_{n-3}) - c_{n-3}(b_n - \beta b_{n-1}) = b_{n-1}c_{n-2} - b_n c_{n-3}$$

$$C = c_{n-2}(b_n - \beta b_{n-1}) - b_{n-1}(c_{n-1} - b_{n-1} - \beta c_{n-2}) =$$

$$= b_n c_{n-2} - b_{n-1}c_{n-1} + b_{n-1}^2$$

$$\Delta = c_{n-2}(c_{n-2} - \beta c_{n-3}) - c_{n-3}(c_{n-1} - b_{n-1} - \beta c_{n-2}) =$$

$$c_{n-2}^2 - c_{n-3}c_{n-1} + c_{n-3}b_{n-1}.$$

Opmerking. De formules voor C en Δ worden wat fraaier indien men schrijft $\bar{c}_{n-1} = c_{n-1} - b_{n-1} = \beta c_{n-2} + \gamma c_{n-3}$. De formules voor $\bar{c}_{n-1}$ wijkt dan af van die voor c_j $(c_j < n-1)$.

Voorbeeld. $x^4 + 5x^3 + 10x^2 + 5x + 10 = 0$.

Pas de methode van Bernoulli toe.

n	S_n	
0	4	$\frac{S_{10}^2 - S_9 S_{11}}{S_9^2 - S_8 S_{10}} = 9.197$
1	-5	
2	5	
3	10	$\frac{S_{11}^2 - S_{10} S_{12}}{S_{10}^2 - S_9 S_{11}} = 9.189$
4	-115	
5	500	
6	-1450	
7	2725	$\frac{S_9 S_{10} - S_8 S_{11}}{S_9^2 - S_8 S_{10}} = -5.056$
8	-495	
9	-22525	
10	118450	$\frac{S_{10} S_{11} - S_9 S_{12}}{S_{10}^2 - S_9 S_{11}} = -5.055$
11	-391775	
12	891950	

Als eerste benadering voor een kwadratische factor hebben we dus $x^2 + 5.055x + 9.189$.

Het Bairstow schema levert

	-5.055	-9.189		-5.055	-9.189
1	0	0	1	0	0
5	1	0	-0.055	1	0
10	-0.055	1	1.089025	-5.110	1
5	1.089025	-0.055	0.000374	17.731	-5.110
10	0.000374	1.089025		-42.674	
	-0.008941				

Hieruit volgt $B = -0.039057$

$C = -0.142572$

$\Delta = 96.324$

en als nieuwe benadering voor de kwadratische factor

$$x^2 + 5.054595x + 9.187520.$$

Deling door deze factor levert

	-5.054595	-9.187520
1	0	0
5	1	0
10	-0.054595	1
5	1.088436	-0.054595
10	-0.000009	1.088436
	+0.000019	

De andere kwadratische factor is dus ongeveer

$$x^2 - 0.054595x + 1.088436.$$

Als controle delen we p(x) door deze factor

	0.054595	-1.088436
1	0	0
5	1	0
10	5.054595	1
5	9.187520	5.054595
10	-0.000010	9.187520
	-0.000029	

We vinden dus (in 6 decimalen) precies de eerste kwadratische factor terug.

Als benaderingen voor de wortels vinden we

$$-2.527298 \pm 1.673406\ i$$
$$0.027298 \pm 1.042924\ i$$

1.6.2.4. Bepaling van de overige wortels

De methode van Bernoulli levert slechts een (of twee) wortel(s), nl. die met de grootste absolute waarde. Er bestaan diverse methoden om de overige wortels te vinden.

a. Is $a_o \neq 0$ (zodat $x = 0$ geen wortel is) dan kan men de wortel(s) met de kleinste absolute waarde vinden door Bernoulli toe te passen op de vergelijking

$$a_n y^n + a_{n-1} y^{n-1} + \ldots a_o = 0$$

(hoe?).

b. Heeft men één (of twee) wortel(s) bepaald dan kan men $p(x)$ delen door de corresponderende lineaire (kwadratische) factor. Dit gebeurt met het schema van Horner (resp. het in 1.6.2.3 gegeven schema). Is het quotiënt $q(x)$ (een polynoom van de graad n-1, resp. n-2) dan bepaalt men vervolgens met Bernoulli (eventueel gevolgd door Newton of Bairstow) de wortel(s) met de grootste absolute waarde van de vergelijking $q(x) = 0$. Deze voldoet ook aan $p(x) = 0$. Etc.

Om de invloed van afbreekfouten (men bepaalt de wortels niet exact en het verkregen quotiënt is dus ook niet exact) te beperken kan men nadat een (of twee) wortel(s) van $q(x) = 0$ met redelijke nauwkeurigheid bepaald is, de nauwkeurigheid van deze wortel(s) controleren, cq. verbeteren door met deze wortel(s) als startwaarde Newton (Bairstow) toe te passen op de oorspronkelijke vergelijking $p(x) = 0$.

c. Men kan ook nadat als boven het quotiënt $q(x)$ bepaald is, meteen Newton of Bairstow toepassen op de vergelijking $q(x) = 0$ met als startwaarde de reeds gevonden wortel(s). In het algemeen convergeert dit proces dan naar de volgende wortel van de oorspronkelijke vergelijking.

d. Bij werk met automatische rekenmachines is het van belang een proces te hebben waarbij niet een herkenning en aparte behandeling van allerlei speciale gevallen nodig is. Het Bernoulli proces voldoet niet aan deze eis.

Men kan echter ook vanaf het eerste begin het Bairstow-proces toepassen, bv. met startwaarden $\beta = \gamma = 0$ (laat zien dat dan

$$\beta_1 = -\frac{a_{n-1}\,a_{n-2} - a_n\,a_{n-3}}{a_{n-2}^2}\,,\quad \gamma_1 = -\frac{a_n}{a_{n-2}}\quad)$$

In het algemeen convergeert de iteratie dan naar de kwadratische factor corresponderend met de kleinste twee wortels. Vaak zijn vele iteratiestappen nodig voordat men in de buurt van deze factor is, maar dat is bij een snelle machine niet zo erg. In enkele gevallen treedt helemaal geen convergentie op of wordt Δ nul(bv. als $a_{n-2}=0$). Men moet dan opnieuw beginnen met andere startwaarden.
Na het uitdelen van de gevonden kwadratische factor begint men weer opnieuw, nu met de coëfficiënten van de zojuist uitgedeelde factor als startwaarden. Etc.

e. Er bestaan ook methoden die voor alle wortels tegelijk benaderingen geven (bv. het proces van Dandelin-Graeffe). Deze hebben echter nogal wat bezwaren.

2. Lineaire vergelijkingen

2.1. Inleiding

In dit hoofdstuk behandelen we het oplossen van stelsels lineaire vergelijkingen van de vorm

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n. \end{aligned} \qquad (1)$$

Voor de zuiver wiskundige is hier geen probleem aangezien volgens de regel van Cramer (1750!) de oplossing gegeven wordt door

$$x_j = \frac{\begin{vmatrix} a_{11} \ldots a_{1,j-1} & b_1 & a_{1,j+1} \ldots a_{1n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\ a_{n1} \ldots a_{n,j-1} & b_n & a_{n,j+1} \ldots a_{nn} \end{vmatrix}}{\begin{vmatrix} a_{11} \ldots\ldots a_{1n} \\ \ldots\ldots\ldots \\ a_{n1} \ldots\ldots a_{nn} \end{vmatrix}}, \quad j=1,\ldots,n$$

en de berekening van determinanten volkomen bepaald is door de regels dat

$$\begin{vmatrix} c_{11} \ldots c_{1n} \\ \ldots\ldots\ldots \\ c_{n1} \ldots c_{nn} \end{vmatrix} = \sum_{j=1}^{n} (-1)^{j-1} c_{1j} \cdot \begin{vmatrix} c_{21} \ldots c_{2,j-1} & c_{2,j+1} \ldots c_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ c_{n1} \ldots c_{n,j-1} & c_{n,j+1} \ldots c_{nn} \end{vmatrix}$$

(ontwikkeling naar de eerste rij waardoor de berekening van $n \times n$-determinanten teruggebracht is tot de berekening van $(n-1) \times (n-1)$-determinanten) en $|c_{11}| = c_{11}$ (berekening van een 1×1-determinant).

De numericus is echter met deze oplossing, waarin is aangegeven hoe de oplossing door eindig veel bewerkingen op de coëfficiënten van (1) gevonden kan worden, niet volledig gelukkig. Want toepassing van deze regels leidt, als n enigzins groot is, tot een astronomisch groot aantal bewerkingen.

Zij nl. het aantal vermenigvuldigingen *) nodig om met behulp van deze regels een n × n-determinant uit te rekenen f(n). Dan is kennelijk

$$f(n) = n + nf(n-1), \quad (n \geqslant 2) \quad \text{en} \quad f(1) = 0.$$

Om uit deze recursiebetrekking f(n) te bepalen stellen we $f(n) = n!.g(n)$. Dan moet

$$g(n) - g(n-1) = \frac{1}{(n-1)!} \quad (n \geqslant 2) \quad \text{en} \quad g(1) = 0,$$

waaruit volgt dat voor $n \geqslant 2$

$$1 \leqslant g(n) = \frac{1}{1!} + \frac{1}{2!} + \ldots + \frac{1}{(n-1)!} < e - 1$$

en dus $\quad n! \leqslant f(n) < (e-1).n!$

Voor het uitrekenen van een 20 × 20-determinant zouden dus ca $20! \sim 2.4 \times 10^{18}$ vermenigvuldigingen nodig zijn. Met een uiterst snelle automatische machine met een vermenigvuldigtijd van 20.10^{-6} sec. zou men dus ca 0.5×10^{13} sec $\sim$ 2.10^5 jaar nodig hebben!

Later zullen we een methode aangeven waarbij voor de berekening van een n × n-determinant slechts ca $\frac{1}{3}\,n^3$ vermenigvuldigingen nodig zijn. Ook dan blijft de regel van Cramer niet aanbevelenswaard aangezien een n × n-stelsel vergelijkingen ook met ca $\frac{1}{3}\,n^3$ vermenigvuldigingen opgelost blijkt te kunnen worden.

---

*) De tijd nodig voor een vermenigvuldiging is - zowel bij het rekenen uit het hoofd, met een tafelmachine of met een automatische rekenmachine - essentieel langer dan die nodig voor een optelling of aftrekking. Daarom telt men meestal alleen het aantal nodige vermenigvuldigingen (en delingen, die meestal met vermenigvuldigingen over één kam geschoren worden).

Een ander probleem dat de numericus bezig houdt is dat van de invloed van afrondingsfouten. Deze invloed kan onder omstandigheden desastreus zijn. Een volledige theorie hierover bestaat niet. Wel kan men een aantal practische regels aangeven.
Speciale moeilijkheden treden op indien het stelsel vergelijkingen zg. "ill-conditioned" is. In dit geval treden bij kleine wijzigingen van de coëfficiënten of van de rechterleden grote variaties in de oplossing op.

Voorbeeld

$$\begin{aligned} 5x_1 + 7x_2 + 6x_3 + 5x_4 &= 23 \\ 7x_1 + 10x_2 + 8x_3 + 7x_4 &= 32 \\ 6x_1 + 8x_2 + 10x_3 + 9x_4 &= 33 \\ 5x_1 + 7x_2 + 9x_3 + 10x_4 &= 31. \end{aligned}$$

De oplossing is $x_1 = x_2 = x_3 = x_4 = 1$.
Verandert men de rechterleden in 23.01, 31.99, 32.99 en 31.01 dan wordt de oplossing (exact)

$$x_1 = 2.36, \quad x_2 = 0.18, \quad x_3 = 0.65, \quad x_4 = 1.21!$$

De vlakken in R_4 die door deze vergelijkingen voorgesteld worden zijn "bijna" evenwijdig. Een kleine verschuiving van de vlakken geeft daarom een zeer grote verplaatsing van het snijpunt. Het is duidelijk dat de invloed van afrondingsfouten hier bijzonder groot kan zijn. Bovendien dient de numericus zijn opdrachtgever te waarschuwen indien hij bv. weet dat de rechterleden door metingen verkregen zijn : deze meetmethode is, tenzij de metingen extreem nauwkeurig worden uitgevoerd, blijkbaar niet geschikt om de onbekenden $x_1, \ldots, x_4$ met enige nauwkeurigheid te bepalen!

Behalve naar de oplossing van het stelsel (1) dat verkort geschreven kan worden als

$$A\underline{x} = \underline{b} \qquad (1a)$$

kan men ook vragen naar de inverse van de matrix A, dat is een matrix A^{-1} zodanig dat

$$AA^{-1} = A^{-1}A = I \qquad (2)$$

(waarin I de n × n-eenheidsmatrix is). In componenten geschreven luidt dit

$$\sum_{j=1}^{n} A_{ij} \, (A^{-1})_{jk} = \sum_{j=1}^{n} (A^{-1})_{ij} \, A_{jk} = \delta_{ik}, \qquad (2a)$$

waarin $\delta_{ik} = \begin{cases} 1 & \text{als } i = k \\ 0 & \text{als } i \neq k \end{cases}$ (Kronecker-symbool)

Kent men de matrix A^{-1} dan is de oplossing van het stelsel (1a) ook direct uit te rekenen :

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

Omgekeerd kan men A^{-1} berekenen door n stelsels van het type (1a) op te lossen. Zijn nl. $\underline{x}_1, \ldots, \underline{x}_n$ de oplossingen van de stelsels

$$A\underline{x}_k = \underline{e}_k, \qquad k = 1, \ldots, n$$

waarin $\underline{e}_k$ de k-de eenheidsvector is ($(\underline{e}_k)_i = \delta_{ik}$), dan zijn $\underline{x}_1, \ldots, \underline{x}_n$ de kolommen van de matrix A^{-1}.

In het voorgaande is steeds verondersteld dat de determinant van de matrix A niet nul is. Is det A wel nul dan is het stelsel (1a) in het algemeen strijdig, d.w.z. heeft het geen oplossing. Voldoet de vector $\underline{b}$ echter aan bepaalde voorwaarden (zodat het stelsel (1a) afhankelijk wordt) dan bestaan wel oplossingen, deze zijn echter niet eenduidig bepaald. Het numeriek bepalen van de oplossingen van afhankelijke stelsels is een uiterst moeilijke zaak aangezien hier in principe volledig exact gerekend moet worden.

De methoden voor het oplossen van stelsels lineaire vergelijkingen kunnen ingedeeld worden in twee klassen : de directe en de indirecte methoden. Met de directe methoden bepaalt men in eindig veel operaties de exacte oplossing, althans indien geen afrondingsfouten gemaakt worden. Bij de iteratieve methoden bepaalt men een rij approximaties $\underline{x}_1, \underline{x}_2, \ldots$ voor de oplossing. Hier is in principe dus steeds een afbreekfout aanwezig. De invloed van afrondingsfouten is echter vaak minder groot dan bij de directe methoden.

2.2. Directe methoden

2.2.1. Triangulaire stelsels

Beschouw een stelsel van de vorm

$$\begin{aligned} x_1 + c_{12}x_2 + \ldots + c_{1n}x_n &= b_1 \\ x_2 + \ldots + c_{2n}x_n &= b_2 \\ \ldots\ldots\ldots\ldots\ldots & \\ x_n &= b_n. \end{aligned}$$

Dit stelsel is onmiddellijk op te lossen :

$$\begin{aligned} x_n &= b_n \\ x_{n-1} &= b_{n-1} - c_{n-1,n}x_n \\ \ldots\ldots\ldots\ldots\ldots & \\ x_i &= b_i - \sum_{j=i+1}^{n} c_{ij}x_j. \end{aligned}$$

Hoeveel vermenigvuldigingen zijn hiervoor nodig? Voor de berekening van x_i (als $x_{i+1},\ldots,x_n$ al bekend zijn) zijn het er $n - i$. In totaal dus

$$\sum_{i=1}^{n} (n - i) = \tfrac{1}{2}\, n(n - 1).$$

Stel dat er bij de berekening van x_{n-1} een afrondingsfout δ_1 gemaakt wordt. Deze plant zich voort (afgezien van verdere afrondingsfouten) in de berekende waarden van $x_{n-2},\ldots,x_1$ met bedragen $-c_{n-2,n}\,\delta_1,\ldots,-c_{1,n}\,\delta_1$.
Analoog voor een additionele afrondingsfout δ_2 bij de berekening van x_{n-2}, etc.
Zonder op details in te gaan merken we op dat het effect van de voortplanting van deze afrondingsfouten in het algemeen geringer zal zijn naarmate de absolute waarde van de coëfficiënten c_{ij} kleiner is.

2.2.2. De eliminatiemethode van Gauss

Beschouw een stelsel van de vorm

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n &= b_1 \\ \ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1 + \ldots + a_{nn}x_n &= b_n. \end{aligned} \qquad (1)$$

Kunnen we dit stelsel in de triangulaire vorm brengen ? Stellig zijn niet alle coëfficiënten $a_{11},\ldots,a_{1n}$ nul(anders was det A nul). Stel dat $a_{11} \neq 0$

(is dit niet het geval dan vernummeren we de onbekenden). Deel de eerste vergelijking door a_{11}. Dit levert een vergelijking van de vorm

$$x_1 + a_{12}^{(1)} x_2 + \ldots + a_{1n}^{(1)} x_n = b_1^{(1)}. \qquad (2a)$$

Vermenigvuldig nu deze vergelijking met a_{j1} en trek hem af van de j-de vergelijking van (1). Dit levert n - 1 vergelijkingen van de vorm

$$\begin{aligned} a_{22}^{(1)} x_2 + \ldots + a_{2n}^{(1)} x_n &= b_2^{(1)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n2}^{(1)} x_2 + \ldots + a_{nn}^{(1)} x_n &= b_n^{(1)}. \end{aligned} \qquad (2b)$$

Samen met (2a) is dit een stelsel vergelijkingen dat equivalent is met (dezelfde oplossing heeft als) het stelsel (1). Merk op dat, indien de matrix van het stelsel (2) $A^{(1)}$ genoemd wordt, geldt

$$\det A = a_{11} . \det A^{(1)}.$$

Behandel nu het stelsel (2b) op dezelfde wijze. Veronderstel dat $a_{22}^{(1)} \neq 0$ (anders worden de onbekenden $x_2,\ldots,x_n$ vernummerd). Deel de eerste vergelijking van (2b) door $a_{22}^{(1)}$. Etc. Dan krijgen we het stelsel *)

$$\begin{aligned} x_1 + a_{12}^{(2)} x_2 + a_{13}^{(2)} x_3 + \ldots + a_{1n}^{(2)} x_n &= b_1^{(2)} \\ x_2 + a_{23}^{(2)} x_3 + \ldots + a_{2n}^{(2)} x_n &= b_2^{(2)} \\ a_{33}^{(2)} x_3 + \ldots + a_{3n}^{(2)} x_n &= b_3^{(2)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n3}^{(2)} x_3 + \ldots + a_{nn}^{(2)} x_n &= b_n^{(2)} \end{aligned} \qquad (3)$$

dat weer equivalent is met (1). Voorts geldt

$$\det A = a_{11} . a_{22}^{(1)} . \det A^{(2)}.$$

*) Indien geen vernummering plaats gehad heeft dan is de eerste vergelijking van (3) dezelfde als (2a).

Het is duidelijk dat men, op deze manier doorgaande, tot een triangulair stelsel van de in 2.2.1 beschouwde vorm komt. Bovendien geldt

$$\det A = a_{11} \cdot a_{22}^{(1)} \cdot a_{33}^{(2)} \cdots a_{nn}^{(n-1)} . \qquad (4)$$

Het aantal operaties om van (1) op (2) te komen is (afgezien van optellingen, aftrekkingen en vernummeringen) : n delingen en (n-1)n vermenigvuldigingen (waarvan 1 deling en n-1 vermenigvuldigingen die betrekking hebben op de kolom der rechterleden). Totaal dus n^2 operaties (vermenigvuldigingen en delingen worden voortaan over een kam geschoren). Van (2) naar (3) zijn $(n-1)^2$ operaties nodig. Etc. Voor de reductie naar de triangulaire vorm zijn dus

$$\sum_{j=1}^{n} j^2 = \frac{1}{3} n^3 + \frac{1}{2} n^2 + \frac{1}{6} n$$

operaties nodig (waarvan $\sum_{1}^{n} j = \frac{1}{2} n(n+1)$ voor de kolom der rechterleden).
Voor het oplossen van het triangulaire stelsel zijn nog $\frac{1}{2} n(n-1)$ operaties nodig, zodat in totaal voor de oplossing van het stelsel (1)

$$\frac{1}{3} n^3 + n^2 - \frac{1}{3} n$$

vermenigvuldigingen en delingen nodig zijn.
Moet men het stelsel (1) oplossen voor m verschillende kolommen van rechterleden (en dezelfde matrix A) dan kan men al deze kolommen tegelijk meebehandelen. Dit kost $(m-1) \cdot \frac{1}{2} n(n+1)$ extra operaties voor de reductie naar de triangulaire vorm zodat hiervoor dan in totaal

$$\frac{1}{3} n^3 - \frac{1}{3} n + \frac{1}{2} mn(n+1) \qquad (5)$$

operaties nodig zijn. De m verschillende triangulaire stelsels moeten afzonderlijk worden opgelost, zodat het totaal nu komt op

$$\frac{1}{3} n^3 - \frac{1}{3} n + mn^2 . \qquad (6)$$

Voor de bepaling van det A met formule (4) behoeft men geen kolom rechterleden mee te nemen bij de reductie op de triangulaire vorm. Deze kost dan dus (volgens (5)) $\frac{1}{3}n^3 - \frac{1}{3}n$ operaties. Daar komen nog n - 1 vermenigvuldigingen bij voor de toepassing van (4) zodat het totaal voor de berekening van det A wordt

$$\frac{1}{3}n^3 + \frac{2}{3}n - 1. \qquad (7)$$

Voor de berekening van A^{-1} door in het rechterlid van (1) achtereenvolgens de n eenheidsvectoren $\underline{e}_1, \ldots, \underline{e}_n$ te nemen zijn volgens (6) $\frac{4}{3}n^3 - \frac{1}{3}n$ operaties nodig. Houdt men er echter rekening mee dat de rechterleden nu hoofdzakelijk uit nullen bestaan dan kan dit aantal nog wat gereduceerd worden. Bij het in triangulaire vorm brengen van het stelsel $A\underline{x}_1 = \underline{e}_1$ speelt vanaf het begin (de stap van (1) naar (2)) de kolom der rechterleden volledig mee. Maar bij het stelsel $A\underline{x}_2 = \underline{e}_2$ spelen de rechterleden voor het eerst pas een rol bij de overgang van (2) naar (3). Etc. Derhalve zijn bij de herleiding tot de triangulaire vorm hier niet $n.\frac{1}{2}n(n+1)$ operaties op de rechterleden nodig, doch slechts

$$\sum_{j=1}^{n} \frac{1}{2}j(j+1) = \frac{1}{6}n^3 + \frac{1}{2}n^2 + \frac{1}{3}n.$$

Brengt men dit in rekening dan wordt het totale aantal operaties voor de berekening van A^{-1} precies n^3. Dit is frappant weinig - wil men na afloop het resultaat verifiëren door AA^{-1} uit te rekenen dan zijn daarvoor nog eens n^3 vermenigvuldigingen nodig! Overigens is het rekening houden met het feit dat er nullen staan in de rechterleden bij het gebruik van automatische rekenmachines niet aanbevelenswaard aangezien dit nogal wat complicaties bij de programmering geeft.

We geven geen behandeling van de invloed van afrondingsfouten. Men kan inzien dat het in het algemeen gunstig is, de vergelijkingen en de onbekenden zo te vernummeren dat a_{11} in absolute waarde de grootste coëfficiënt van de matrix A is. Na de reductie van (1) op (2) vernummert men de laatste n - 1 vergelijkingen en de laatste n - 1 onbekenden zodat

$|A_{22}^{(1)}| \geqslant |A_{ij}^{(1)}|$ voor $i \geqslant 2$, $j \geqslant 2$. Etc. Met name bereikt men op deze manier dat alle coëfficiënten van het resulterende triangulaire stelsel in absolute waarde hoogstens één zijn. Men noemt deze aanpak van de eliminatie <u>pivotal condensation</u> (pivot = vleugelman (militaire term)). Vaak past men deze strategie slechts gedeeltelijk toe, door bv. bij de overgang van (1) op (2) alleen te kijken naar de grootste coëfficiënt van de eerste rij van A (de vergelijkingen worden dus niet onderling verwisseld).

Bij de practische uitvoering van de eliminatie noteert men alleen de coëfficiënten en laat men de verwisseling van rijen en kolommen achterwege. Vaak neemt men nog een controle-kolom mee : de getallen $c_i = \sum_{j=1}^{n} a_{ij} + b_i$. Voert men op deze kolom dezelfde bewerkingen uit als op de kolom der rechterleden dan blijkt dat de eigenschap dat de termen uit deze kolom de som zijn van de overige termen uit dezelfde rij van het schema, behouden blijft. Dit berust in wezen op het feit dat uit

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad \text{volgt dat} \quad \sum_{j=1}^{n} a_{ij}(x_j + 1) = b_i + \sum_{j=1}^{n} a_{ij} = c_i.$$

<u>Voorbeeld</u>

$$\begin{aligned} 3.21\,x_1 + 4.25\,x_2 + 6.71\,x_3 &= 11.05 \\ 6.38\,x_1 - 0.92\,x_2 + 4.36\,x_3 &= 4.71 \\ 2.37\,x_1 - 5.11\,x_2 + 1.03\,x_3 &= 6.38 \end{aligned}$$

Men noteert dit als volgt :

	a_{i1}	a_{i2}	a_{i3}	b_i	c_i	
1)	3.21	4.25	<u>6.71</u>	11.05	25.82	
2)	6.38	-0.92	4.36	4.71	14.53	
3)	2.37	-5.11	-1.03	6.38	2.61	
4)	0.4784	0.6334	1	1.6468	3.7586	(1) : 6.71
5)	4.2942	-3.6816		-2.4700	-1.8575	(2) - 4.36 x (4)
6)	2.8628	<u>-4.4576</u>		8.0762	6.4814	(3) + 1.03x (4)
7)	-0.6422	1		-1.8118	-1.4540	(6) : (-4.4576)
8)	<u>1.9299</u>			-9.1403	-7.2105	(5) + 3.6816 x (7)
9)	1			-4.7362	-3.7362	(8) : 1.9299

De rijen (4), (5) en (6) corresponderen met het stelsel $A^{(1)}\underline{x} = \underline{b}^{(1)}$, de rijen (4), (7) en (8) met $A^{(2)}\underline{x} = \underline{b}^{(2)}$, de rijen (4), (7) en (9) met het stelsel $A^{(3)}\underline{x} = \underline{b}^{(3)}$, dat triangulair is. De pivots waren resp. 6.71, -4.4576 en 1.9299.
De oplossing van het triangulaire stelsel (9), (7), (4) is $x_1 = -4.7362$, $x_2 = -4.8524$, $x_3 = 6.9867$.
Substitutie in het oorspronkelijke stelsel levert voor de rechterleden 11.0506, 4.7102, 6.3798. Hieruit volgt dat (daar het stelsel niet ill-conditioned is) dat de fouten in de oplossing niet groter dan enkele eenheden van de laatste decimaal zullen zijn.
De determinant van het stelsel is het product van de pivots dus -57.7423 (dus redelijk groot ten opzichte van de coëfficiënten, hetgeen een aanwijzing voor een goede conditie is).

2.2.3. Varianten op de eliminatie methode

Er bestaan vele varianten op het in 2.2.2 behandelde eliminatieschema. Een voor de hand liggende variant is de zg. Gauss-Jordan eliminatie. Hier herleidt men het stelsel $A\underline{x} = \underline{b}$ meteen tot een zg. diagonaal-stelsel (en niet slechts tot de triangulaire vorm). Zij de eliminatie volgens 2.2.2 gevorderd tot

$$\begin{array}{rcl}
x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + \dots + a_{1n}^{(1)} x_n &=& b_1^{(1)} \\
x_2 + a_{23}^{(2)} x_3 + \dots + a_{2n}^{(2)} x_n &=& b_2^{(2)} \\
a_{32}^{(1)} x + a_{33}^{(1)} x_3 + \dots + a_{3n}^{(1)} x_n &=& b_3^{(1)} \\
\dots\dots\dots\dots\dots\dots\dots\dots & & \\
a_{n2}^{(1)} x_2 + a_{n3}^{(1)} x_3 + \dots + a_{nn}^{(1)} x_n &=& b_n^{(1)}
\end{array}$$

Dan kan men met behulp van de tweede vergelijking x_2 niet alleen uit de derde t/m n-de vergelijking elimineren (zodat het stelsel (3) uit 2.2.2 zou ontstaan) maar ook uit de eerste vergelijking. Gaat men zo door dan is de matrix van het gereduceerde stelsel de eenheids matrix, zodat de rechterleden van dit stelsel de oplossing x_1 , $x_2, \dots, x_n$ vormen.

Het Gauss-Jordan schema eist iets meer vermenigvuldigingen dan het schema van Gauss. De procedure is echter iets systematischer.

Een andere variant is de zg. reductiemethode van Crout. Deze methode is er op gericht het aantal te noteren tussen-resultaten zo kleine mogelijk te maken. Dit is met name van belang bij het werken met tafelmachines. Met een normale tafelmachine (waarmee vormen als $a + \sum_{j=1}^{k} b_j c_j$ uitgerekend kunnen worden zonder tussenresultaten te noteren) komt men uit met het noteren van $n(n + 1)$ tussenresultaten (bij de Gauss-methode zijn dit er $\frac{1}{3} n^3 + \frac{1}{2} n^2 + \frac{1}{6} n$). Een min of meer ernstig bezwaar van de methode is dat pivotal condensation slechts beperkt mogelijk is.

Tenslotte nog iets over partitioning. Beschouw het stelsel

$$A\underline{x} = \underline{b}, \qquad (1)$$

waarin A een $(m + n) \times (m + n)$ matrix en $\underline{x}$ en $\underline{b}$ vectoren met $m + n$ componenten zijn. Schrijf

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad \underline{x} = \begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix}, \quad \underline{b} = \begin{pmatrix} \underline{b}_1 \\ \underline{b}_2 \end{pmatrix}$$

Hierin is A_{11} een $m \times m$-matrix, A_{12} een $m \times n$-matrix (m rijen, n kolommen) A_{21} een $n \times m$-matrix en A_{22} een $n \times n$-matrix ; de vectoren $\underline{x}_1$ en $\underline{b}_1$ hebben m, de vectoren $\underline{x}_2$ en $\underline{b}_2$ n componenten.
Uit de regels voor de matrix-vermenigvuldiging volgt dat het stelsel (1) equivalent is met

$$\left.\begin{aligned} A_{11}\underline{x}_1 + A_{12}\underline{x}_2 &= \underline{b}_1 \\ A_{21}\underline{x}_1 + A_{22}\underline{x}_2 &= \underline{b}_2. \end{aligned}\right\} \qquad (2)$$

Veronderstel dat A_{11}^{-1} bestaat. Vermenigvuldig de eerste regel van (2) hiermee :

$$\underline{x}_1 + A_{11}^{-1}A_{12}\underline{x}_2 = A_{11}^{-1}\underline{b}_1. \qquad (3a)$$

Vermenigvuldig (3a) met A_{21} en trek af van de tweede regel van (2)

$$(A_{22} - A_{21}A_{11}^{-1}A_{12})\,\underline{x}_2 = \underline{b}_2 - A_{21}A_{11}^{-1}\underline{b}_1. \qquad (3b)$$

Los nu $\underline{x}_2$ op uit (3b) en vervolgens $\underline{x}_1$ uit (3a)

$$\left.\begin{aligned} \underline{x}_1 &= A_{11}^{-1}\underline{b}_1 - A_{11}^{-1}A_{12}\underline{x}_2 \\ \underline{x}_2 &= (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}\,(\underline{b}_2 - A_{21}A_{11}^{-1}\underline{b}_1). \end{aligned}\right\} \qquad (4)$$

Hiermee is dus de oplossing $\underline{x}$ van (1) bepaald.
In totaal zijn hierbij n + 1 m × m-stelsels opgelost (nl. de stelsels $A_{11}\underline{u} = \underline{b}$ en $A_{11}U = A_{12}$) en één n × n-stelsel (het stelsel 3b)). Verder één vermenigvuldiging van matrices met elkaar en twee vermenigvuldigingen van matrices met vectoren.
Het voordeel van deze methode (waarbij precies hetzelfde aantal vermenigvuldigingen en delingen nodig zijn als bij de directe behandeling van (1)) is dat men per stap van de berekening slechts te maken heeft met m × m- of n × n-stelsels. Dit kan een groot voordeel zijn bij automatische rekenmachines met beperkte geheugencapaciteit.

Opmerkingen

1. Merk op dat de overgang van (2) naar (3) formeel precies dezelfde is als bij de Gaussreductie van een 2 × 2-stelsel naar de triangulaire vorm. En de overgang van (3) naar (4) correspondeert met de oplossing van het triangulaire stelsel. Het is duidelijk dat men de matrix A ook in meer delen kan splitsen.
2. Het is noodzakelijk dat A_{11}^{-1} bestaat. Is dit niet het geval dan moetem de onbekenden vernummerd worden. Uit het bestaan van A^{-1} en A_{11}^{-1} volgt dat ook $(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$ bestaat.
3. Uit (4) volgt dat

$$A^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

met

$$\begin{aligned} B_{11} &= A_{11}^{-1} + A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} \\ B_{12} &= -A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \\ B_{21} &= -(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} \\ B_{22} &= (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}\,. \end{aligned}$$

2.3. Iteratieve methoden

2.3.1. Normen voor vectoren en matrices

2.3.1.1. Beschouw een vectorruimte R. R heet genormeerd indien er een voorschrift is dat aan iedere $\underline{x} \in R$ een reeel getal $\|\underline{x}\|$ (de norm van $\underline{x}$) toevoegt, zodanig dat

1) $\|\underline{x}\| > 0$ als $\underline{x} \neq \underline{o}$ $\quad \|\underline{o}\| = 0$
2) $\|\alpha \underline{x}\| = |\alpha| . \|\underline{x}\|$ (α een getal)
3) $\|\underline{x} + \underline{y}\| \leqslant \|\underline{x}\| + \|y\|$.

Voorbeelden. Zij R de n-dimensionale cartesische ruimte R_n (met als elementen de rijtjes van n getallen $\underline{x} = (x_1, \ldots, x_n)$). Dan kunnen we bv. als norm kiezen

$$\|\underline{x}\|_1 = \sum_j |x_j|$$

of $$\|\underline{x}\|_2 = \left(\sum_j |x_j|^2\right)^{\frac{1}{2}}$$

of $$\|\underline{x}\|_\infty = \max_j |x_j| .$$

Dat deze normen aan de eisen 1) en 2) voldoen is duidelijk. Ook aan de eis 3) (de zg. driehoeksongelijkheid) is voldaan. Voor de norm $\| \ \|_2$ (de euclidische lengte) is dit bekend.
En

$$\|\underline{x} + \underline{y}\|_1 = \sum_j |x_j + y_j| \leqslant \sum_j \{|x_j| + |y_j|\} = \|\underline{x}\|_1 + \|\underline{y}\|_1$$

$$\|\underline{x} + \underline{y}\| = \max_j |x_j + y_j| \leqslant \max_j \{|x_j| + |y_j|\} \leqslant$$

$$\leqslant \max_{j,k} \{|x_j| + |y_k|\} = \|\underline{x}\|_\infty + \|\underline{y}\|_\infty .$$

N.b. Zij n = 2. Wat is de meetkundige plaats van de uiteinden van de vectoren $\underline{x}$ waarvoor geldt $\|\underline{x}\|_1 = 1$? Idem als $\|\underline{x}\|_\infty = 1$.

Stelling. Zij $\| \ \|$ en $\| \ \|'$ twee normen in een eindig-dimensionale vectorruimte R. Dan zijn er positieve getallen m en M zodanig dat voor alle $\underline{x} \in R$

$$m \|\underline{x}\| \leq \|\underline{x}\|' \leq M \|\underline{x}\|$$

(en dus ook $M^{-1}\|\underline{x}\|' \leq \|\underline{x}\| \leq m^{-1}\|\underline{x}\|'$).

Deze stelling bewijzen we niet. Voor de boven ingevoerde normen in R_n blijkt eenvoudig dat

$$\frac{1}{n}\|\underline{x}\|_1 \leq \frac{1}{n}\|\underline{x}\|_2 \leq \|\underline{x}\|_\infty \leq \|\underline{x}\|_2 \leq \|\underline{x}\|_1 \leq \sqrt{n}\,\|x\|_2 \leq n\,\|x\|_\infty$$

waarmee voor deze gevallen de stelling dus bewezen is.

2.3.1.2. In een genormeerde vectorruimte kunnen we over <u>limieten</u> spreken. Zij $\underline{x}^{(1)}, \underline{x}^{(2)}, \ldots$ een rij elementen uit R. We zeggen dat

$$\lim_{k\to\infty} \underline{x}^{(k)} = \underline{x}$$

(waarbij $\underline{x}$ een element van R is) indien

$$\lim_{k\to\infty} \|\underline{x}^{(k)} - \underline{x}\| = 0.$$

Uit de boven aangehaalde stelling blijkt dat in een eindig-dimensionale vectorruimte het al dan niet bestaan van een limiet onafhankelijk is van de keuze van de norm.
In R_n geldt kennelijk dat

$$\lim \underline{x}^{(k)} = \underline{x} \quad \text{dan en slechts dan indien}$$
$$\lim x_j^{(k)} = x_j, \quad j = 1,\ldots,n.$$

2.3.1.3. Zoals bekend is een lineaire afbeelding A van een vectorruimte in zichzelf een voorschrift dat aan iedere $\underline{x} \in R$ een $\underline{y} = A\underline{x} \in R$ toevoegt zodanig dat

$$A(\alpha\underline{x}) = \alpha A\underline{x} \qquad (\alpha \text{ een getal})$$
$$A(\underline{x} + \underline{y}) = A\underline{x} + A\underline{y}.$$

Het is duidelijk dat het product van een getal α en een lineaire afbeelding A (gedefinieerd door $(\alpha A)\underline{x} = \alpha(A\underline{x})$) en de som van twee lineaire afbeeldingen A en B (gedefinieerd door $(A + B)\underline{x} = A\underline{x} + B\underline{x}$) weer lineaire afbeeldingen zijn. De lineaire afbeeldingen vormen dus zelf weer een

vectorruimte. In deze vectorruimte der lineaire afbeeldingen kunnen we op vele manieren een norm invoeren. Maar het is vooral interessant om dit zo te doen dat er samenhang is met de norm in de oorspronkelijke vectorruimte. Kunnen we niet een norm $\|A\|$ van A zo definiëren dat $\|A\underline{x}\| \leqslant \|A\|.\|\underline{x}\|$ voor alle $\underline{x} \in R$?
Dit kan zeker als R eindig-dimensionaal is. Want met $\|\ \|$ is ook $\|\ \|'$ gedefinieerd door

$$\|\underline{x}\|' = \|A\underline{x}\| + \|\underline{x}\|$$

een norm in R (ga na : waarom kunnen we niet nemen $\|\underline{x}\|' = \|A\underline{x}\|$?)
Volgens de stelling is er dus een (eindig) positief getal M zodanig dat voor alle $\underline{x} \neq \underline{o}$

$$\frac{\|\underline{x}\|'}{\|\underline{x}\|} = \frac{\|A\underline{x}\|}{\|\underline{x}\|} + 1 \leqslant M, \text{ dus } \frac{\|A\underline{x}\|}{\|\underline{x}\|} \leqslant M - 1.$$

Als norm van A kiezen we nu het kleinste getal M - 1 waarvoor de laatste ongelijkheid nog geldt (voor alle $\underline{x} \neq \underline{o}$) :

$$\|A\| = \sup_{x \neq \underline{o}} \frac{\|A\underline{x}\|}{\|\underline{x}\|}$$

hetgeen (per definitie) betekent dat $\|A\underline{x}\| \leqslant \|A\|\,\|\underline{x}\|$ voor alle $\underline{x} \in R$ terwijl er bij iedere $\varepsilon > 0$ een $\underline{x} \in R$ is zodanig dat $\|A\underline{x}\| > (\|A\| - \varepsilon)\,\|\underline{x}\|$.
Is dit nu een norm in de vectorruimte der lineaire afbeelding ? Ja. Want

1) $\|A\| \geqslant 0$ en $\|A\| = 0$ impliceert $\|A\underline{x}\| = 0$, (dus $A\underline{x} = \underline{0}$) voor alle $\underline{x}$, dus A = 0 (de afbeelding die heel R op het nul-element van R afbeeldt)
2) Voor alle $\underline{x} \in R$ is $\|\alpha A\underline{x}\| = |\alpha|\ \|A\underline{x}\|$, waaruit volgt dat $\|\alpha A\| = |\alpha|\ \|A\|$.
3) Voor alle $x \in R$ is $\|(A + B)\underline{x}\| = \|A\underline{x} + B\underline{x}\| \leqslant \|A\underline{x}\| + \|B\underline{x}\| \leqslant$
$\leqslant (\|A\| + \|B\|)\,\|\underline{x}\|$, dus $\|A + B\| \leqslant \|A\| + \|B\|$.

Het product AB van twee lineaire afbeeldingen A en B is gedefinieerd door $(AB)\underline{x} = A(B\underline{x})$. Daar dan voor alle $\underline{x} \in R$ $\|(AB)\underline{x}\| = \|A(B\underline{x})\| \leqslant \|A\|.\|B\underline{x}\| \leqslant$
$\leqslant \|A\|.\|B\|.\|\underline{x}\|$ geldt voor de norm van een product

$$\|AB\| \leqslant \|A\|.\|B\|.$$

In de n-dimensionale cartesische ruimte R_n behoort bij een lineaire afbeelding A een $n \times n$-matrix $\{A_{ij}\}$ zodanig dat

$$(A\underline{x})_i = \sum_j A_{ij} x_j .$$

En omgekeerd kunnen we aan iedere $n \times n$ matrix op deze wijze een lineaire afbeelding toevoegen.

Kunnen we $\|A\|$ uitdrukken in de kentallen van de matrix $\{A_{ij}\}$? Dit is betrekkelijk eenvoudig als we in R_n de $\| \ \|_1$, of de $\| \ \|_\infty$ -norm gebruiken.

<u>Stelling</u> $\quad \|A\|_1 = \max_j \sum_i |A_{ij}|,$

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

<u>Bewijs</u>. a) Noem $\max_j \sum_i |A_{ij}| = M_1$.

Voor iedere $\underline{x} \in R_n$ geldt $\|A\underline{x}\|_1 = \sum_i |(A\underline{x})_i| =$

$$= \sum_i \sum_j |A_{ij} x_j| \leqslant \sum_{ij} |A_{ij}| \ |x_j| =$$

$$= \sum_j |x_j| \sum_i |A_{ij}| \leqslant M_1 \sum_j |x_j| = M_1 \|x\|_1 ,$$

waaruit volgt dat $\|A\|_1 \leqslant M_1$.
Zij k zo dat $\sum_i |A_{ik}| = M_1$. Kies $\underline{x}$ zodat $x_j = \delta_{jk}$.
Dan is $\|\underline{x}\|_1 = 1$ en $\|A\underline{x}\|_1 = \sum_i \left| \sum_j A_{ij} x_j \right| = \sum_i |A_{ik}| = M$.

Dus $\|A\|_1 \geqslant M$.

b) Noem $\max_i \sum_j |A_{ij}| = M_\infty$
Dan is voor iedere $\underline{x} \in R_n$

$$\|A\underline{x}\|_\infty = \max_i \left| \sum_j A_{ij} x_j \right| \leqslant \max_i \sum_j |A_{ij}| |x_j| \leqslant$$

$$\leqslant \|\underline{x}\|_\infty \max_i \sum_j |A_{ij}| = M_\infty \|\underline{x}\|_\infty . \quad \text{Dus } \|A\|_\infty \leqslant M_\infty .$$

Zij k zo dat $\sum_j |A_{kj}| = M_\infty$. Kies $\underline{x}$ zo dat $x_j = \frac{|A_{kj}|}{A_{kj}}$ als $A_{kj} \neq 0$ en $x_j = 1$ als $A_{kj} = 0$. Dan is $\|\underline{x}\|_\infty = 1$. En $|(A\underline{x})_k| = |\sum_j A_{kj}x_j| = \sum |A_{kj}|_n = M_\infty$, terwijl voor $i \neq k$ $|(A\underline{x})_i| = |\sum_j A_{ij}x_j| \leqslant \sum_j |A_{ij}| \leqslant M_\infty$, dus $\|A\underline{x}\| = \max_i |(A\underline{x})_i| = M_\infty$, waaruit volgt dat $\|A\|_\infty \geqslant M_\infty$.

2.3.1.4. Nu we een norm voor lineaire afbeeldingen hebben kunnen we ook spreken over de <u>limiet</u> van een rij lineaire afbeeldingen : $A^{(1)}$, $A^{(2)}$,... . We zeggen

$$\lim_{k\to\infty} A^{(k)} = A$$

(waarbij A een lineaire afbeelding is) indien

$$\lim_{k\to\infty} \|A^{(k)} - A\| = 0.$$

Dit impliceert kennelijk dat voor iedere $\underline{x} \in R$ $\lim A^{(k)}\underline{x} = A\underline{x}$.

Men kan bewijzen dat in een eindig dimensionale R dit limietbegrip weer onafhankelijk is van de keuze van de norm. In R_n geldt voor een lineaire afbeelding A met matrix $\{A_{ij}\}$

$$\|A\|_1 \leqslant \sum_{ij} |A_{ij}| \leqslant n\,\|A\|_1 \quad \text{en}$$

$$\|A\|_\infty \leqslant \sum_{ij} |A_{ij}| \leqslant n\,\|A\|_\infty.$$

Hieruit volgt met name dat $\lim A^{(k)} = A$ dan en slechts dan als $\lim A_{ij}^{(k)} = A_{ij}$ voor alle i en j.

2.3.1.5. Zoals bekend kan een lineaire afbeelding A een inverse hebben, d,w,z. kan er een afbeelding bestaan zodanig dat $BA = AB = I$, men schrijft dan $B = A^{-1}$. Is R eindig dimensionaal dan bestaat A^{-1} dan en slechts dan als uit $A\underline{x} = \underline{o}$ volgt dat $\underline{x} = \underline{o}$ (A heet dan niet-singulier). In R_n is dit equivalent met : $\det\{A_{ij}\} \neq 0$.

Stelling. Zij de lineaire afbeelding B zodanig dat $\lim_{k\to\infty} B^k = 0$. Dan bestaat $(I - B)^{-1}$ en $(I - B)^{-1} = \sum_{k=0}^{\infty} B^k$.

Bewijs. a) Zij $(I - B)\underline{x} = 0$. Dan is $\underline{x} = B\underline{x} = B^2\underline{x} = \ldots = B^k\underline{x} = \ldots$ Daar uit $\lim B^k = 0$ volgt dat $\lim B^k\underline{x} = \underline{o}$ is dus $\underline{x} = \underline{o}$. Dus $(I - B)^{-1}$ bestaat.

b) Voor $k = 1,2,\ldots$ geldt $(I - B)(I + B + \ldots + B^{k-1}) = I - B^k$ en dus

$$I + B + \ldots + B^{k-1} = (I - B)^{-1} - (I - B)^{-1} B^k.$$

Daar $\|(I - B)^{-1} B^k\| \leqslant \|(I - B)^{-1}\| \, \|B^k\|$ volgt uit het gegeven dat

$$\lim_{k\to\infty} (I + B + \ldots + B^{k-1}) = (I - B)^{-1},$$

d.w.z. dat (met de gebruikelijke notatie voor een oneindige reeks)

$$\sum_{0}^{\infty} B^k = (I - B)^{-1}.$$

Opmerking. Het is duidelijk dat de voorwaarde $\lim B^k = 0$ niet alleen voldoende maar ook nodig is voor de convergentie van de reeks ΣB^k. Een wat beter hanteerbare voorwaarde die voldoende(doch niet nodig) is, is $\|B\| < 1$. Want daar $\|B^k\| \leqslant \|B\|^k$, volgt hieruit dat $\lim B^k = 0$.

2.3.2. De iteratieprocessen van Gauss-Jacobi en Gauss-Seidel

Beschouw het stelsel

$$\begin{aligned} A_{11}x_1 + \ldots + A_{1n}x_n &= b_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ A_{n1}x_1 + \ldots + A_{nn}x_n &= b_n \end{aligned} \qquad (1)$$

en veronderstel dat $A_{jj} \neq 0$, $j = 1,\ldots,n$. Dan kunnen we de vergelijkingen schrijven als

$$x_i = A_{ii}^{-1}(b_i - \sum_{j\neq i} A_{ij}x_j), \quad i = 1,\ldots,n \qquad (2)$$

waarin met $\sum_{j\neq i}$ bedoeld wordt dat we sommeren over alle $j \neq i$.

Zij $x_1^{(k)},\ldots,x_n^{(k)}$ een benadering van de oplossing.

Is dan

$$x_i^{(k+1)} = A_{ii}^{-1}[b_i - \sum_{j\neq i} A_{ij}x_j^{(k)}] \quad , \; i = 1,\ldots,n. \qquad (3)$$

een betere benadering ? En kunnen we zo doorgaan ? Men noemt dit proces het iteratieproces van Gauss-Jacobi.

Een voor de hand liggende variant is de volgende.

Stel $x_1^{(k)},\ldots,x_n^{(k)}$ bekend. Bereken hiermee met (3) $x_1^{(k+1)}$. Maar gebruik voor de berekening van $x_2^{(k+1)}$ nu $x_1^{(k+1)}$, $x_3^{(k)},\ldots,x_n^{(k)}$ in plaats van $x_1^{(k)},\ldots,x_n^{(k)}$. Etc. Dit betekent dat (3) vervangen wordt door

$$x_i^{(k+1)} = A_{ii}^{-1}[b_i - \sum_{j=i+1}^{n} A_{ij}x_j^{(k)} - \sum_{j=1}^{i-1} A_{ij}x_j^{(k+1)}] \; . \qquad (4)$$

Dit is het iteratieproces van Gauss-Seidel.

Om de convergentie van deze processen te onderzoeken stellen we

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ & & \\ A_{n1} & \cdots & A_{nn} \end{pmatrix}, \quad D = \begin{pmatrix} A_{11} & & 0 \\ & \ddots & \\ 0 & & A_{nn} \end{pmatrix}$$

$$L = \begin{pmatrix} 0 & & \\ A_{21} & \ddots & 0 \\ \cdots\cdots & \ddots & \\ A_{n1} & \cdots & A_{n,n-1} \; 0 \end{pmatrix} \quad R = \begin{pmatrix} 0 & A & \cdots & A_{1n} \\ & \ddots & \ddots & \\ & & \ddots & A_{n-1,n} \\ 0 & & & 0 \end{pmatrix}$$

Dan is $A = D + L + R$. En (1), (3) en (4) zijn equivalent met resp.

$$A\underline{x} = \underline{b} \qquad (1a)$$

$$\underline{x}^{(k+1)} = D^{-1}[\underline{b} - (A - D)\underline{x}^{(k)}] \qquad \text{Jacobi} \qquad (3a)$$

$$\underline{x}^{(k+1)} = D^{-1}[\underline{b} - R\underline{x}^{(k)} - L\,\underline{x}^{(k+1)}] \qquad \text{Gauss-Seidel} \qquad (4a)$$

Uit (4a) volgt *)

←

andere bewijsmethode
geschrapt dictaat p. 14

*) D + L heeft een inverse daar de corresponderende matrix triangulair is en geen nullen in de diagonaal heeft.

(3a) $\underline{x}^{(k+1)} = D^{-1}\underline{b} - D^{-1}(A-D)\underline{x}^{(k)}$

$$\underline{x}^{(k+1)} = (D + L)^{-1}\,\underline{b} - (D + L)^{-1}\,R\underline{x}^{(k)} \qquad (4b)$$

De recursiebetrekkingen (3a) en (4b) zijn van de vorm

$$\underline{x}^{(k+1)} = P\underline{b} + Q\underline{x}^{(k)} \qquad (5)$$

waarbij P en Q zo zijn dat $P^{-1}(I - Q) = A.$ (6)

controle

Stelling. Het iteratieproces (5) convergeert dan en slechts dan voor iedere beginvector $\underline{x}^{(o)}$ naar $A^{-1}\underline{b}$ indien $\lim Q^k = 0$.

Bewijs. Uit (5) volgt dat

$$\underline{x}^{(k+1)} - A^{-1}\underline{b} = Q[\underline{x}^{(k)} - A^{-1}\underline{b}] + [P - A^{-1} + QA^{-1}]\underline{b}$$

$$= Q[\underline{x}^{(k)} - A^{-1}\underline{b}],$$

daar (6) equivalent is met $P = (1 - Q)A^{-1}$.
Door volledige inductie volgt hieruit

$$\underline{x}^{(k)} - A^{-1}\underline{b} = Q^k[\underline{x}^{(o)} - A^{-1}\underline{b}], \qquad (7)$$

waaruit direct de bewering volgt.

Opmerking. Is $\|Q\| < 1$, dan volgt uit (7) dat $\|\underline{x}^{(k)} - A^{-1}\underline{b}\|$ minstens naar nul gaat als de termen van een meetkundige reeks met reden $\|Q\|$. Men kan $\|Q\|$ de convergentie-factor noemen.

Stelling. Het iteratieproces van Gauss-Jacobi convergeert zeker indien

$$\sum_{j\neq i} |A_{ij}| \leq \gamma\,|A_{ii}|\,, \qquad i = 1,\ldots,n$$

met $\gamma < 1$. De convergentiefactor is minstens γ.

Bewijs. Bij Gauss-Jacobi is $Q = I - D^{-1}A$.
Nu is $(D^{-1})_{ij} = A_{ii}^{-1}\delta_{ij}$, waaruit volgt dat

$$Q_{ij} = \delta_{ij} - \frac{A_{ij}}{A_{ii}} = \begin{cases} 0 & \text{als } i = j \\ -\dfrac{A_{ij}}{A_{ii}} & \text{als } i \neq j. \end{cases}$$

Derhalve is $\|Q\|_\infty = \max_i \sum_j |Q_{ij}| =$

$$= \max_i \sum_{j\neq i} \left|\frac{A_{ij}}{A_{ii}}\right| = \max_i \frac{1}{|A_{ii}|} \sum_{j\neq i} |A_{ij}| \leq \gamma.$$

Opmerking. Men kan diverse analoge convergentie voorwaarden geven, bv. :

$$\sum_{i\neq j} |A_{ij}| < |A_{jj}|, \qquad j = 1,\ldots,n$$

$$\sum_{j\neq i} \left|\frac{A_{ij}}{A_{jj}}\right| < 1 \quad , \qquad i = 1,\ldots,n$$

$$\sum_{i\neq j} \left|\frac{A_{ij}}{A_{ii}}\right| < 1 \quad , \qquad j = 1,\ldots,n$$

Stelling. Het iteratieproces van Gauss-Seidel convergeert zeker indien

$$\sum_{j\neq i} |A_{ij}| \leq \gamma\, |A_{ii}| \quad , \; i = 1,\ldots,n$$

met $\gamma < 1$. De convergentiefactor is minstens γ.

Bewijs. Bij Gauss-Seidel is $Q = -(D + L)^{-1} R$. We zullen bewijzen dat uit het gegeven volgt dat $\|Q\|_\infty \leq \gamma$.
Zij $Q\underline{u} = \underline{v}$. Dan is $R\underline{u} = -(D + L)\underline{v}$, en dus

$$\underline{v} = -D^{-1}[R\underline{u} + L\underline{v}].$$

Hieruit volgt

$$v_i = -A_{ii}^{-1}\left[\sum_j R_{ij}u_j + \sum_j L_{ij}v_j\right] =$$

$$= -A_{ii}^{-1}\left[\sum_{j>i} A_{ij}u_j + \sum_{j<i} A_{ij}v_j\right]$$

Derhalve geldt

$$|v_i| \leq \{|A_{ii}|^{-1} \sum_{j\neq i} |A_{ij}|\} \cdot \max\{\|\underline{u}\|_\infty, \|\underline{v}\|_\infty\},$$

dus $$\|\underline{v}\|_\infty \leq \gamma \max\{\|\underline{u}\|_\infty, \|\underline{v}\|_\infty\}. \qquad (8)$$

Veronderstel dat $\|\underline{v}\|_\infty \geq \|\underline{u}\|_\infty$. Dan moet volgens (8) $\|\underline{v}\|_\infty \leq \gamma \|\underline{v}\|_\infty$ en dus daar $\gamma < 1$, $\|\underline{v}\|_\infty = 0$, dus ook $\|\underline{u}\|_\infty = 0$, dus $\underline{u} = \underline{o}$.
Is dus $\underline{u} \neq \underline{o}$, dan moet $\|\underline{v}\|_\infty < \|\underline{u}\|_\infty$ en dan is volgens (8) $\|\underline{v}\|_\infty \leq \gamma \|\underline{u}\|_\infty$.
Hieruit volgt dat $\|Q\|_\infty \leq \gamma$.

Opmerkingen.

1. De in deze stellingen gegeven voorwaarden zijn voldoende, niet nodig.
2. Men kan voorbeelden geven van stelsels, waarbij Gauss-Jacobi wel en Gauss-Seidel niet convergeert en ook van stelsels waarbij het tegengestelde geldt.
3. Men kan bewijzen dat Gauss-Seidel altijd convergeert indien de matrix A symmetrisch (d.w.z. $A_{ij} = A_{ji}$) en positief definiet (d.w.z. $(A\underline{x},\underline{x}) = \sum_{ij} A_{ij}x_i x_j > 0$ tenzij $\underline{x} = \underline{o}$) is.
4. Het is (zeker bij het werken met automatische rekenmachines) nuttig de formule (4) voor het Gauss-Seidel proces te schrijven als

$$x_i^{(k+1)} = x_i^{(k)} + A_{ii}^{-1}\left[b_i - \sum_{j=i}^{n} A_{ij}x_j^{(k)} - \sum_{j=1}^{i-1} A_{ij}x_j^{(k+1)}\right]$$

Binnen de vierkante haken staat dan het zg. residu dat men verkrijgt door de benaderingsvector $(x_1^{(k+1)},\ldots,x_{i-1}^{(k+1)}, x_i^{(k)},\ldots,x_n^{(k)})$ waarover men beschikt op het moment dat men $x_i^{(k+1)}$ gaat berekenen, te substitueren in de i^e vergelijking. Vermenigvuldigd met A_{ii}^{-1} levert dit de correctie voor $x_i^{(k)}$. Bij een automatische rekenmachine (of als men met potlood en vlakgom werkt) vervangt men nu $x_i^{(k)}$ door $x_i^{(k+1)}$ waardoor de nieuwe benaderingsvector verkregen is. Deze wordt gesubstitueerd in de i + 1-ste vergelijking. Etc. Men kan bv. afspreken (de rekenmachine de opdracht geven) om door te werken tot n opeenvolgende correcties kleiner zijn dan een voorgeschreven tolerantie.

2.3.3. Een kwadratisch convergent proces voor de bepaling van A^{-1}.

Zij X_0 een benadering voor A^{-1}, d.w.z. zij

$$R_0 = I - X_0 A \qquad X_0 A = I - R_0 \qquad (1)$$

"klein". Daar uit (1) volgt dat

$$A^{-1}X_0^{-1} = (I-R_0)^{-1}$$

$$A^{-1} = (I-R_0)^{-1}X_0 \approx (I+R_0)X_0$$

$$A^{-1} = (I - R_0)^{-1}X_0$$

kunnen we vermoeden dat

$$X_1 = (I + R_o)X_o = 2X_o - X_oAX_o$$

een betere benadering voor A^{-1} is.
Algemeen kunnen we definiëren

$$\begin{aligned} R_k &= I - X_kA, \\ X_{k+1} &= (I + R_k)X_k = 2X_k - X_k AX_k . \end{aligned} \tag{2}$$

<u>Stelling</u>. Zij de rij $X_1, X_2, \ldots$ bepaald (uitgaande van zekere X_o) door (2).
Zij $\lim R_k = 0.$ (3)
Dan bestaat A^{-1} en

$$\lim X_k = A^{-1} \tag{4}$$

Bovendien geldt

$$\|R_{k+1}\| \leq \|R_k\|^2 \tag{5}$$

$$\|A^{-1} - X_{k+1}\| \leq \|A\| . \|A^{-1} - X_k\|^2 , \tag{6}$$

d.w.z. als het proces convergeert dan convergeert het kwadratisch.
<u>Bewijs</u>. a. Zij $A\underline{x} = 0$. Dan is $\underline{x} = (I - X_kA)\underline{x} = R_k\underline{x}$. Dus $\underline{x} = 0$ (waarom), dus A^{-1} bestaat.
b. Uit (2) volgt $A^{-1} - X_k = R_kA^{-1}$. (7)
Uit (3) volgt dus (4).
c. Uit (2) volgt ook

$$R_{k+1} = I - (I + R_k)X_kA = I - (I + R_k)(I - R_k) = R_k^2. \tag{8}$$

Hieruit volgt (5).
d. Uit (7), (8) en (2) volgt

$$A^{-1} - X_{k+1} = R_{k+1}\, A^{-1} = R_k^2A^{-1} = (I - X_kA)(I - X_kA)A^{-1} =$$

$$= (A^{-1} - X_k)A\,(A^{-1} - X_k).$$

Hieruit volgt (6).

Opmerkingen.

1. Uit (8) volgt door volledige inductie

$$R_k = R_o^{2^k} \tag{9}$$

Het proces convergeert dus stellig indien $\|R_o\| < 1$.

2. De formule (2) is formeel dezelfde als de formule die we krijgen bij Newton-iteratie van de vergelijking $a - \frac{1}{x} = 0$ (vgl. 1.3.2, voorbeeld 2).

3. Uit (2) en (9) volgt

$$X_{k+1} = (1 + R_k)X_k = \ldots = (1 + R_k)\ldots(1 + R_o)X_o =$$

$$= (1 + R_o^{2^k})\ldots(1 + R_o^2)(1 + R_o)X_o. \tag{10}$$

Men zou dus, uitgaande van X_o, R_o kunnen berekenen, vervolgens R_o^2, R_o^4, etc. en dan met (10) de rij X_1, X_2,... . Het proces is dan echter geen iteratieproces meer (omdat R_k niet als een residu berekend wordt) en afrondingsfouten worden niet weggedempt.

4. Uit (1) volgt $A^{-1} = (I - R_o)^{-1} X_o$. Formule (10) is dus formeel dezelfde als de bekende formule

$$(1 - z)^{-1} = (1 + z)(1 + z^2)\ldots(1 + z^{2^k})\ldots$$

(die geldt voor $|z| < 1$).

5. In plaats van (2) kan men ook het proces

$$R_k = I - X_k A$$

$$X_{k+1} = (I + R_k + \ldots + R_k^{p-1})X_k$$

gebruiken. Bewijs dat dit proces de orde p heeft.

6. Het hier beschreven proces wordt met name toegepast om de invloed van afrondingsfouten op een met behulp van een eliminatie-methode berekende "schatting" van A^{-1} te verminderen.

Helaas bestaat er geen analoog proces voor het oplossen van een enkel stelsel lineaire vergelijkingen.

7. Zij A een matrix met diagonaaltermen ongelijk aan nul. Kies $X_o = D^{-1}$ (D is het diagonaal gedeelte van A). Bewijs dat het proces zeker convergeert (en - althans op de duur - heel wat sneller dan Gauss-Jacobi of Gauss-Seidel) indien

$$\sum_{j \neq i} |A_{ij}| < |A_{ii}|, \qquad i = 1,\ldots,n.$$

3. Het bepalen van eigenwaarden van een matrix

3.1. Inleiding

Zij A een lineaire afbeelding van een vectorruimte R in zichzelf. Een getal λ heet eigenwaarde van A indien $A - \lambda I$ (meestal kortheidshalve geschreven als $A - \lambda$) singulier is, d.w.z. indien er een $\underline{x} \neq \underline{o}$ in R is zodanig dat

$$A\underline{x} = \lambda\underline{x}.$$

De vector $\underline{x}$ heet eigenvector.

We beperken ons verder tot het geval dat R de n-dimensionale cartesische ruimte R_n is.

Zoals bekend, is λ dan en slechts dan een eigenwaarde van een matrix (lineaire afbeelding) A indien

$$\det (A - \lambda I) = 0.$$

Dit is een n-de graads vergelijking in λ, (de zgn. karakteristieke vergelijking) die dus n oplossingen heeft, waaronder echter gelijke kunnen voorkomen.
Zijn $\lambda_1,\ldots,\lambda_k$ onderling verschillende eigenwaarden van A en $\underline{x}_1,\ldots,\underline{x}_k$ bijbehorende eigenvectoren dan zijn $\underline{x}_1,\ldots,\underline{x}_k$ onafhankelijk.
Bij een enkelvoudige wortel van de karakteristieke vergelijking behoort slechts één onafhankelijke eigenvector (die op een getal factor na eenduidig bepaald is). Bij een m-voudige wortel van de karakteristieke vergelijking kunnen m onafhankelijke eigenvectoren behoren, maar het kan ook zijn dat er minder zijn (d.w.z. dat er m_1 ($m_1 < m$) onafhankelijke eigenvectoren bij deze eigenwaarde zijn zodanig dat iedere andere eigenvector bij deze eigenwaarde een lineaire combinatie is van deze m_1 eigenvectoren). Deze laatste - uiterst onplezierige, want allerlei uitzonderingsgevallen in het leven roepende - situatie treedt niet op indien A symmetrisch (d.w.z. $A_{ij} = A_{ji}$) is.

Indien er n onafhankelijke eigenvectoren $\underline{x}_1,\ldots,\underline{x}_n$ van A bestaan (dit is altijd het geval indien alle eigenwaarden van A verschillend zijn en ook als A symmetrisch is ; het kan maar hoeft niet het geval te zijn in de overige gevallen) dan vormen deze een basis voor R_n, d.w.z. iedere $\underline{x} \in R_n$ kan geschreven worden als

$$\underline{x} = \sum_1^n \alpha_i \underline{x}_i . \qquad (1)$$

En dan geldt $A\underline{x} = \sum \lambda_i \alpha_i \underline{x}_i$ (2)

als λ_i de eigenwaarde is die behoort bij $\underline{x}_i$).

Op deze basis voor R_n heeft A dus een zeer eenvoudig karakter.

Is A symmetrisch dan zijn, zoals bekend, alle eigenwaarden reeel (in het algemeen kan een reële A ook paren toegevoegd complexe eigenwaarden hebben), eigenvectoren bij verschillende eigenwaarden zijn onderling orthogonaal en bij een m-voudige eigenwaarde kunnen m onderling orthogonale eigenvectoren gekozen worden. Dit berust allemaal op het feit dat voor een symmetrische A voor iedere $\underline{x}$ en $\underline{y}$ geldt dat

$$(A\underline{x},\underline{y}) = \sum_{ij} A_{ij}x_j y_i = \sum_{ij} A_{ji}x_j y_i = (\underline{x},A\underline{y}).$$

Is A niet symmetrisch dan definieert men de matrix A* (ook wel $\tilde{A}$ geschreven) door

$$(A^*)_{ij} = A_{ji}.$$

Er geldt (ga na) $(A^*)^* = A$, $(AB)^* = B^*A^*$.

Daar $\det(A - \lambda) = \det(A^* - \lambda)$ hebben A en A* dezelfde eigenwaarden (maar in het algemeen verschillende eigenvectoren).

<u>Stelling 1</u>. Er geldt $\lim A^m = 0$ dan en slechts dan als $|\lambda| < 1$ voor alle eigenwaarden λ van A.

<u>Bewijs</u>. a. Zij $\lim A^m = 0$, zij λ eigenwaarde van A met eigenvector $\underline{x}$. Dan is $\lim (\lambda^m \underline{x}) = \lim A^m \underline{x} = 0$ en dus $|\lambda| < 1$, daar $\underline{x} \neq \underline{o}$. De voorwaarde is dus nodig.

b. Dat de voorwaarde ook voldoende is bewijzen we alleen voor het geval dat A n onafhankelijke eigenvectoren $\underline{x}_1,\ldots,\underline{x}_n$ heeft (de stelling geldt ook in de andere gevallen).

Uit (1) en (2) volgt dan nl. dat iedere vector $\underline{x}$ geschreven kan worden als

$$\underline{x} = \sum_{1}^{n} \alpha_j \underline{x}_j$$

en dat

$$A^m \underline{x} = \sum_{1}^{n} \lambda_j^m \alpha_j \underline{x}_j$$

Uit $|\lambda_j| < 1$ $(j = 1,\ldots,n)$ volgt dus $\lim A^m \underline{x} = 0$ voor iedere $\underline{x}$ en dus $\lim A^m = 0$.

Stelling 2. Voor iedere keuze van de norm geldt dat $|\lambda| \leqslant \|A\|$ voor alle eigenwaarden λ van A.

Bewijs. Uit $A\underline{x} = \lambda \underline{x}$ volgt $|\lambda| = \frac{\|A\underline{x}\|}{\|\underline{x}\|} \leqslant \|A\|$.

Opmerking. Het $<$-teken kan voor alle eigenwaarden gelden. Voorbeeld : $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$: beide eigenwaarden zijn nul en in iedere norm is $\|A\| > 0$ (daar $A \neq 0$).

Stelling 3. Zij μ_1 de grootste eigenwaarde van de matrix A*A (alle eigenwaarden van deze matrix zijn reeel en $\geqslant 0$). Dan is

$$\|A\|_2 = \sqrt{\mu_1}$$

Bewijs. A*A is symmetrisch. Zij $\mu_1 \geqslant \mu_2 \ldots > \mu_n$ de eigenwaarden en $\underline{x}_1 \ldots \underline{x}_n$ bijbehorende onderling orthogonale eigenvectoren.
Dan is $\mu_k(\underline{x}_k, \underline{x}_k) = (A^*A\underline{x}_k, \underline{x}_k) = (A\underline{x}_k, A\underline{x}_k)$, waaruit volgt dat voor alle k $0 \leqslant \mu_k \leqslant (\|A\|_2)^2$.
Anderzijds geldt, voor $\underline{x} = \sum \alpha_k \underline{x}_k$ (3)

$$(A\underline{x}, A\underline{x}) = (A^*A\underline{x}, \underline{x}) = \sum \mu_k \alpha_k^2 (\underline{x}_k, \underline{x}_k)$$

$$\leqslant \mu_1 \sum \alpha_k^2 (\underline{x}_k, \underline{x}_k) = \mu_1 (\underline{x}, \underline{x}).$$

Daar iedere $\underline{x}$ in de vorm (3) geschreven kan worden volgt hieruit dat $\|A\|_2 \leqslant \sqrt{\mu_1}$.

Stelling 4. Zij $|A_{ii}| > \sum_{j \neq i} |A_{ij}|$, $i = 1,\ldots,n$.

Dan is A niet singulier.

Bewijs. Zij $A\underline{x} = 0$. Dan is

$$x_i = -A_{ii}^{-1} \sum_{j \neq i} A_{ij} x_j, \qquad i = 1,\ldots,n$$

en dus $|x_i| \leq |A_{ii}|^{-1} \sum_{j \neq i} |A_{ij}|\,|x_j|$, waaruit volgt

$$\|\underline{x}\|_\infty \leq \|\underline{x}\|_\infty \cdot \max_i \left(|A_{ii}|^{-1} \sum_{j \neq i} |A_{ij}| \right)$$

Was $\|\underline{x}\|_\infty \neq 0$ dan volgt hieruit een tegenspraak met het gegeven.

Gevolg. Indien het complexe getal λ zo is dat

$$|A_{ii} - \lambda| > \sum_{j \neq i} |A_{ij}| , \quad i = 1,..,n$$

dan is λ geen eigenwaarde van A. Want volgens de stelling is $A - \lambda I$ niet singulier.

Positiever gezegd (stelling van Gershgorin) :

Iedere eigenwaarde van A ligt binnen of op ten minste één van de n cirkels (in het complexe vlak)

$$|A_{ii} - \lambda| = \sum_{j \neq i} |A_{ij}| , \quad i = 1,\ldots,n.$$

En ook (daar de eigenwaarden van A^* dezelfde zijn als die van A) binnen of op ten minste een van de cirkels

$$|A_{jj} - \lambda| = \sum_{i \neq j} |A_{ij}| , \quad j = 1,\ldots,n.$$

Opmerkingen

1. De stelling van Gershgorin geeft een (in het algemeen zeer ruwe) localisatie van de eigenwaarden van A.
Men kan ook bewijzen dat als de vereniging van k van de Gershgorincirkels (van dezelfde soort) een samenhangend gebied vormt terwijl de overige cirkels daarbuiten liggen, er in dit gebied precies k eigenwaarden liggen.

2. Zij $a_1,\ldots,a_n$ reële getallen. Beschouw de matrix

$$A = \begin{pmatrix} -a_1 & -a_2 & \ldots\ldots\ldots & -a_n \\ 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Laat zien dat de karakteristieke vergelijking van deze matrix is

$$(-1)^n [\lambda^n + a_1 \lambda^{n-1} + \ldots + a_n] = 0.$$

Toepassing van stelling 2 (met de $\| \ \|_1$ - of de $\| \ \|_\infty$ - norm) en van de stelling van Gershgorin op deze matrix leidt tot de schattingen uit 1.6.1.3. voor de grootste wortel van een n^e graadsvergelijking!

3.2. Een proces voor de bepaling van de grootste eigenwaarde

Veronderstel dat de matrix A n onafhankelijke eigenvectoren $\underline{x}_1,\ldots,\underline{x}_n$ heeft, behorend bij eigenwaarden $\lambda_1,\ldots,\lambda_n$. Veronderstel verder dat

$$|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|. \tag{1}$$

Zij $\underline{y}$ een willekeurige vector. Dan zijn er getallen $\alpha_1,\ldots,\alpha_n$ zodanig dat

$$\underline{y} = \alpha_1 \underline{x}_1 + \ldots + \alpha_n \underline{x}_n. \tag{2}$$

Dan geldt voor $m = 1,2,\ldots$

$$A^m \underline{y} = \alpha_1 \lambda_1^m \underline{x} + \ldots + \alpha_n \lambda_n^m \underline{x}_n. \tag{3}$$

Met (1) volgt hieruit

$$\lim_{m \to \infty} \lambda_1^{-m} A^m \underline{y} = \alpha_1 \underline{x}_1. \tag{4}$$

Globaal gezegd : als $\alpha_1 \neq 0$ dan heeft, als m groot is, $A^m \underline{y}$ "vrijwel" de richting van $\underline{x}_1$ en $A^{m+1} \underline{y}$ is "vrijwel" λ maal zo lang als $A^m \underline{y}$. Op deze manier kunnen we dus λ_1 en $\underline{x}_1$ benaderen.
Is $|\lambda_1| > 1$ dan worden op de duur de kentallen van $A^m \underline{y}$ erg groot, is $|\lambda_1| < 1$ dan worden ze erg klein. We kunnen dit verhelpen door na iedere stap te normeren, bv. zo dat we $A^m \underline{y}$ delen door een factor zodanig dat een vast gekozen kental - bij voorkeur het grootste - één wordt. D.w.z. in plaats van de rij $\underline{y}$, $A\underline{y}$, $A^2\underline{y}$,.. beschouwen we de rij $y_0, y_1, y_2, \ldots$ gedefinieerd door

$$\underline{y}_0 = \frac{\underline{y}}{(\underline{y}, \underline{e}_k)}$$

$$\underline{y}_{m+1} = \frac{A\underline{y}_m}{(A\underline{y}_m, \underline{e}_k)}, \quad m = 0,1,\ldots \tag{5}$$

($\underline{e}_k$ is de k-de eenheidsvector, de k^e component van $\underline{y}_{m+1}$ is dus 1)

Uit (5) volgt dat (als $\alpha_1 \neq 0$ en $(\underline{x}, \underline{e}_k) \neq 0$)

$$\underline{y}_m = \frac{A^m \underline{y}}{(A^m \underline{y}, \underline{e}_k)} = \frac{\sum_1^n \alpha_j \lambda_j^m \underline{x}_j}{\sum_1^n \alpha_j \lambda_j^m (\underline{x}_j, \underline{e}_k)} =$$

$$= \frac{\underline{x}_1 + \sum_2^n \frac{\alpha_j}{\alpha_1} \left(\frac{\lambda_j}{\lambda_1}\right)^m \underline{x}_j}{(\underline{x}_1, \underline{e}_k)\left[1 + \sum_2^n \frac{\alpha_j}{\alpha_1} \left(\frac{\lambda_j}{\lambda_1}\right)^m \frac{(\underline{x}_j, \underline{e}_k)}{(\underline{x}_1, \underline{e}_k)}\right]} \qquad (6)$$

en

$$(A\underline{y}_m, \underline{e}_k) = \lambda_1 \cdot \frac{1 + \sum_2^n \frac{\alpha_j}{\alpha_1} \left(\frac{\lambda_j}{\lambda_1}\right)^{m+1} \frac{(\underline{x}_j, \underline{e}_k)}{(\underline{x}_1, \underline{e}_k)}}{1 + \sum_2^n \frac{\alpha_j}{\alpha_1} \left(\frac{\lambda_j}{\lambda_1}\right)^m \frac{(\underline{x}_j, \underline{e}_k)}{(\underline{x}_1, \underline{e}_k)}} . \qquad (7)$$

Derhalve geldt

$$\lim \underline{y}_m = \frac{\underline{x}_1}{(\underline{x}_1, \underline{e}_k)}$$

$$\lim (A\underline{y}_m, \underline{e}_k) = \lambda_1 .$$

We bespreken nog even de veronderstellingen waarop deze afleiding berust.

a. A heeft n onafhankelijke eigenvectoren. Dit is zeker het geval als alle eigenwaarden verschillend zijn of als A symmetrisch is. Is dit niet het geval (doch wel $|\lambda_1| > |\lambda_2|$) dan blijft het proces geldig, de afleiding wordt wat anders.

b. $|\lambda_1| > |\lambda_2|$. Is dit niet het geval (zodat er twee of meer dominante eigenwaarden zijn) dan kan men diverse gevallen onderscheiden, bv. $\lambda_2 = \lambda_1$ $\lambda_2 = -\lambda_1$, $\lambda_2 = \bar{\lambda}_1$, etc. en voor deze gevallen speciale processen ontwerpen. We gaan hier niet verder op in.

c. $\alpha_1 \neq 0$. Is dit niet het geval (t.g.v. een ongelukkige keuze van de beginvector $\underline{y}$) dan zou het proces theoretisch convergeren naar λ_2 en $\underline{x}_2$.

In de practijk krijgt $\underline{y}_m$ op de duur door afrondingsfouten een - aanvankelijk kleine - component in de richting van $\underline{x}_1$ en op de - lange - duur gaat deze het winnen. Bij handrekenen kan men, indien men ziet dat de getallen $(A\underline{y}_m, \underline{e}_k)$ langzaam, maar monotoon, variëren, beter opnieuw beginnen met een andere beginvector.
d. $(\underline{x}_1, \underline{e}_k) \neq 0$. Daar op de duur $\underline{y}_m$ vrijwel de richting van $\underline{x}_1$ heeft zal, als we k zo kiezen dat de k-de component van $\underline{y}_m$ de grootste of ongeveer de grootste is, ook stellig $(\underline{x}_1, \underline{e}_k) \neq 0$ zijn.

Het is duidelijk dat het proces sneller convergeert naarmate $|\frac{\lambda_2}{\lambda_1}|$ (de convergentie-factor) kleiner is. Soms kan men de convergentie versnellen door te werken met de matrix $A_1 = A - \alpha I$. De eigenwaarden hiervan zijn (ga na) $\lambda_1 - \alpha, \ldots, \lambda_n - \alpha$ en de eigenvectoren $\underline{x}_1, \ldots, \underline{x}_n$. Men tracht α zo te kiezen dat $\max_{j=2,\ldots,n} \left|\frac{\lambda_j - \alpha}{\lambda_1 - \alpha}\right|$ zo klein mogelijk is. Hiertoe is een ruwe schatting van de eigenwaarden van A nodig.
Een andere manier om de convergentie te versnellen is het δ^2-proces van Aitken toe te passen op de componenten van drie opvolgende vectoren $\underline{y}_m$, $\underline{y}_{m+1}$ en $\underline{y}_{m+2}$ en de resulterende vector als nieuwe startvector te gebruiken.

Is de matrix A symmetrisch dan kan men de convergentie van de benaderingen voor de eigenwaarde (niet van die voor de eigenvector) als volgt versnellen. We mogen in dit geval veronderstellen dat de eigen vectoren $\underline{x}_1, \ldots, \underline{x}_n$ een orthonormaalstelsel vormen (d.w.z. $(\underline{x}_i, \underline{x}_j) = \delta_{ij}$). Uit (6) volgt dan

$$\frac{(A\underline{y}_m, \underline{y}_m)}{(\underline{y}_m, \underline{y}_m)} = \lambda_1 \cdot \frac{1 + \sum_2^n \left(\frac{\alpha_j}{\alpha_1}\right)^2 \left(\frac{\lambda_j}{\lambda_1}\right)^{2m+1}}{1 + \sum_2^n \left(\frac{\alpha_j}{\alpha_1}\right) \left(\frac{\lambda_j}{\lambda_1}\right)^{2m}} .$$

De convergentiefactor is dan dus $|\frac{\lambda_2}{\lambda_1}|^2$.

Opmerking. Iets dergelijks geldt algemener :
Zij A symmetrisch, zij λ een eigenwaarde met eigenvector $\underline{x}$.

Zij $\underline{y} = \underline{x} + \varepsilon\underline{z}$ ($\underline{z}$ een willekeurige vector).
Dan is (daar $(A\underline{z},\underline{x}) = (\underline{z},A\underline{x}) = \lambda(\underline{z},\underline{x})$)

$$\frac{(A\underline{y},\underline{y})}{(\underline{y},\underline{y})} = \frac{\lambda[(\underline{x},\underline{x}) + \varepsilon(\underline{x},\underline{z}) + \varepsilon(\underline{z},\underline{x})] + \varepsilon^2(A\underline{z},\underline{z})}{(\underline{y},\underline{y})} =$$

$$= \lambda + \varepsilon^2 \frac{(A\underline{z},\underline{z}) - \lambda(\underline{z},\underline{z})}{(\underline{y},\underline{y})}. \qquad (8)$$

Zij nu ε klein. Dan is $\underline{y}$ een benadering voor de eigenvector $\underline{x}$ met een fout "van de orde ε". En uit (8) blijkt dat $\frac{(A\underline{y},\underline{y})}{(\underline{y},\underline{y})}$ dan een benadering voor de bijbehorende eigenwaarde λ is met een fout "van de orde ε^2".

3.3. De overige eigenwaarden

We veronderstellen dat A symmetrisch is met eigenwaarden $\lambda_1,\ldots,\lambda_n$ en onderling orthogonale eigenvectoren $\underline{x}_1,\ldots,\underline{x}_n$.
Veronderstel dat $|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots$ en dat λ_1 en $\underline{x}_1$ al bepaald zijn.
Indien men het proces uit 3.2. zou beginnen met een vector $\underline{y}$ waarvoor geldt dat

$$(\underline{y},\underline{x}_1) = 0 \qquad (1)$$

dan convergeert dit proces (als $(\underline{y},\underline{x}_2) \neq 0$ en als exact gerekend wordt) naar λ_2 en $\underline{x}_2$. Dit volgt uit de herleidingen in 3.2. daar (dank zij de orthogonaliteit van $\underline{x}_1,\ldots,\underline{x}_n$) (1) impliceert dat

$$\underline{y} = \alpha_2\underline{x}_2 + \ldots + \alpha_n\underline{x}_n.$$

Theoretisch zou men dus op deze wijze λ_2 en $\underline{x}_2$ kunnen bepalen. In de practijk is deze methode echter om twee redenen niet bruikbaar.
1. Het voorgestelde proces is z.g. numeriek instabiel. D.w.z., bij exact rekenen convergeert het naar de gewenste grootheden. Doch door afrondingsfouten krijgt $\underline{y}_m$ op de duur een component in de $\underline{x}_1$-richting en daar $|\lambda_1| > |\lambda_2|$ wordt deze op de lange duur meer versterkt dan de component in de $\underline{x}_2$-richting. Rekent men "bijna exact", dus met een groot aantal extra-cijfers dan zal men na een beperkt aantal stappen vrij dicht in de buurt van λ_2 en $\underline{x}_2$ komen, gaat men echter langer door dan gaan de uitkomsten

verlopen en uiteindelijk komt men en blijft men in de buurt van λ_1 en $\underline{x}_1$ ("in de buurt" omdat er bij aanwezigheid van afrondingsfouten geen echte limiet waarde is - vergelijk ook opmerking 3 in 1.2.). Numeriek instabiele processen zijn in de practijk ongewenst, soms echter niet beslist onbruikbaar, mits men met enige voorzichtigheid (extra-cijfers meenemen en niet te lang doorgaan, zodat de geaccumuleerde afrondingsfout bij het beëindigen nog redelijk klein is) te werk gaat.

2. In de practijk kent men $\underline{x}_1$ niet exact doch slechts bij benadering. Want als regel zullen $\underline{x}_1$ en λ_1 benaderd worden met een infiniet proces dat na eindig veel stappen afgebroken wordt. Natuurlijk kan men de daardoor optredende afbreekfouten klein maken door dit proces ver door te zetten. Maar dit kost veel tijd. Bovendien is men als regel primair geïnteresseerd in de eigenwaarden van A en aan het eind van 3.2 bleek dat na een zeker aantal iteratiestappen de benadering voor de eigenwaarden van een symmetrische matrix essentieel beter is dan die voor de eigenvector (dit geldt ook bij andere methoden). Het is dus realistisch om er van uit te gaan dat voor $\underline{x}_1$ slechts een matig nauwkeurige benadering $\tilde{\underline{x}}_1$ bekend is. Kiezen we nu $\underline{y}$ zo dat $(\underline{y},\tilde{\underline{x}}_1) = 0$ (dit is het beste wat we kunnen doen) dan is de component van $\underline{y}$ in de (exacte) $\underline{x}_1$ richting niet nul, doch slechts matig dicht bij nul en bij iteratie zal deze component het na een matig groot aantal stappen gaan winnen van de component in de $\underline{x}_2$-richting.

De volgende modificatie van het bovenstaande proces is wel bruikbaar. Zij $\tilde{\underline{x}}_1$ een benadering voor $\underline{x}_1$ zodanig dat $(\tilde{\underline{x}}_1,\tilde{\underline{x}}_1) = 1$. Kies een willekeurige beginvector $\underline{y}$. Bepaal nu de vectoren $\underline{z}_0,\underline{z}_1,\ldots$ en $\underline{y}_0,\underline{y}_1,\ldots$ als volgt*)

$$\underline{z}_0 = \underline{y} - (\underline{y},\tilde{\underline{x}}_1)\,\tilde{\underline{x}}_1 \qquad (2a)$$

$$\underline{y}_m = \frac{\underline{z}_m}{(\underline{z}_m,\underline{e}_k)}\,, \qquad m \geqslant 0 \qquad (2b)$$

$$\underline{z}_{m+1} = A\underline{y}_m - (A\underline{y}_m,\tilde{\underline{x}}_1)\,\tilde{\underline{x}}_1\,, \qquad m \geqslant 0 \qquad (2c)$$

*) $\underline{e}_k$ is weer een geschikt gekozen basis-vector.

Dan is $\lim_{m\to\infty} \underline{y}_m$ een benadering voor $\frac{\underline{x}_2}{(\underline{x}_2,\underline{e}_k)}$ en $\lim_{m\to\infty} \frac{(\underline{z}_{m+1},\underline{y}_m)}{(\underline{y}_m,\underline{y}_m)}$ is een benadering voor λ_2 .

Bewijs. Definieer de lineaire afbeelding $\tilde{A}_1$ door

$$\tilde{A}_1\underline{y} = A\underline{y} - (A\underline{y},\tilde{\underline{x}}_1)\,\tilde{\underline{x}}_1. \qquad (3)$$

Het is duidelijk dat dit een lineaire afbeelding is. Voor (2b) en (2c) kunnen we dan schrijven

$$\underline{y}_{m+1} = \frac{\tilde{A}_1\underline{y}_m}{(A\underline{y}_m,\underline{e}_k)} .$$

Derhalve passen we in wezen het proces uit 3.2. (met de aan het eind gegeven modificatie voor de bepaling van de eigenwaarde voor het geval van een symmetrische *) matrix : merk op dat

$$\frac{(\underline{z}_{m+1},\underline{y}_m)}{(\underline{y}_m,\underline{y}_m)} = \frac{(\tilde{A}_1\underline{y}_m,\underline{y}_m)}{(\underline{y}_m,\underline{y}_m)} \qquad)$$

Het proces zal dus convergeren naar de grootste eigenwaarde van $\tilde{A}_1$ en de bijbehorende eigenvector.

Zij A_1 gedefinieerd door $A_1\underline{y} = A\underline{y} - (\underline{y},\underline{x}_1)\,\underline{x}_1$. Dan is $A_1\underline{x}_1 = A\underline{x}_1 - (A\underline{x}_1,\underline{x}_1)\,\underline{x}_1 = \lambda_1\underline{x}_1 - \lambda_1(\underline{x}_1,\underline{x}_1)\,\underline{x}_1 = \underline{o}$ en voor $j \geqslant 2$

$$A_1\underline{x}_j = A\underline{x}_j - (A\underline{x}_j,\underline{x}_1)\,\underline{x}_1 = \lambda_j\underline{x}_j - \lambda_j(\underline{x}_j,\underline{x}_1)\,\underline{x}_1 = \lambda_j\underline{x}_j.$$

De eigenwaarden van A_1 zijn dus $0,\lambda_2,\ldots,\lambda_n$ met eigenvector $\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n$. Ligt nu $\tilde{\underline{x}}_1$ dicht bij $\underline{x}_1$ dan ligt $\tilde{A}_1$ "dicht bij" A_1 en het is plausibel dat de eigenwaarden van $\tilde{A}_1$ dan dicht bij $0,\lambda_2,\ldots,\lambda_n$ zullen liggen (en analoog voor de eigenwaarden °)). De grootste eigenwaarde van $\tilde{A}_1$ ligt dus in de buurt van λ_2 .

---

*) De lineaire afbeelding A_1 is niet symmetrisch in R_u. Maar wel in de n-1-dimensionale deelruimte R_n bestaande uit alle vectoren $\underline{y}$ waarvoor $(\underline{y},\tilde{\underline{x}}_1) = s$ - alle vectoren $\underline{y}_m$ liggen in deze deelruimte.

°) Als A niet singulier is dan heeft $\tilde{A}$ een eigenwaarde nul met eigenvector $A^{-1}\tilde{\underline{x}}_1$.

Opmerkingen

1. Men kan bewijzen dat, als $\tilde{\underline{x}}_1 - \underline{x}_1$ "klein is van de orde ε", de afwijkingen tussen de eigenwaarden van $\tilde{A}_1$ en die van A_1 klein zijn van de orde ε^2, terwijl de verschillen tussen de corresponderende eigenvectoren klein zijn van de orde ε. Dit is belangrijk, daar volgens 3.2. uit $\tilde{\underline{x}}_1$ dan voor λ_1 een benadering met fout van de orde ε^2 gevonden kan worden. De fout in λ_2 tengevolge van de onnauwkeurigheid in $\tilde{\underline{x}}_1$ is dus van dezelfde orde als die in λ_1 en analoog voor de eigenvectoren.

2. Het proces berust in wezen hierop dat men na iedere iteratiestap zorgt dat $\underline{y}_m$ loodrecht is op $\tilde{\underline{x}}_1$, d.w.z. ongeveer loodrecht op $\underline{x}_1$, zodat de component in de $\underline{x}_1$-richting geen kans krijgt om te groeien (ook niet door afrondingsfouten).

3. Man kan op deze manier verder gaan. Voor de bepaling van λ_3 gebruikt men (als $\tilde{\underline{x}}_1$ en $\tilde{\underline{x}}_2$ bekend zijn) in plaats van (2c) de formule

$$\underline{z}_{m+1} = A\underline{y}_m - (A\underline{y}_m, \tilde{\underline{x}}_1)\,\tilde{\underline{x}}_1 - (A\underline{y}_m, \tilde{\underline{x}}_2)\,\tilde{\underline{x}}_2.$$

D.w.z. men werkt met een lineaire afbeelding waarvan de eigenwaarden in de buurt liggen van $0, 0, \lambda_3, .., \lambda_n$. Etc.

4. Voor de bepaling van alle eigenwaarden van een vrij grote matrix (in de practijk is men vaak slechts geïnteresseerd in enkele eigenwaarden, meestal in de grootste) is dit proces niet het meest geschikte.

3.4. Triagonaalmatrices

In een aantal toepassingen komen zg. tridiagonaalmatrices voor. Dat zijn matrices waarbij $A_{ij} = 0$ als $|i-j| \geqslant 2$. We schrijven deze matrices in de vorm

$$A = \begin{pmatrix} a_1 & -b_1 & & & 0 \\ -c_1 & a_2 & -b_2 & & \\ & -c_2 & \ddots & \ddots & \\ & & \ddots & \ddots & -b_{n-1} \\ 0 & & & -c_{n-1} & a_n \end{pmatrix}$$

We beperken ons tot het geval dat

$$b_j c_j > 0, \qquad j = 1, \ldots, n-1. \tag{1}$$

Dit geval komt in de practijk veel voor. Merk op dat bij symmetrische matrices $b_j = c_j$, zodat aan (1) voldaan is tenzij voor zekere j $b_j = 0$; in dit laatste geval valt de matrix echter uiteen in deelmatrices die onafhankelijk behandeld kunnen worden (ga na!).
We zoeken de eigenwaarden van de matrix A. Dat zijn, zoals bekend, die getallen λ waarvoor het homogene stelsel $(A - \lambda I)\underline{x} = \underline{o}$ een niet-triviale oplossing heeft. Een kleine generalisatie van dit probleem - die voor sommige toepassingen nuttig is - is de volgende :
Zij D een diagonaalmatrix (d.w.z. $D_{ij} = 0$ als $i \neq j$) met diagonaalelementen $d_1,\ldots,d_n$. We veronderstellen

$$d_j > 0 , \quad j = 1,\ldots,n. \tag{2}$$

We vragen nu naar de getallen λ waarvoor het homogene stelsel $(A - \lambda D)\underline{x} = \underline{o}$ een niet-triviale oplossing heeft. Deze getallen kan men de eigenwaarden van A ten opzichte van D noemen en de bijbehorende vectoren $\underline{x}$ de eigenvectoren van A ten opzichte van D (is D = I dan ontstaan de gewone eigenwaarden en eigenvectoren).

Het is duidelijk dat λ dan en slechts dan eigenwaarde van A ten opzichte van D is indien

$$\Delta_n(\lambda) = \det\{A - \lambda D\} = 0.$$

Zij algemeen, voor $j = 1,\ldots,n$

$$\Delta_j(\lambda) = \begin{vmatrix} a_1 - \lambda d_1 & -b_1 & & 0 \\ -c_1 & \ddots & \ddots & \\ & \ddots & \ddots & -b_{j-1} \\ 0 & & -c_j & a_j - \lambda d_j \end{vmatrix}$$

Ontwikkeling naar de laatste rij levert

$$\Delta_j(\lambda) = (a_j - \lambda d_j)\,\Delta_{j-1}(\lambda) - b_{j-1}c_{j-1}\Delta_{j-2}(\lambda) \tag{3}$$

Deze formule geldt voor $j \geq 2$ mits we stellen

$$\Delta_0(\lambda) = 1.$$

Daar verder

$$\Delta_1(\lambda) = a_1 - \lambda, \qquad a_1 - \lambda d_1 \text{ denk ik.}$$

kunnen $\Delta_2(\lambda)\ldots,\Delta_n(\lambda)$ voor een gegeven waarde van λ eenvoudig berekend worden. Voorts constateren we dat (daar $d_j \neq 0$) Δ_j een polynoom van de

graad j is met kopterm $d_1,\ldots,d_j(-\lambda)^j$. Merk op dat hieruit volgt (met de veronderstelling (2)) dat

$$\operatorname{sgn}\ \Delta_j(\lambda) = \begin{cases} +1 & \text{voor } \lambda \to -\infty \\ (-1)^j & \text{voor } \lambda \to +\infty \end{cases} \qquad (4)$$

Dank zij de veronderstelling (1) bezitten de polynomen Δ_j een aantal fraaie eigenschappen (men noemt een rij polynomen met deze eigenschappen wel een Sturm-rij).

a. Twee opvolgende polynomen Δ_{j-1} en Δ_j hebben geen gemeenschappelijke nulpunten.

Bewijs : Stel $\Delta_j(\lambda_o) = \Delta_{j-1}(\lambda_o) = 0$. Als $j \geqslant 2$ dan volgt uit (3) en (1) dat ook $\Delta_{j-2}(\lambda_o) = 0$. Etc. Dus $\Delta_o(\lambda_o) = 0$. Tegenspraak!

b. De nulpunten van Δ_j zijn alle reeel en enkelvoudig. Noem ze $\lambda_1^{(j)},\ldots,\lambda_j^{(j)}$, gerangschikt in toenemende grootte. Dan geldt

$$\lambda_1^{(j)} < \lambda_1^{(j-1)} < \lambda_2^{(j)} < \ldots < \lambda_{j-1}^{(j-1)} < \lambda_j^{(j)} \qquad (5)$$

(de nulpunten van Δ_j en Δ_{j-1} liggen dus om en om).

Bewijs. Voor j = 1 geldt de bewering triviaal (Δ_o heeft geen nulpunten). Stel dat de bewering geldt voor tot en met een zekere $j \geqslant 1$. We bewijzen dat hij dan ook geldt als j door j+1 vervangen wordt.

Volgens de inductieveronderstelling heeft Δ_{j-1} j-1 reële enkelvoudige nulpunten. Gaat λ van $-\infty$ naar $+\infty$ (reeel) dan wisselt $\Delta_{j-1}(\lambda)$ dus van teken in $\lambda_1^{(j-1)},\ldots,\lambda_{j-1}^{(j-1)}$ en ook nergens anders. Uit (4) en (5) volgt derhalve dat Δ_{j-1} positief is in $\lambda_1^{(j)}$, negatief in $\lambda_2^{(j)}$, etc. We schrijven dit in een formule :

$$\operatorname{sgn}\ \Delta_{j-1}(\lambda_k^{(j)}) = (-1)^{k-1}, \quad k=1,\ldots,j.$$

Met (1) en (3) volgt hieruit

$$\operatorname{sgn}\ \Delta_{j+1}(\lambda_k^{(j)}) = (-1)^{k}, \quad k=1,\ldots,j.$$

inductie bewijs

$\Delta_{j+1}(\lambda_k^j) = (\lambda_k^j - d_{j+1}/d_j)\Delta_j(\lambda_k^j) + 0$ waarom ? → $- \underbrace{b_j c_j}_{>0} \Delta_{j-1}(\lambda_k^j)$

Combineren we dit met (4) (geschreven met j+1 in plaats van j) dan zien we dat als λ van $-\infty$ naar $+\infty$ loopt, $\Delta_{j+1}(\lambda)$ minstens j+1 maal van teken wisselt, en wel minstens een maal in ieder van de intervallen $(-\infty, \lambda_1^{(j)})$, $(\lambda_1^{(j)},\lambda_2^{(j)}),\ldots,(\lambda_{j-1}^{(j)},\lambda_j^{(j)})$ en $(\lambda_j^{(j)},\infty)$. Maar Δ_{j+1} heeft (als polynoom van de graad j+1 hoogstens j+1 verschillende nulpunten. Deze moeten dus alle reeel en enkelvoudig zijn en voldoen aan

$$-\infty < \lambda_1^{(j+1)} < \lambda_1^{(j)} < \lambda_2^{(j+1)} < \ldots < \lambda_j^{(j)} < \lambda_{j+1}^{(j+1)} < \infty$$

Q.e.d.

c. Zij λ zo dat $\Delta_j(\lambda) \neq 0$, $j=1,\ldots,n$. Zij $T(\lambda)$ het aantal tekenwisselingen in de rij getallen $\Delta_0(\lambda), \Delta_1(\lambda),\ldots,\Delta_n(\lambda)$. Dan is $T(\lambda)$ gelijk aan het aantal der nulpunten van Δ_n die links van λ liggen.

Bewijs : $T(\lambda)$ is constant in ieder interval waarin geen nulpunten van Δ_0, $\Delta_1,\ldots,$ of Δ_n liggen.

Zij λ_o een nulpunt van den Δ_j met $j < n$. Dit nulpunt is enkelvoudig dus Δ_j wisselt hier van teken. Maar volgens a. is $\Delta_{j-1}(\lambda_o) \neq 0$ en $\Delta_{j+1}(\lambda_o) \neq 0$ dus Δ_{j-1} en Δ_{j+1} hebben een constant teken in een omgeving van λ_o. Hieruit volgt dat $T(\lambda)$ links en rechts van λ_o dezelfde waarde heeft.

$T(\lambda)$ kan dus alleen van waarde veranderen in de nulpunten $\lambda_k^{(n)}$ van $\Delta_n(\lambda)$. Maar uit het bewijs van b. volgt eenvoudig dat Δ_{n-1} en Δ_n links van $\lambda_k^{(n)}$ hetzelfde en rechts van $\lambda_k^{(n)}$ tegengesteld teken hebben. In ieder punt $\lambda_k^{(n)}$ neemt $T(\lambda)$ dus met 1 toe. En daar uit (4) volgt dat $T(\lambda)=0$ voor $\lambda \to -\infty$ (en $T(\lambda)=n$ voor $\lambda \to +\infty$), volgt hieruit de bewering.

Op grond van deze eigenschappen zijn de eigenwaarden van A ten opzichte van D (de nulpunten van $\Delta_n(\lambda)$), die dus alle reeel en enkelvoudig zijn, vrij eenvoudig te bepalen. Men kan bv. als volgt te werk gaan. Men bepaalt eerst met de stelling van Gershgorin *) een interval (α,β) waarbinnen alle eigenwaarden van A liggen. Zij c_o het midden van (α,β), Bepaal $\Delta_1(c_o),\ldots,\Delta_n(c_o)$ en hiermee $T(c_o)$. Dan weten we hoeveel eigenwaarden er in (α,c_o) en hoeveel er in (c_o,β) liggen. Halveer nu (α,c_o) en (c_o,β). Etc. Zo kan men doorgaan tot men n intervallen heeft die ieder precies een eigenwaarde bevatten. Zij (γ,δ) zo'n interval. Dan hebben $\Delta_n(\gamma)$ en $\Delta_n(\delta)$ verschillend teken. Met regula falsi vindt men dan betere benaderingen voor deze eigenwaarde, die in (γ,δ) ligt. Werkt men met een tafelmachine dan is het zinvol, ook een ruwe grafiek van Δ_n te tekenen en hieruit schattingen voor de eigenwaarde af te leiden.

---

*) Ga na dat (ook als $D \neq I$) de Gershgorincirkels zijn

$$|a_1 - \lambda d_1| = |b_1|$$
$$|a_j - \lambda d_j| = |b_{j-1}| + |b_j|, \quad j=2,\ldots,n-1$$
$$|a_n - \lambda d_n| = |c_{n-1}|$$

Zij λ eigenwaarde van A ten opzichte van D (dus nulpunt van Δ_n). Kunnen we dan eenvoudig de bijbehorende eigenvector vinden ? We beweren : de eigenvector heeft componenten $\xi_1, \ldots, \xi_n$, waarin

$$\left.\begin{aligned} \xi_1 &= 1 \qquad (= \Delta_0(\lambda)) \\ \xi_j &= \frac{\Delta_{j-1}(\lambda)}{b_1 b_2 \ldots b_{j-1}} \qquad j=2,\ldots,n \end{aligned}\right\} \tag{6}$$

Bewijs . De componenten van de eigenvector moeten voldoen aan

$$\begin{pmatrix} a_1-\lambda d_1 & -b_1 & & \\ -c_1 & \ddots & \ddots & \\ & \ddots & \ddots & -b_{n-1} \\ & & -c_{n-1} & a_n-\lambda d_n \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \tag{7}$$

Hieruit volgt

$$\left.\begin{aligned} b_1 \xi_2 &= (a_1 - \lambda d_1)\xi_1 \\ b_j \xi_{j+1} &= (a_j - \lambda d_j)\xi_j - c_{j-1}\xi_{j-1}, \qquad j=2,\ldots,n-1, \\ 0 &= (a_n - \lambda d_n)\xi_n - c_{n-1}\xi_{n-1} \end{aligned}\right\} \tag{8}$$

Vermenigvuldig de j-de vergelijking met $b_1 b_2 \ldots b_{j-1}$ (dus de eerste met 1, en stel $b_1 b_2 \ldots b_{j-1}\xi_j = \zeta_j$ (dus $\xi_1 = \zeta_1$). Dan gaan de vergelijkingen (8) over in

$$\begin{aligned} \zeta_2 &= (a_1 - \lambda d_1)\zeta_1 \\ \zeta_{j+1} &= (a_j - \lambda d_j)\zeta_j - b_{j-1}c_{j-1}\zeta_{j-1} \\ 0 &= (a_n - \lambda d_n)\zeta_n - b_{n-1}c_{n-1}\zeta_{n-1} \end{aligned}$$

Vergelijking met (3) leert dat, als we $\zeta_1 = \xi_1 = 1$ stellen

$$\zeta_j = \Delta_{j-1}(\lambda), \qquad j=1,\ldots,n.$$

En aan de laatste vergelijking is voldaan daar

$$(a_n - \lambda d_n)\Delta_{n-1}(\lambda) - b_{n-1}c_{n-1}\Delta_{n-2}(\lambda) = \Delta_n(\lambda) = 0.$$

We beschouwen nu een variant van de bovenstaande methode die in de technische mechanica bekend is als de methode van Holzer. Beschouw voor zekere

reële λ het volgende stelsel vergelijkingen

$$\begin{pmatrix} a_1 - \lambda d_1 & -b_1 & & \\ -b_1 & \ddots & \ddots & \\ & \ddots & \ddots & -b_{n-1} \\ & & -c_{n-1} & a_n - \lambda d_n \end{pmatrix} \begin{pmatrix} 1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \alpha \end{pmatrix} \qquad (9)$$

Dit is een - op een wat ongebruikelijke wijze opgeschreven- stelsel van n lineaire vergelijkingen met n onbekenden $(\eta_2,\ldots,\eta_n\alpha)$, dat mechanisch geïnterpreteerd kan worden. Het is (daar $b_j \neq 0$) eenvoudig op te lossen : η_2 uit de eerste ,...,η_n uit de (n-1)-ste en α uit de laatste vergelijking :

$$\eta_2 = \frac{a_1 - \lambda d_1}{b_1}$$

$$\eta_{j+1} = \frac{(a_j - \lambda d_j)\,\eta_j - c_{j-1}\eta_{j-1}}{b_j}, \quad j=2,\ldots,n-1$$

$$\alpha = (a_n - \lambda d_n)\,\eta_n - c_{n-1}\eta_{n-1}.$$

Bij iedere λ vinden we dus een $\alpha = \alpha(\lambda)$. We beweren : $\alpha(\lambda) = 0$ dan en slechts dan als λ eigenwaarde is van A ten opzichte van D.

<u>Bewijs</u> : Zij $\alpha(\lambda) = 0$. Dan volgt uit (9) dat λ eigenwaarde is met eigenvector $(1,\eta_2,\ldots,\eta_n)$.

Zij λ eigenwaarde. Dan heeft het homogene stelsel (7) een niet-triviale oplossing (de eigenvector). Deze oplossing ligt na keuze van ξ_1 eenduidig vast (ξ_2 bepalen uit eerste vergelijking, etc.) Hieruit volgt dat de eigenvector op een factor na eenduidig bepaald is en dat $\xi_1 \neq 0$ is (want anders zou ook $\xi_2 = \ldots = \xi_n = 0$ zijn). Kies $\xi_1 = 1$. Dan blijkt door vergelijking van (7) en (9) dat $\eta_2 = \xi_2,\ldots,\eta_n = \xi_n$. En dus is $\alpha = -c_{n-1}\eta_{n-1} + (a_n - \lambda d_n)\,\eta_n = -c_{n-1}\xi_{n-1} + (a_n - \lambda d_n)\,\xi_n = 0$.

Volgens Holzer bepaalt men nu $\alpha(\lambda)$ voor een aantal waarden van λ en hieruit vindt men bv. grafisch de nulpunten van $\alpha(\lambda)$, d.w.z. de eigenwaarden en de bijbehorende eigenvectoren $(1,\eta_1,\ldots,\eta_n)$.

Het verband tussen deze methode en de voorgaande is als volgt. Uit (9) volgt, net als bij het bewijs van (6), dat

$$\eta_j = \frac{\Delta_{j-1}(\lambda)}{b_1 b_2 \cdots b_{j-1}} \quad , \qquad j=2,\ldots,n.$$

En uit de laatste vergelijking volgt dan

$$\alpha(\lambda) = \frac{\Delta_n(\lambda)}{b_1 b_2 \cdots b_{n-1}}.$$

De functie $\alpha(\lambda)$ is dus op een factor na gelijk aan $\Delta_n(\lambda) = \det(A - \lambda D)$. Hieruit volgt de equivalentie van beide methoden.
In de mechanische toepassingen is meestal $b_j > 0$ $(j=1,\ldots,n-1)$. Dan heeft dus η_j hetzelfde teken als $\Delta_{j-1}(\lambda)$ en α hetzelfde teken als $\Delta_n(\lambda)$. Voor de bepaling van $T(\lambda)$ (d.i. het aantal eigenwaarden links van λ) kunnen we dus ook kijken naar het aantal tekenwisselingen in de rij $1, \eta_1, \ldots, \eta_n, \alpha$. En we vinden nog het volgende resultaat (ga na) :
Als λ_k de k-de eigenwaarde is (bij rangschikking naar opvolgende grootte) met eigenvector $(\xi_1, \ldots, \xi_n)$ dan heeft de rij $\xi_1, \ldots, \xi_n$ k-1 tekenwisselingen.

Opmerkingen

1. Er bestaan processen om een willekeurige symmetrische matrix A in eindig veel stappen te reduceren tot een symmetrische tridiagonaalmatrix. Combineert men dit met de bovenstaande methode dan ontstaat een zeer aantrekkelijke methode om alle eigenwaarden en eigenvectoren van een symmetrische matrix te bepalen.
2. Is A symmetrisch dan vindt men als volgt een kwadratisch convergent proces om een (niet te slechte) schatting voor een eigenwaarde te verbeteren.
Zij λ een benadering voor een eigenwaarde λ_k. Bepaal uit (9) de getallen $\eta_2, \ldots, \eta_n$ en α. Zou $\lambda = \lambda_k$ zijn dan is $(1, \eta_2, \ldots, \eta_n)$ de eigenvector. Is λ dicht bij λ_k (afwijking van de orde van ε) dan ligt $(1, \eta_2, \ldots, \eta_n)$ dicht bij de eigenvector (afwijking van de orde ε daar $\eta_2, \ldots, \eta_n$ continu differentieerbaar van λ afhangen).

Nu geldt algemeen : is $\underline{y}$ een benadering voor een eigenvector van A (symmetrisch) ten opzichte van D met afwijking van de orde ε dan is $\frac{(A\underline{y},\underline{y})}{(D\underline{y},\underline{y})}$ een benadering voor de bijbehorende eigenwaarde met afwijking van de orde ε^2. Dit wordt op dezelfde manier bewezen als op pag. 68 voor het geval D = I.

In ons geval is $\underline{y} = (1, \eta_2, \ldots, \eta_n)$, $(A - \lambda D)\underline{y} = (0, \ldots, 0, \alpha)$ en de nieuwe benadering voor de eigenwaarde wordt dus

$$\frac{(A\underline{y},\underline{y})}{(D\underline{y},\underline{y})} = \frac{((A - \lambda D)\underline{y},\underline{y})}{(D\underline{y},\underline{y})} + \lambda = \lambda + \frac{\alpha \eta}{d_1 + d_2 \eta_2^2 + \ldots + d_n \eta_n^2} =$$

$$= \lambda + \frac{\Delta_n(\lambda)\, \Delta_{n-1}(\lambda)}{d_1 (b_1 \ldots b_{n-1})^2 + d_2 (b_2 \ldots b_{n-1})^2 \Delta_1^2(\lambda) + \ldots + d_n \Delta_{n-1}^2(\lambda)}$$

4. Interpolatie

4.1. Interpolatie volgens Lagrange

Zij van een functie f van x $n+1$ functiewaarden $f(x_0)$, $f(x_1), \ldots, f(x_n)$ gegeven. Kunnen we dan een polynoom p vinden dat in de punten $x_0, \ldots, x_n$ dezelfde waarde heeft als f? En is een dergelijk probleem eenduidig bepaald?

Het antwoord op deze vragen hangt af van de beperkingen die we stellen aan de graad van het polynoom. Eisen we dat de graad van het polynoom hoogstens n is, dan is de oplossing - als er een is - stellig eenduidig. Want stel dat zowel p als q voldoen. Dan geldt (B.v. door n punten hoogstens één eerstegraads kromme)

$$p(x_j) = q(x_j) = f(x_j), \qquad j = 0, 1, \ldots, n.$$

Het polynoom p-q, dat hoogstens de graad n heeft, heeft dus minstens n+1 nulpunten, d.w.z. het is identiek gelijk aan nul.

Omgekeerd lukt het ook steeds om een polynoom met graad hoogstens n te vinden dat voldoet. Definieer nl. de n+1 polynomen $L_0, L_1, \ldots, L_n$ door

$$L_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \left(\frac{x - x_j}{x_k - x_j} \right), \qquad k = 0, 1, \ldots, n. \tag{1}$$

Dit zijn polynomen van de graad n (n factoren) met de eigenschap dat

$$L_k(x_j) = \begin{cases} 0 & \text{als } j \neq k \\ 1 & \text{als } j = k \end{cases}.$$

En hieruit volgt direct dat

$$p(x) = \sum_{k=0}^{n} f(x_k)\, L_k(x)$$

voldoet aan de gestelde eisen (graad $\leq n$ en $p(x_k) = f(x_k)$). Men noemt het polynoom (1) het interpolatiepolynoom van Lagrange.

Kunnen we iets zeggen over het verschil tussen $p(x)$ en $f(x)$ als $x \neq x_j$ $(j = 0,1,\ldots,n)$? Is $p(x)$ een goede benadering voor $f(x)$? In het algemeen is hier niets van te zeggen - f kan een "wilde" functie zijn. We kunnen wel iets zeggen als we iets weten over de afgeleiden van f :

Stelling. Zij f continu in een gesloten interval $[a,b]$ dat de punten $x_0,\ldots,x_n$ bevat. Zij f in het open interval (a,b) n+1 maal differentieerbaar.
Dan is er bij iedere x uit $[a,b]$ een getal ξ, dat voldoet aan

$$\min[x_0,x_1,\ldots,x_n x] < \xi < \max[x_0,x_1,\ldots,x_n x]$$

zodanig dat

$$f(x) = \sum_{k=0}^{n} f(x_k)\, L_k(x) + (x - x_0)\ldots(x - x_n)\, \frac{f^{(n+1)}(\xi)}{(n+1)!} \qquad (3)$$

Bewijs. Het bewijs berust op het theorema van Rolle :
Zij $F(t)$ continu in een gesloten interval $[t_1,t_2]$ en differentieerbaar in het open interval (t_1,t_2). Dan is er een punt τ in (t_1,t_2) zodanig dat $F'(\tau) = 0$.

Definieer voor $x \in [a,b]$, $x \neq x_j$ de functie $S(x)$ door

$$f(x) = \sum_{k=0}^{n} f(x_k)\, L_k(x) + (x - x_0)\ldots(x - x_n)\, S(x).$$

Kies een $x \neq x_j$ uit $[a,b]$ en beschouw de functie

$$F(t) = f(t) - \sum_{k=0}^{n} f(x_k)\, L_k(t) - (t - x_0)\ldots(t - x_n)\, S(x)$$

in het interval $\alpha \leq t \leq \beta$, waarin

$$\alpha = \min[x_0, x_1, \ldots, x_n, x], \quad \beta = \max[x_0, x_1, \ldots, x_n, x].$$

Dan is $F(t) = 0$ in de $n+2$ verschillende punten $x_0, x_1, \ldots, x_n$ en x. Daar F continu is in $[\alpha, \beta]$ en differentieerbaar in (α, β) is, volgt hieruit met Rolle dat F' minstens $n+1$ verschillende nulpunten heeft in (α, β) (verschillend en in het open interval (α, β) omdat het punt τ van Rolle in het open interval ligt).

Daar F' differentieerbaar en dus continu is in (α, β) volgt hieruit met Rolle dat F'' minstens n verschillende nulpunten in (α, β) heeft. Etc. Zo vinden we dat $F^{(n+1)}$ minstens een nulpunt $t = \xi$ heeft in (α, β). Maar (ga na)

$$F^{(n+1)}(\xi) = f^{(n+1)}(\xi) - (n+1)!\, S(x).$$

Hieruit volgt de bewering voor het geval dat $x \neq x_j$.
Het andere geval is triviaal.

Opmerkingen.

1. In het algemeen hebben we niet zoveel aan het resultaat van deze stelling daar er meestal niet veel over de afgeleiden van f bekend is.
2. Formule (3) heet de interpolatie formule van Lagrange.
3. Formules (1) en (2) zijn niet zo erg geschikt om het interpolatie polynoom (2) uit te rekenen. Men kan beter werken met zgn. gedeelde differenties (interpolatie formule van Newton). We gaan daar hier niet op in.
4. Voor het geval dat de interpolatiepunten $x_0, \ldots, x_n$ equidistant zijn (d.w.z. $x_k = x_0 + kh$) bestaan er tabellen van de polynomen $L_k(x)$.
5. In het geval dat $n = 1$ is $L_0(x) = \dfrac{x - x_1}{x_0 - x_1}$, $L_1(x) = \dfrac{x - x_0}{x_1 - x_0}$ en formule (2) wordt

$$p(x) = \frac{(x_1 - x)\, f(x_0) + (x - x_0)\, f(x_1)}{x_1 - x_0} =$$

$$= f(x_0) + \frac{x - x_0}{x_1 - x_0} \cdot [\, f(x_1) - f(x_0)\,] =$$

$$= f(x_1) + \frac{x_1 - x}{x_1 - x_0} \cdot [\, f(x_0) - f(x_1)\,]$$

(lineaire interpolatie).

4.2. Interpolatie bij gelijke intervallen met behulp van differenties

4.1.1. Voorwaartse, achterwaartse en centrale differenties

Zij h een vast gekozen positief getal. We definiëren de voorwaartse differenties Δf, $\Delta^2 f$,... van een functie f door

$$\begin{aligned}\Delta f(x) &= f(x+h) - f(x)\\ \Delta^2 f(x) &= \Delta(\Delta f)(x) = \Delta f(x+h) - \Delta f(x)\\ &= f(x+2h) - 2f(x+h) + f(x)\\ \Delta^{k+1} f(x) &= \Delta(\Delta^k f)(x) = \Delta^k f(x+h) - \Delta^k f(x).\end{aligned}$$

Meestal kiezen we een vast punt x_o en stellen $x_k = x_o + kh$ (k geheel). We schrijven dan f_o in plaats van $f(x_o)$, f_k in plaats van $f(x_k)$. Voorts Δf_j in plaats van $\Delta f(x_j)$ etc. We hebben dan

$$\Delta f_o = f_1 - f_o, \quad \Delta f_j = f_{j+1} - f_j$$
$$\Delta^{k+1} f_j = \Delta^k f_{j+1} - \Delta^k f_j, \text{ etc.}$$

De differenties worden gerangschikt in een tabel :

x_o	f_o			
		Δf_o		
x_1	f_1		$\Delta^2 f_o$	
		Δf_1		$\Delta^3 f_o$
x_2	f_2		$\Delta^2 f_1$	
		Δf_2		....
x_3	f_3		....	
		....		
....	....			

Zoals bekend worden de binomiaal coëfficiënten $\binom{s}{j}$ gedefinieerd door

$$\binom{s}{j} = \frac{s(s-1)\ldots(s-j+1)}{j!} \qquad (1)$$

Hierin is j een niet-negatief geheel getal en s een willekeurig getal (0! = 1 en als j = 0 dan is de teller van (1) per definitie gelijk aan 1, nl. nul factoren!). Is j geheel dan is (1) equivalent met

$$\binom{s}{j} = \begin{cases} 0 & \text{als } j > s \\ \dfrac{s!}{j!(s-j)!} & \text{als } 0 \leq j \leq s. \end{cases}$$

Voorts is het prettig om $\binom{s}{j} = 0$ te stellen voor $j < 0$.
Een belangrijke eigenschap van binomiaalcoëfficiënten (geldig voor willekeurige s en gehele j) is

$$\binom{s}{j} = \binom{s-1}{j} + \binom{s-1}{j-1}.$$

Uit deze eigenschap volgen door volledige inductie de formules (ga na)

$$\Delta^k f_i = \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} f_{i+j} \qquad (2)$$

$$\sum_{j=0}^{k} \binom{k}{j} \Delta^j f_i = f_{i+k}. \qquad (3)$$

<u>Opmerking</u>. Deze en dergelijke formules leidt men handig af met behulp van de volgende operatorenmethode.
Definieer de verplaatsings operator E door

$$E\, f(x) = f(x + h), \quad \text{of} \quad E\, f_j = f_{j+1}$$

(de grafiek van Ef ontstaat uit die van f door de laatste over de afstand h naar links te schuiven of door de x-as met z'n coördinaten over h naar rechts te schuiven).
Analoog $E^2 f = E(Ef)$, dus $E^2 f(x) = f(x + 2h)$, of $E^2 f_j = f_{j+2}$. Etc.
Dan is $\Delta f = (E - 1)f$, of $\Delta f_j = (E - 1)f_j$.
Derhalve is

$$\Delta^k f_i = (E - 1)^k f_i = \sum_{j=0}^{k} \binom{k}{j} E^j (-1)^{k-j} f_i =$$

$$= \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} f_{i+j}.$$

En $\sum_{j=0}^{k} \binom{k}{j} \Delta^j f_i = (1 + \Delta)^k f_i = E^k f_i = f_{i+k}$.

Hierbij is gebruik gemaakt van de binomium formule

$$(a + b)^k = \sum_{j=0}^{k} \binom{k}{j} a^j b^{k-j}$$

die, zoals men eenvoudig inziet, ook voor de operatoren E,Δ, etc. gebruikt mag worden.

Geheel analoog definieert men achterwaartse differenties :

$$\nabla f(x) = f(x) - f(x - h)$$

$$\nabla^{k+1} f(x) = \nabla(\nabla^k f)(x) = \nabla^k f(x) - \nabla^k f(x - h).$$

Dus ook $\nabla f_j = f_j - f_{j-1}$ etc.

De tabel wordt

$$\begin{array}{llllll}
\ldots. & \ldots. & & & & \\
 & & \ldots. & & & \\
x_{-3} & f_{-3} & & \ldots. & & \\
 & & \nabla f_{-2} & & \ldots. & \\
x_{-2} & f_{-2} & & \nabla^2 f_{-1} & & \ldots. \\
 & & \nabla f_{-1} & & \nabla^3 f_0 & \\
x_{-1} & f_{-1} & & \nabla^2 f_0 & & \\
 & & \nabla f_0 & & & \\
x_0 & f_0 & & & &
\end{array}$$

Merk op dat de getallen in de tabel precies dezelfde zijn als bij voorwaartse differenties, alleen de benaming is anders. We hebben

$$\nabla f_j = \Delta f_{j-1}, \quad \nabla^k f_j = \Delta^k f_{j-k}.$$

Hieruit (of rechtstreeks) volgt

$$\nabla^k f_i = \sum_{j=0}^{k} \binom{k}{j} (-1)^j f_{i-j}$$

$$\sum_{j=0}^{k} \binom{k}{j} (-1)^j \nabla^j f_i = f_{i-k}.$$

<u>Opmerking</u>. Definiëren we E^{-1} door $E^{-1} f(x) = f(x - h)$ of $E^{-1} f_j = f_{j-1}$ (zodat inderdaad $E^{-1}.E = E.E^{-1} = 1$) dan is $\nabla = 1 - E^{-1} = E^{-1}.\Delta = \Delta.E^{-1}$.

Tenslotte hebben we centrale differenties :

$$\delta f(x) = f(x + \tfrac{1}{2}h) - f(x - \tfrac{1}{2}h)$$

$$\delta^{k+1} f(x) = \delta^k f(x + \tfrac{1}{2}h) - \delta^k f(x - \tfrac{1}{2}h)$$

of $\delta f_j = f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}$, etc.

Meestal berekent men niet δf_0, δf_1, ..., doch $\delta f_{\frac{1}{2}}$ $(= f_1 - f_0)$, $\delta f_{\frac{3}{2}}$,... vervolgens $\delta^2 f_0$, $\delta^2 f_1$,..., dan $\delta^3 f_{\frac{1}{2}}$, $\delta^3 f_{\frac{3}{2}}$,... etc. (waarom?).
De tabel wordt dan

....	....				
		....			
x_{-2}	f_{-2}		....		
		$\delta f_{-\frac{3}{2}}$		....	
x_{-1}	f_{-1}		$\delta^2 f_{-1}$		....
		$\delta f_{-\frac{1}{2}}$		$\delta^3 f_{-\frac{1}{2}}$	
x_0	f_0		$\delta^2 f_0$		$\delta^4 f_0$ ....
		$\delta f_{\frac{1}{2}}$		$\delta^3 f_{\frac{1}{2}}$	
x_1	f_1		$\delta^2 f_1$		....
		$\delta f_{\frac{3}{2}}$		....	
x_2	f_2		....		
		....			
....	....				

Merk op : weer dezelfde getallen als bij de vorige tabellen doch met andere namen. We hebben

$$\delta f_{j+\frac{1}{2}} = \Delta f_j = \nabla f_{j-1}$$

$$\delta^{2k} f_i = \Delta^{2k} f_{i-k} = \nabla^{2k} f_{i+k} = \sum_{j=0}^{2k} \binom{2k}{j} (-1)^j f_{i-k+j}$$

$$\delta^{2k+1} f_{i+\frac{1}{2}} = \Delta^{2k+1} f_{i-k} = \nabla^{2k+1} f_{i+k+1} = \sum_{j=0}^{2k+1} \binom{2k+1}{j} (-1)^{j+1} f_{i-k+j}.$$

<u>Opmerking</u>. Definieert men $E^{\frac{1}{2}}$ door $E^{\frac{1}{2}} f(x) = f(x + \frac{1}{2}h)$ of $E^{\frac{1}{2}} f_j = f_{j+\frac{1}{2}}$ (zodat $E^{\frac{1}{2}} . E^{\frac{1}{2}} = E$) en analoog $E^{-\frac{1}{2}}$, dan is

$$\delta = E^{\frac{1}{2}} - E^{-\frac{1}{2}}.$$

Een operator die in verband met centrale differenties vaak handig is, is de operator μ, gedefinieerd door

$$\mu f(x) = \tfrac{1}{2} [f(x + \tfrac{1}{2}h) + f(x - \tfrac{1}{2}h)].$$

$$\mu f_j = \tfrac{1}{2} [f_{j+\frac{1}{2}} + f_{j-\frac{1}{2}}].$$

Dus $$\mu = \tfrac{1}{2} [E^{\frac{1}{2}} - E^{-\frac{1}{2}}]$$

Met deze operator kan men half-tallige indices vaak wegwerken. Bv.

$$\mu \delta f_0 = \tfrac{1}{2} [\delta f_{\frac{1}{2}} + \delta f_{-\frac{1}{2}}] \quad (= \tfrac{1}{2}(f_1 - f_{-1}))$$

4.2.2. Interpolatieformules van Gregory-Newton

Zij van een functie f de waarden

$$f_o = f(x_o),\ f_1 = f(x_1), \ldots, f_n = f(x_n)$$

gegeven (waarin weer $x_j = x_o + jh$).
Definieer het polynoom p door

$$p(x) = f_o + \frac{x-x_o}{1!h}\Delta f_o + \frac{(x-x_o)(x-x_1)}{2!h^2}\Delta^2 f_o - \frac{(x-x_o)\ldots(x-x_{n-1})}{n!h^n}\Delta^n f_o.$$

Een handiger notatie krijgen we door $x = x_o + sh$ te stellen

$$p_s = p(x_o+sh) = f_o + \frac{s}{1!}\Delta f_o + \frac{s(s-1)}{2!}\Delta^2 f_o + \ldots + \frac{s\ldots(s-n+1)}{n!}\Delta^n f_o$$

$$= f_o + \binom{s}{1}\Delta f_o + \binom{s}{2}\Delta^2 f_o + \ldots + \binom{s}{n}\Delta^n f_o.$$

Dan is voor $k=0,1,\ldots,n$

$$p_k = p(x_o + kh) \quad \text{gelijk aan } f_k,$$

d.w.z. p (een polynoom met graad $\leq n$) is het (volgens 4.1 eenduidig bepaalde) interpolatiepolynoom door de punten $(x_o,f_o),\ldots,(x_n,f_n)$.
Want volgens 4.2.1. (formule (3)) is

$$p_k = \sum_{j=o}^{n}\binom{k}{j}\Delta^j f_o = \sum_{j=o}^{k}\binom{k}{j}\Delta^j f_o = f_k,$$

daar $\binom{k}{j} = 0$ als $j > k$ (k is geheel).

Dit is een eenvoudige methode om (bij aequidistante abscissen) het interpolatiepolynoom te bepalen!
Een geheel analoge formule is

$$p_{-s} = p(x_o - sh) = f_o - \binom{s}{1}\nabla f_o + \ldots + (-1)^n\binom{s}{n}\nabla^n f_o$$

voor het interpolatiepolynoom door de punten $(x_o,f_o),\ldots(x_{-n},f_{-n})$.

Schrijven we deze formule in de vorm

$$f_s = f_o + \binom{s}{1} \Delta f_o + \ldots + \binom{s}{n} \Delta^n f_o + R_s^{(n)} \qquad (1)$$

$$f_{-s} = f_o - \binom{s}{1} \nabla f_o + \ldots + (-1)^n \binom{s}{n} \nabla^n f_o + R_{-s}^{(n)} \qquad (2)$$

dan heten zij voorwaartse, resp. achterwaartse interpolatieformules van Newton-Gregory.

Kunnen we nu iets over de restterm zeggen? Zij f n+1 maal differentieerbaar. Pas op de functie $f(x_o + sh)$, beschouwd als functie van s en met interpolatiepunten $s_o = 0$, $s_1 = 1, \ldots, s_n = n$, de stelling uit 4.1 toe. Dan vinden we, daar

$$\frac{d^{n+1}}{ds^{n+1}} f(x_o + sh) = h^{n+1} f^{(n+1)}(x_o + sh),$$

voor de voorwaartse formule (1)

$$R_s^{(n)} = \frac{s(s-1)\ldots(s-n)}{(n+1)!} h^{n+1} f^{(n+1)}(x_o + \sigma h), \qquad (3)$$

waarin $\min[0,s] < \sigma < \max[n,s]$.

Analoog vinden we voor de achterwaartse formule (2)

$$R_{-s}^{(n)} = (-1)^{n+1} \frac{s(s-1)\ldots(s-n)}{(n+1)!} h^{n+1} f^{(n+1)}(x_o + \sigma h)$$

met $\min[-n,-s] < \sigma < \max[0,-s]$.

<u>Opmerking</u>. Uit de formules(1) en (3) leiden we nog het volgende resultaat af. Neem in (1) $s = n + 1$. Dan ontstaat

$$f_{n+1} = f_o + \binom{n+1}{1} \Delta f_o + \ldots + \binom{n+1}{n} \Delta^n f_o + h^{n+1} f^{(n+1)}(x_o + \sigma h),$$

met $0 < \sigma < n+1$.

Met formule (3) uit 4.2.1 volgt hieruit

$$\Delta^{n+1} f_o = h^{n+1} f^{(n+1)}(x_o + \sigma h) \qquad (4)$$

Hieruit volgt weer : is f een polynoom van de n-graad dan zijn alle differenties hoger dan de n-de nul (en de n-de differentie is constant, gelijk aan $n!h^n a_o$, als a_o de coëfficiënt van x^n is).

Verder is formule (4) handig om de restterm (3) te schatten : varieert $f^{(n+1)}(x)$ slechts weinig in het interval (x_0, x_{n+1}) dan is, als $0 < s < n+1$, $\frac{s(s-1)\ldots(s-n)}{(n+1)!}\,\Delta^{n+1}f_0$ een redelijke benadering voor $R_s^{(n)}$. Merk op dat dit juist de eerste weggelaten term is!

4.2.3. Interpolatieformules met centrale differenties

De voorwaartse interpolatieformule van Newton is geschikt voor interpolatie in het begin van een tabel, de achterwaartse voor het eind van een tabel. Voor interpolatie midden in een tabel gebruiken we liever interpolatieformules met centrale differenties.
Interpoleren we bv. in de buurt van x_0 (bv. tussen $x_0-\frac{1}{2}h$ en $x_0+\frac{1}{2}h$) met behulp van een tweede graadspolynoom (drie basispunten) dan ligt het voor de hand om als basispunten x_{-1}, x_0 en x_1 te kiezen. Gebruiken we de voorwaartse formule van Newton, dan krijgen we

$$f_s = f_{-1+(s+1)} = f_{-1} + \binom{s+1}{1}\Delta f_{-1} + \binom{s+1}{2}\Delta^2 f_{-1} + \frac{(s+1)s(s-1)}{3!}h^3 f^{(3)}(\xi)$$

met (als $|s| \leq 1$) $\quad x_{-1} < \xi < x_1$.
We kunnen dit reorganiseren tot de meer symmetrische formule

$$f_s = f_0 + \tfrac{1}{2}s[\,\delta f_{\frac{1}{2}} + \delta f_{-\frac{1}{2}}\,] + \tfrac{1}{2}s^2\delta^2 f_0 - \frac{s(1-s^2)}{6}h^3 f^{(3)}(\xi). \qquad (1)$$

Dit is de interpolatieformule van Stirling (afgebroken na de tweede differentie).

Willen we tussen x_0 en x_1 interpoleren met een derde graads polynoom (vier basispunten) dan kiezen we als basispunten x_{-1}, x_0, x_1 en x_2. We kunnen weer uitgaan van de voorwaartse formule van Newton met beginpunt x_{-1} en afbreken na de derde differenties.
Na reorganisatie vinden we

$$f_s = (1-s)f_0 + s\,f_1 + \binom{2-s}{3}\delta^2 f_0 + \binom{1+s}{3}\delta^2 f_1 +$$

$$+ \frac{s(1-s)(1+s)(2-s)}{4!}h^4 f^{(4)}(\xi) \qquad (2)$$

waarin (als $-1 < s < 2$) $x_{-1} < \xi < x_2$.

Dit is de interpolatieformule van Everett (afgebroken na de tweede differenties).
Voor $s = \frac{1}{2}$ vinden we met name

$$f_{\frac{1}{2}} = \tfrac{1}{2}[f_0 + f_1] - \frac{1}{16}[\delta^2 f_0 + \delta^2 f_1] + \frac{3h^4}{128} f^{(4)}(\xi), \qquad (3)$$

een prettige formule voor halvering van het interval.

Opmerking. In sommige tabellen vindt men naast de functiewaarden de tweede differenties vermeld. In andere vindt men de zg. gemodificeerde tweede differenties (met $\bar{\delta}^2$ aangegeven). Dit berust hierop. Nemen we in de formule van Everett ook de vierde differenties mee dan krijgen we

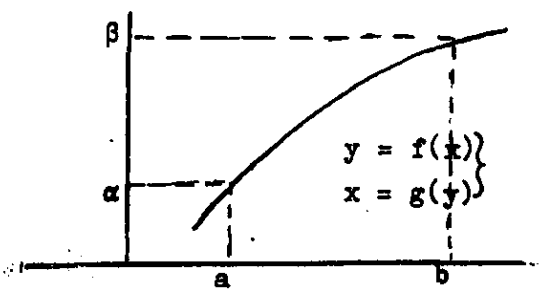
$$f_s = (1-s) f_0 + s f_1 - \frac{s(1-s)(2-s)}{6}\left[\delta^2 f_0 - \frac{(1+s)(3-s)}{20}\delta^4 f_0\right] +$$

$$- \frac{s(1-s)(1+s)}{6}\left[\delta^2 f_1 - \frac{(2+s)(2-s)}{20}\delta^4 f_1\right] + R_5^{(6)}. \qquad (4)$$

Varieert s tussen 0 en 1 dan variëren de factoren $\frac{(1+s)(3-s)}{20}$ en $\frac{(2+s)(2-s)}{20}$ beide tussen $\frac{3}{20}$ en $\frac{4}{20}$, dus betrekkelijk weinig. Men maakt dus geen erge fout (of beter gezegd - de vierde differenties worden bijna correct in rekening gebracht) indien de factoren tussen vierkante haken vervangen worden door $\bar{\delta}^2 f_0 = \delta^2 f_0 - c\delta^4 f_0$, resp. $\bar{\delta}^2 f_1 = \delta^2 f_1 - c\delta^4 f_1$, waarin c een geschikte constante is. Op historische gronden kiest men $c = 0.184$ (het gemiddelde van de genoemde factoren is $\frac{11}{60} = 0.18333..$). In plaats van met (4) werkt men dus met (2) (en een andere restterm) waarbij echter de tweede differenties vervangen zijn door gemodificeerde tweede differenties. Men noemt dit proces dat door Comrie geïntroduceerd is "throwing back the forth difference on the second".

4.3. Inverse interpolatie

Zij f in een interval (a,b) monotoon stijgend, $f(a) = \alpha$, $f(b) = \beta$. Dan is er in het interval (α,β) een functie g, zodanig dat $g(f(x)) = x$ voor $x \in (a,b)$ en $f(g(y)) = y$ voor $y \in (\alpha,\beta)$, de zgn inverse functie.



Zij f equidistant getabelleerd - met basispunten $x_0 = a, x_1, \ldots, x_N = b$. Noem $f(x_0) = y_0 (= \alpha)$, $f(x_1) = y_1, \ldots$ $f(x_N) = y_N (= \beta)$.

Dan is $g(y_o) = x_o$, $g(y_1) = x_1, \ldots$. Met de tabel van f hebben we dus ook een tabel van g. Maar de laatste heeft in het algemeen geen equidistante abscissen.

Hoe kunnen we bij gegeven η uit (α,β) $\xi = g(\eta)$ het best bepalen ?

We kunnen (dank zij de monotonie) steeds twee punten, stel x_k en x_{k+1} aangeven zodanig dat $y_k = f(x_k) \leq \eta < f(x_{k+1}) = y_{k+1}$.

Er zijn nu diverse mogelijkheden om verder te gaan (als $\eta \neq y_k$).

a. Leg een interpolatiepolynoom p (als functie van x - dus een benadering voor f) door een aantal punten (x_{k+j}, y_{k+j}) rond (x_k,y_k). De abscissen zijn dan equidistant. We moeten dan de vergelijking $p(x) = \eta$ oplossen - een hogeregraads vergelijking.

b. Leg een interpolatiepolynoom q (als functie van y, dus een benadering voor g) door een aantal punten (y_{k+j}, x_{k+j}). We moeten dan $q(\eta)$ berekenen. Men noemt dit inverse interpolatie. De abscissen zijn nu niet equidistant. Bovendien is, als $f'(\xi)$ klein is ten opzichte van 1, $g'(\xi)$ groot ten opzichte van 1 en de benadering van g door een polynoom is dan meestal slecht.

Leg een eerstegraads polynoom $x = q_1(y)$ door (y_k,x_k) en (y_{k+1}, x_{k+1}). Bepaal $\xi_1 = q_1(\eta)$ (dit is dus lineaire inverse interpolatie). Bepaal met behulp van een interpolatiepolynoom $y = p(x)$ (rechtstreekse interpolatie van hogere orde) een nauwkeurige benadering η_1 voor $f(\xi_1)$. Leg nu een eerstegraads polynoom $x = q_2(y)$ hetzij door (y_k,x_k) en (η_1,ξ_1), hetzij door (η_1,ξ_1) en (y_{k+1},x_{k+1}). Bepaal $\xi_2 = q_2(\xi\eta)$. Bepaal $\eta_2 = p(\xi_2)$. Etc.

Dit is veelal het meest aan te bevelen proces. Ga na dat het neerkomt op het oplossen van de vergelijking $p(x) = \eta$ met behulp van regula falsi !

4.4. Algemenere approximatiemethoden.

Bij interpolatie bepaalt men een n-de graads polynoom p als benadering voor een functie f met de eis dat in n+1 voorgeschreven punten $x_o \ldots x_n$ geldt $p(x_j) = f(x_j)$, $j=0,\ldots,n$. In de tussen gelegen punten eist men niets (kan men ook niets meer eisen). Dit is ehigzins vergelijkbaar met de reeks van Taylor.

Is nl. f voldoende vaak differentieerbaar in een interval $x_o-p < x < x_o+p$ dan is het n-de graads polynoom

$$p(x) = f(x_o) + (x - x_o)\, f'(x_o) + \ldots + \frac{x-x_o)^n}{n!}\, f^{(n)}(x_o)$$

een benadering voor f in dit interval welke eenduidig bepaald is door de eis dat

$$\frac{d^j p}{dx^j}(x_o) = \frac{d^j f}{dx^j}(x_o), \qquad j=0,\ldots,n.$$

De restterm is hier te schrijven als $\frac{x-x_o)^{n+1}}{(n+1)!}\, f^{(n+1)}(\xi)$

met $x_o - \rho < \xi < x_o + \rho$; (vergelijk dit met de restterm bij interpolatie). In de andere punten van het interval wordt niets geëist.

Een andere methode van approximeren is de methode der kleinste kwadraten.
a. Discreet. Men kent f in een vrij groot aantal punten $x_0,\ldots,x_N$ (bv. door meting) en bepaalt de coëfficiënten van een n-de graadspolynoom p (met $n < N$) zodanig dat

$$\sum_{j=o}^{N} [f(x_j) - p(x_j)]^2$$

zo klein mogelijk is. Deze eis leidt tot n+1 lineaire vergelijkingen waaruit de n+1 coëfficiënten van p bepaald kunnen worden.
b. Continu. Men kent f in een interval [a,b] en bepaalt het n-de graadspolynoom p door de eis dat

$$\int_a^b [f(x) - p(x)]^2 dx$$

zo klein mogelijk is. Ook dit leidt tot n+1 lineaire vergelijkingen voor de coëfficiënten van p.
Men kan deze eis ook als volgt formuleren.
De continue functies f op het interval [a,b] vormen een lineaire ruimte (optellen, vermenigvuldigen met een getal gaat goed).
In deze ruimte definiëren we als norm

$$\|f\|_2 = \left[\int_a^b f^2(x)\, dx\right]^{\frac{1}{2}}.$$

Men kan bewijzen dat aan de eisen voor een norm (vgl.2.3.1.1) voldaan is. We bepalen nu het polynoom p zo dat $\|f - p\|_2$ (de "afstand" van f en p) zo klein mogelijk is.

Het bij de metnode der kleinste kwadraten gehanteerde afstands begrip voor functies leidt tot eenvoudige vergelijkingen voor de coëfficiënten van p. Het heeft echter ook nadelen. Het klein zijn van $\|f - p\|_2$ garandeert namelijk niet dat voor alle x uit [a,b] $|f(x) - p(x)|$ ook klein is. Een met het oog op numerieke approximatie meer bevredigend afstandsbegrip is gebaseerd op een andere norm voor de ruimte der lineaire functies :

$$\|f\|_\infty = \max_{a \leq x \leq b} |f(x)|$$

Dan immers impliceert $\|f - p\|_\infty = \varepsilon$ dat $|f(x) - p(x)| \leq \varepsilon$ voor alle x uit [a,b]*).

Dat approximatie in deze zin steeds mogelijk is volgt uit de volgende

Stelling van Weierstrasz :

Zij f continu in [a,b]. Dan is er bij iedere $\varepsilon > 0$ een polynoom p zodanig dat

$$\|f - p\|_\infty < \varepsilon$$

Het polynoom p hangt natuurlijk van ε af. (ook de graad - bij kleinere ε zal de graad hoger worden).

Men kan nu trachten bij een gegeven f een polynoom p met voorgeschreven graad n te vinden zodanig dat $\|f - p\|_\infty$ zo klein mogelijk is.

Approximaties in deze zin noemt men Chebyshev-approximaties. Men kan bewijzen dat er steeds een eenduidig bepaald polynoom p* is dat aan deze eis voldoet. Dit polynoom heeft bovendien de volgende eigenschap : er zijn in [a,b] n+2 punten $\xi_0 < \xi_1 < \ldots < \xi_{n+1}$ zodanig dat $f(x) - p^*(x)$ in deze punten afwisselende tekens en de absolute waarde $\|f - p^*\|$ heeft.

Helaas is het slechts zelden mogelijk om bij een gegeven functie f het optimale polynoom expliciet te bepalen. Wel kan men (met vrij veel moeite) goede benaderingen voor p* vinden.

---

*) Op dit afstandsbegrip is ook het begrip uniforme convergentie van een reeks van functies gebaseerd !

Voorbeeld *)

$$f(x) = \log(1 + x), \qquad 0 \leqslant x \leqslant 1.$$

De Chebyshev-approximatie met behulp van een vierde-graads polynoom is

$$p(x) = 0.997444x - 0.471284x^2 + 0.225668x^3 - 0.058753x^4.$$

De maximale fout is 0.000072 (vergelijk dit met de eerste vier termen van de machtreeks $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots$; de afbreekfout is hier 0.11).

Bij Chebyshev-approximaties met graad 5,6,7, resp. 8 vindt men als maximale fout resp. 0.000010, 0.0000014, 0.00000021, 0.000000032.

Het is duidelijk dat dergelijke approximaties erg nuttig zijn bij het werken met automatische rekenmachines (waarbij het gebruik van grote tabellen in het algemeen lastig is).

Men kan op deze wijze nl. ingewikkelde functies die in bepaalde berekeningen voor vele verschillende waarden van x nodig zijn, vervangen door een approximatie die in het gehele interval een nauwkeurigheid heeft, die correspondeert met het aantal cijfers waarmee men werkt (eventueel verschillende approximaties in verschillende intervallen).

Tot nu toe werd steeds gesproken over de approximaties met behulp van polynomen. Natuurlijk kan men ook met andere functies approximeren. Van belang zijn met name trigonometrische functies (Fourier reeksen). Ook gebroken rationale functies van de vorm

$$g(x) = \frac{a_o x^k + \ldots + a_k}{b_o x^m + \ldots + b_m}$$

zijn vaak geschikt (vooral bij Chebyshev-approximaties).

---

*) Uit Hastings, Approximations for digital computers, Princeton N.J., 1955. Hierin vindt men ook vele andere voorbeelden.

5. Numerieke integratie

5.1. Inleiding

Zoals bekend wordt een bepaalde integraal gedefinieerd door een limietproces. In een beperkt aantal gevallen kan deze limiet langs analytische weg bepaald worden (bv. omdat men een primitieve van de integrand kan vinden). In de andere gevallen is het de taak van de numerieke wiskunde, een benadering voor de integraal te vinden die in een eindig aantal stappen verkregen kan worden.

Vrijwel alle numerieke integratiemethoden geven benaderingen van de volgende vorm

$$\int_a^b f(x)\,dx = c_0 f(x_0) + \ldots + c_N f(x_N) + R_N$$

Hierin is (a,b) een gegeven integratie interval, $x_0,\ldots,x_N$ zijn gekozen punten (vgl. de basispunten bij interpolatie). De coëfficiënten $c_0,\ldots,c_N$ hangen af van a,b en de gekozen punten $x_0,\ldots,x_N$, doch niet van de te integreren functie f(x). Men bepaalt $c_0,\ldots,c_N$ zodanig dat voor "fatsoenlijke" functies R_N "zo klein mogelijk" is. Deze eis is rijkelijk vaag. Er bestaan dan ook zeer vele integratieformules die ieder hun eigen voordelen en nadelen hebben.

We beginnen met een uiterst simpele integratie formule :

$$\int_a^b f(x)\,dx = \frac{b-a}{N} \sum_{k=1}^{N} f\left(a + (k - \tfrac{1}{2}) \frac{b-a}{N}\right) + R_N, \qquad (1)$$

of, als we $\frac{b-a}{N} = h$, $a = x_0$, $a + h = x_1$, ... stellen

$$\int_{x_0}^{x_0+Nh} f(x)\,dx = h\left[f_{\frac{1}{2}} + f_{\frac{3}{2}} + \ldots + f_{N-\frac{1}{2}}\right] + R_N$$

Kunnen we iets over R_N zeggen ?

Men kan bewijzen dat voor alle continue functies $\lim_{N\to\infty} R_N = 0$ (in wezen

djimmetschietleeg
drekschietleeg.

is dit een van de manieren waarop men de bepaalde integraal van een continue functie kan definiëren). Maar meer valt er zonder nadere veronderstellingen omtrent f niet te zeggen.
Veronderstel nu dat f'' bestaat en continu is in [a,b] . Dan kunnen we schrijven (partiële integratie)

$$\int_{x_j}^{x_{j+1}} f(x)\,dx - h\,f_{j+\frac{1}{2}} =$$

$$= \int_{x_j}^{x_{j+\frac{1}{2}}} f(x)\,d(x - x_j) + \int_{x_{j+\frac{1}{2}}}^{x_{j+1}} f(x)\,d(x - x_{j+1}) - h\,f_{j+\frac{1}{2}} =$$

$$= -\int_{x_j}^{x_{j+\frac{1}{2}}} (x - x_j)\,f'(x)\,dx - \int_{x_{j+\frac{1}{2}}}^{x_{j+1}} (x - x_{j+1})\,f'(x)\,dx =$$

$$= \tfrac{1}{2}\int_{x_j}^{x_{j+\frac{1}{2}}} (x - x_j)^2 f''(x)\,dx + \tfrac{1}{2}\int_{x_{j+\frac{1}{2}}}^{x_{j+1}} (x - x_{j+1})^2 f''(x)\,dx. \qquad (2)$$

Zij $m_{j+\frac{1}{2}} = \underset{x_j \leq x \leq x_{j+1}}{\text{Min}} f''(x)$; $M_{j+\frac{1}{2}} = \underset{x_j \leq x \leq x_{j+1}}{\text{Max}} f''(x)$.
Dan volgt uit (2), daar $\frac{1}{2}(x - x_j)^2 \geq 0$ en $\frac{1}{2}(x - x_{j+1})^2 \geq 0$

$$\int_{x_j}^{x_{j+1}} f(x)\,dx - h f_{j+\frac{1}{2}} \leq \tfrac{1}{2}M_{j+\frac{1}{2}}\left[\int_{x_j}^{x_{j+\frac{1}{2}}} (x - x_j)^2\,dx + \int_{x_{j+\frac{1}{2}}}^{x_{j+1}} (x - x_{j+1})^2\,dx\right]$$

$$= \frac{1}{24}\,h^3 M_{j+\frac{1}{2}}.$$

Analoog : $\int_{x_j}^{x_{j+1}} f(x)\,dx - h\,f_{j+\frac{1}{2}} \geq \frac{1}{24}\,h^3\,m_{j+\frac{1}{2}}.$

Daar f'' continu is, is er bij ieder getal η dat voldoet aan $m_{j+\frac{1}{2}} \leq \eta \leq M_{j+\frac{1}{2}}$ een ξ uit $[x_j, x_{j+1}]$ zodanig dat $f''(\xi) = \eta$.

$$\frac{1}{24}\,h^3 m_{j+\frac{1}{2}} \leq \int \leq \frac{1}{24}\,h^3 M_{j+\frac{1}{2}}$$

We vinden dus :
er is een ξ in $[x_j, x_{j+1}]$ zodanig dat

$$\int_{x_j}^{x_{j+1}} f(x)\,dx = h\, f_{j+\frac{1}{2}} + \frac{1}{24} h^3\, f''(\xi). \qquad (3)$$

En geheel analoog geldt :
er is een ξ in $[a,b]$ zodanig dat voor de restterm (R_N) in (1) geldt

$$R_N = \frac{1}{24} N h^3\, f''(\xi) = \frac{1}{24} (b - a)\, h^2 f''(\xi). \qquad (4)$$

Voor"nette" functies is hiermee een betere schatting voor de restterm gevonden.

Opmerkingen

1. Er zijn diverse andere methoden om formules voor de restterm af te leiden (niet : bewijzen). Bv.

a) Schrijf formule (3) als

$$\int_{x_j}^{x_{j+1}} f(x)\,dx = h\, f_{j+\frac{1}{2}} + R.$$

$R = 0$ indien f een nulde- of een eerstegraads polynoom is. Is f een tweede graads polynoom, bv

$$f(x) = \tfrac{1}{2}(x - x_{j+\frac{1}{2}})^2$$

dan vinden we

$$R = \frac{1}{24} h^3.$$

Indien er dus een formule van de vorm

$$R = C.f''(\xi)$$

bestaat dan moet $C = \frac{1}{24} h^3$ zijn.

b) Schrijf de Taylorreeks van f op (we doen het expres onhandig om duidelijk te maken hoe het in ingewikkelder gevallen gaat) :

$$f(x) = f(x_j) + (x - x_j)\, f'(x_j) + \tfrac{1}{2}(x - x_j)^2\, f''(x_j) + \ldots.$$

Dan wordt $\int_{x_j}^{x_{j+1}} f(x)\,dx = h\,f_j + \frac{1}{2}h^2 f'_j + \frac{1}{6}h^3 f''_j + \ldots,$

$$h\,f_{j+\frac{1}{2}} = h\,f_j + \tfrac{1}{2}h^2 f'_j + \tfrac{1}{8}h^3 f''_j + \ldots$$ (nieuwe reeksontwikkeling)

En dus $$R = \frac{h^3}{24} f''_j + \ldots$$

2. In veel gevallen kent men f'' niet, zodat R_N niet met (4) geschat kan worden. Men kan dan als volgt een schatting krijgen.
Noem de schatting voor de integraal, verkregen uit (1) S_N (dat is dus het rechterlid van (1), zonder de term R_N). Bereken op analoge wijze, nl. door (a,b) in 2N delen te verdelen, S_{2N}. Noem de bijbehorende fout R_{2N}.
We beweren : als f'' in ieder der intervallen (x_j, x_{j+1}) slechts weinig varieert, dan geldt bij benadering

$$R_{2N} = \tfrac{1}{4} R_N. \qquad (5)$$

Hieruit volgt, als we de integraal I noemen, zodat $R_{2N} = I - S_{2N}$, $R_N = I - S_N$, dat bij benadering

$$I = \tfrac{1}{3}[4\,S_{2N} - S_N] =$$

$$= S_{2N} + \tfrac{1}{3}[S_{2N} - S_N] =$$

$$= S_N + \tfrac{4}{3}[S_{2N} - S_N]$$

en dus ook $R_{2N} = \frac{1}{3}(S_{2N} - S_N)$; $R_N = \frac{4}{3}(S_{2N} - S_N)$.

Op deze wijze kan dus uit twee benaderingen een betere benadering voor I en (wat nog belangrijker is) een schatting voor R_N en R_{2N} gevonden worden.
We moeten formule (5) nog bewijzen.
Beschouw een interval (x_j, x_{j+1}) van de oorspronkelijke verdeling. Dit levert tot R_N een bijdrage $\frac{1}{24}h^3 f''(\xi_1)$, met $x_j \leq \xi_1 \leq x_{j+1}$. De bijdrage tot R_{2N} is (vgl. formule (4) met N = 2 en h vervangen door $\frac{1}{2}h$) is $\frac{1}{96} h^3 f''(\xi_2)$, met $x_j \leq \xi_2 \leq x_{j+1}$. Varieert nu f'' weinig in het interval (x_j, x_{j+1}) dan

is $f''(\xi_1)$ ongeveer gelijk aan $f''(\xi_2)$ en dus de bijdrage van het interval (x_j, x_{j+1}) tot R_N ongeveer vier maal zo groot als die tot R_{2N}.

5.2. Hogere orde integratieformules

De integratieformule van 5.1 wordt verkregen door f in het interval (x_j, x_{j+1}) te benaderen door het nulde-graads polynoom $p(x) = f_{j+\frac{1}{2}}$.
Het ligt voor de hand dat we betere benaderingen krijgen door f te benaderen door polynomen van hogere graad.
Een belangrijke klasse van integratieformules (de zg. formule- van Newton-Cotes) verkrijgt men als volgt. Kies een positief geheel getal n (bv. n = 2,3,4...). Verdeel het interval (a,b) in nN gelijke delen met lengte h (dus nN h = (b-a)). Noem weer $x_o = a$, $x_1 = x_o + h$. Etc.
Benader nu f in het interval (x_o,x_n) door een interpolatiepolynoom (met graad $\leq$ n) door de punten $(x_o,f_o),\ldots(x_n,f_n)$. En in het interval (x_n,x_{2n}) door een interpolatiepolynoom door de punten $(x_n,f_n),\ldots(x_{2n},f_{2n})$. Etc.
In het interval (x_o,x_n) vervangen we f dus door het polynoom p, gedefinieerd door (vgl. 4.2.2)

$$p(x_o + sh) = f(x_o) + \binom{s}{1}\Delta f(x_o) + \ldots + \binom{s}{n}\Delta^n f(x_o).$$

En als benadering voor de bijdrage van het interval (x_o,x_n) tot de integraal hebben we dan

$$\int_{x_o}^{x_n} p(x)\,dx = h\int_{o}^{n} p(x_o + sh)ds.$$

Drukken we het resultaat uit in de functiewaarden $f_o,\ldots,f_n$, dan vinden we

n=1 $$\int_{x_o}^{x_1} f(x)\,dx = \frac{h}{2}[f_o + f_1] - \frac{h^3}{12}f''(\xi)$$

n=2 $$\int_{x_o}^{x_2} f(x)\,dx = \frac{h}{3}[f_o + 4f_1 + f_2] - \frac{h^5}{90}f^{(4)}(\xi)$$

$$n=3 \quad \int_{x_o}^{x_3} f(x)\,dx = \frac{3h}{8}\left[f_o + 3f_1 + 3f_2 + f_3\right] - \frac{3h^5}{80} f^{(4)}(\xi)$$

$$n=4 \quad \int_{x_o}^{x_4} f(x)\,dx = \frac{2h}{45}\left[7f_o + 32f_1 + 12f_2 + 32f_3 + 7f_4\right] - \frac{8h^7}{945} f^{(6)}(\xi)$$

etc.

Bij deze formules is tevens een restterm toegevoegd, die op dezelfde wijze als in 5.1 gevonden kan worden. Natuurlijk is de formule voor n = 2 (de zg. trapezium regel) exact als f zelf een eerstegraads polynoom is (interpolatie met eerstegraads polynoom) en die voor n = 3 als f een tweedegraads polynoom is. Deze formule is blijkens de restterm ook nog goed voor alle derdegraads polynomen. Dit wordt veroorzaakt (hoe?) door het feit dat $\int_{x_o}^{x_2} (x - x_1)^3 dx = 0$.

De formule met n = 2 (regel van Simpson) wordt zeer veel gebruikt. Rijgen we bij deze formule alle intervallen aan elkaar dan krijgen we (als $a = x_o$, $b = x_{2N}$)

$$\int_a^b f(x)\,dx = \frac{h}{3}\left[f_o + 4f_1 + 2f_2 + 4f_3 + \ldots + 4f_{2N-1} + f_{2N}\right] + R_N$$

met

$$R_N = -\frac{Nh^5}{90} f^{(4)}(\xi) = -\frac{(b-a)h^4}{180} f^{(4)}(\xi).$$

Als in 5.1 beredeneren we dat, als $f^{(4)}(x)$ in ieder der intervallen (x_{2j}, x_{2j+2}) weinig varieert, bij benadering

$$R_{2N} = \frac{1}{16} R_N.$$

Hieruit volgt (met dezelfde notatie als in 5.1) dat

$$I = \frac{1}{15}\,[16S_{2N} - S_N] =$$

$$= S_{2N} + \frac{1}{15}\,[S_{2N} - S_N] =$$

$$= S_N + \frac{16}{15}\,[S_{2N} - S_N],$$

$$R_{2N} = \frac{1}{15}\,[S_{2N} - S_N] \quad ; \quad R_N = \frac{16}{15}\,[S_{2N} - S_N].$$

Op grond van welke criteria kiezen we de orde n van een te gebruiken integratieformule ? We zien dat de formule met n = 3 een grotere fout heeft dan die met n = 2 (ook als we rekening houden met het feit dat het interval bij n = 3 anderhalf maal zo groot is als bij n = 2). De formule met n = 3 zullen we dus alleen gebruiken als het gewenst is, het totale aantal intervallen tussen a en b oneven te nemen.
De keuze tussen n = 2 en n = 4 hangt grotendeels af van de gladheid van de te integreren functie. Verwacht men dat de hogere afgeleiden groot zijn dan kiest men n = 2. In het algemeen moet men voorzichtig zijn het gebruik van integratie formules van hoge orde.

Opmerking

Moeten we berekenen $I = \int_0^1 f(x) \log x\,dx$ met een "gladde" f dan ontstaan bij iedere integratieformule moeilijkheden, daar de integrand oneindig wordt bij x = 0 (tenzij f(0) = 0). Men kan deze moeilijkheden omzeilen door te schrijven

$$I = \int_0^1 g(x) \log x\,dx + \int_0^1 [f(x) - g(x)] \log x\,dx,$$

waarin $g(x) = f(0) + xf'(0) + \ldots + \frac{x^4}{4!} f^{(4)}(0)$.

De eerste integraal kan expliciet uitgerekend worden, daar

$$\int_0^1 x^k \log x\,dx = \frac{1}{(k+1)^2},$$

de tweede integraal kan met Simpson behandeld worden, daar de integrand nu een continue vierde afgeleide heeft.

5.3. Andere integratieformules met equidistante abscissen

Behalve de in 5.2 besprokene bestaan nog vele andere integratieformules die werken met equidistante abscissen. Voor het numeriek oplossen van differentiaalvergelijkingen zijn vooral van belang formules die de integraal over het interval (x_o, x_n) uitdrukken in de functiewaarden $f_1, \ldots, f_{n-1}$. Dit zijn formules van het zg. open type, in tegenstelling tot die uit 5.2, die van het zg. gesloten type zijn.
We noemen

$$\int_{x_o}^{x_2} f(x)\,dx = 2h\, f_1 + \tfrac{1}{3}h^3 f''(\xi)$$

$$\int_{x_o}^{x_4} f(x)\,dx = \frac{4h}{3}(2f_1 - f_2 + 2f_3) + \frac{14h^5}{45} f^{(4)}(\xi).$$

De eerste van deze formules is dezelfde als die uit 5.1. We leiden de tweede formule af op een enigzins andere wijze dan in 5.2 (het kan natuurlijk op dezelfde manier.
We gaan coëfficiënten a_1, a_2 en a_3 zoeken zodanig dat

$$\int_{x_o}^{x_4} f(x)\,dx = a_1 f_1 + a_2 f_2 + a_3 f_3$$

geldig is voor polynomen van zo hoog mogelijke graad.

$f(x) = 1$ geeft

$$4h = a_1 + a_2 + a_3.$$

$f(x) = x - x_2$ geeft

$$0 = -a_1 h + a_3 h.$$

$f(x) = (x - x_2)^2$ geeft

$$\frac{16}{3} h^3 = a_1 h^2 + a_3 h^2$$

Hieruit volgt : $a_1 = a_3 = \frac{8}{3} h$, $a_2 = -\frac{4}{3} h$.

De coëfficiënt van de restterm vinden we door te nemen $f = (x - x_2)^4$ (waarom niet $f = (x - x_2)^3$?).

Dit levert

$$R = \frac{64}{5} h^5 - \frac{16}{3} h^5 = \frac{112}{15} h^5 = \frac{14}{45} h^5 f^{(4)}(x)$$

(daar $f^{(4)}(x) = 24$). Is er dus een restterm van de vorm $C.f^{(4)}(\xi)$ dan moet $C = \frac{14}{45} h^5$ zijn.

5.4. Integratieformules van Gauss

Veronderstel dat we niet gebonden zijn aan equidistante abscissen. We kunnen dan de vraag stellen : hoe moeten de abscissen $x_1,\ldots,x_n$ en de coëfficiënten $c_1,\ldots,c_n$ gekozen worden zodanig dat in de n-punts formule

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} c_k f(x_k) + R \qquad (1)$$

de restterm voor "nette" functies zo klein mogelijk is. De abscissen $x_1,\ldots,x_n$ en de coëfficiënten $c_1,\ldots,c_n$ moeten weer onafhankelijk van de integrand f zijn.

Stel dat bij zekere keuze van $x_1,\ldots,x_n$ en $c_1,\ldots,c_n$ de restterm in (1) nul is indien f een polynoom met graad m is, doch niet voor alle polynomen met graad m + 1. We zeggen dan : (1) heeft de *precisiegraad m*. We interpreteren de bovengestelde eis nu (enigzins arbitrair) : bepaal $x_1,\ldots,x_n$ en $c_1,\ldots,c_n$ zo dat de precisiegraad van (1) zo hoog mogelijk is.

Een precisiegraad n-1 kunnen we steeds halen, zelfs als $x_1,\ldots,x_n$ voorgeschreven zijn. Want benader f maar door een n-punts Lagrange polynoom (van de graad n-1) met basispunten $x_1,\ldots,x_n$:

$$p(x) = \sum_{k=1}^{n} f(x_k)\, L_k(x),$$

waarin (vgl.4.1)

$$L_k(x) = \prod_{\substack{j=1 \\ j\neq k}}^{n} \frac{x - x_j}{x_k - x_j} \qquad (2)$$

en neem $c_k = \int_a^b L_k(x)\,dx.$

Dan is R = 0 als f een polynoom is met graad $\leq$ n-1, want voor deze polynomen is p(x) = f(x). (Stel dat je de punten $x_1 \ldots x_n$ gevonden hebt. Doe dan ....
Anderzijds kunnen we een precisiegraad 2n niet halen. Zij nl. π het n^e graadspolynoom gedefinieerd door

$$\pi(x) = (x - x_1)\ldots(x - x_n).$$

Dan is $\pi(x_k) = 0$. Voor de functie $f(x) = \pi^2(x)$ is dus

$$R = \int_a^b \pi^2(x)\, dx > 0.$$

We bewijzen dat de precisiegraad 2n-1 haalbaar is en wel op eenduidige wijze. Een dergelijke integratieformule heet integratieformule van Gauss.
Zij f een polynoom met graad $\leq$ 2n-1.
Schrijf

$$f(x) = \sum_1^n f(x_k)\, L_k(x) + g(x)\, \pi(x).$$

Dan is g een polynoom met graad $\leq$ n-1.
Substitutie in (1) levert

$$R = \sum_1^n f(x_k)\left\{\int_a^b L_k(x)\, dx - c_k\right\} + \int_a^b g(x)\, \pi(x)dx.$$

Daar zeker R = 0 moet zijn als g(x) = 0, moet beslist (ga na)

$$c_k = \int_a^b L_k(x)\, dx. \qquad (3)$$

En we zien : (1) heeft de precisiegraad 2n-1 dan en slechts dan als de punten x_k zo gekozen zijn dat

$$\int_a^b g(x)\, \pi(x)\, dx = 0 \qquad (4)$$

voor alle polynomen g met graad $\leq$ n-1.

Hierdoor is π eenduidig bepaald en de nulpunten $x_1,\ldots,x_n$ van π dus ook. Want stel dat π_1 en π_2 voldoen. Dan is ook

$$\int_a^b g(x)\,[\pi_1(x) - \pi_2(x)]\,dx = 0$$

voor alle polynomen g met graad $\leq n-1$. Neem $g = \pi_1 - \pi_2$ (een polynoom met graad $\leq n-1$). Tegenspraak, tenzij $\pi_1 = \pi_2$!
Hoe kunnen we $\pi(x)$ bepalen uit (3) ? Bij voorbeeld zo (het kan geraffineerder). Stel

$$\pi(x) = x^n + a_1 x^{n-1} + \ldots + a_n.$$

Dan kunnen we n lineaire vergelijkingen vinden voor $a_1,\ldots,a_n$, nl.

$$\int_a^b x^k\,\pi(x)\,dx = 0, \quad k = 0,\ldots,n-1.$$

Deze vergelijkingen hebben steeds een oplossing. Want indien niet dan had het bijbehorende homogene stelsel een niet triviale oplossing en dat zou betekenen dat er een n-1 ste graads polynoom was waarvoor (4) geldt.
Zijn de nulpunten van het zo bepaalde n-de graads polynoom wel alle reeel en liggen ze tussen a en b ?
Ja. Stel dat π in het open interval (a,b) m maal van teken wisselt en wel in de punten $\xi_1,\ldots,\xi_m$ (dit zijn dus nulpunten van π). Stel $m < n$. Neem dan in (4) $g(x) = (x - \xi_1)\ldots(x - \xi_m)$. Dan is $g(x)\,\pi(x)$ in (a,b) of steeds $\geqslant 0$ of steed $\leqslant 0$ (en beslist niet $\equiv 0$). Tegenspraak.
π moet in (a,b) dus minstens n maal van teken wisselen, maar dat betekent dat de n nulpunten van π alle verschillend en reeel zijn en in het open interval (a,b) liggen. We kunnen ze dan als punten $x_1,\ldots,x_n$ in (1) gebruiken. De coëfficiënten c_k volgen nu uit (3) (met L_k volgens (2)).

Opmerkingen

1. Alle coëfficiënten c_k zijn positief. Beschouw nl. het polynoom $L_k(x)\,[L_k(x) - 1]$. Dit heeft de graad 2n-2 en is nul in $x_1,\ldots,x_n$ (daar

$L_k(x_j) = 0$ als $j \neq k$ en $L_k(x_k) = 1$). Dus kunnen we schrijven

$$L_k(x)\,[L_k(x)-1] = g(x)\,\pi(x),$$

waarin g een polynoom met graad n-2 is. Uit (4) volgt nu

$$\int_a^b L_k(x)\,[L_k(x)-1]\,dx = \int_a^b g(x)\,\pi(x) = 0$$

en dus is $\quad c_k = \int_a^b L_k(x)\,dx = \int_a^b L_k^2(x)\,dx > 0.$

2. Zij $S_n[f]$ de benadering voor $\int_a^b f(x)\,dx$ die men krijgt met de n-punts integratieformule van Gauss.
Dan geldt voor iedere continue functie f

$$\lim_{n\to\infty} S_n[f] = \int_a^b f(x)\,dx.$$

Bewijs : Zij $\varepsilon > 0$. Dan is er volgens de approximatiestelling van Weierstrasz (zie 4.4) een polynoom p zodanig dat

$$\|f - p\|_\infty = \max_{a\leq x\leq b} |f(x) - p(x)| \leq \varepsilon$$

Stel dat p de graad N heeft. Dan is, daar de n-punts formule van Gauss de precisiegraad 2n-1 heeft stellig

$$S_n[p] = \int_a^b p(x)\,dx$$

zodra $2n-1 \geq N$.
Verder is

$$|S_n[f - p]| = \left|\sum_1^n c_k[f(x_k) - p(x_k)]\right| \leq \varepsilon\sum_1^n c_k \quad \text{(daar alle } c_k > 0\text{)}.$$