Colloquium "Capita uit de Numerieke Wiskunde"

Enige beschouwingen over iteratieve processen en niet lineaire transformaties

door

Dr. J. Berghuis

Zoals bekend kunnen iteratieve processen voor het vinden van (een) wortel(s) van een vergelijking $f(x) = 0$ gegeven worden door

$$x_{n+1} = x_n - c_n f(x_n) \qquad (1)$$

waarbij x_n de n^e iterant is en de coefficient c_n mag afhangen van x_n.

Wij veronderstellen voor het gemak dat de gezochte wortel 0 is; dit kan zonder bezwaar gedaan worden: Zij de wortel n.l. a en $\varepsilon_n = x_n - a$, dan gelden de volgende beschouwingen voor ε_n i.p.v. voor x_n.

De orde α van het iteratieve proces is de exponent voorkomende in de relatie

$$x_{n+1} = A\, x_n^{\alpha} + O(x_n^{\alpha+\beta}) \qquad (2)$$

waarin A een van x_n onafhankelijke constante is en $\beta > 0$.

Hierbij is gebruik gemaakt van het symbool O ook voorkomende bij asymptotische ontwikkelingen, hetwelk aangeeft dat er een vast van x_n onafhankelijk getal C bestaat zodanig dat de restterm in absolute waarde kleiner is dan $C|x_n^{\alpha+\beta}$

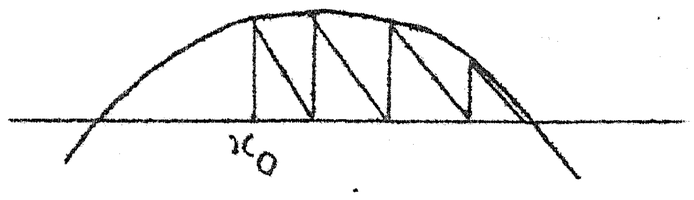
Met behulp van (1) kan indien $f(x)$ en c_n een asymptotische ontwikkeling naar x_n bezitten de relatie (2) worden afgeleid; zelfs merken wij op dat al voldoende is een relatie van de vorm

$$g(x) = \sum_{h=0}^{H-1} g_h\, x^{\varphi h} + O(x^{\Phi H}) \qquad (3)$$

voor één waarde van H.

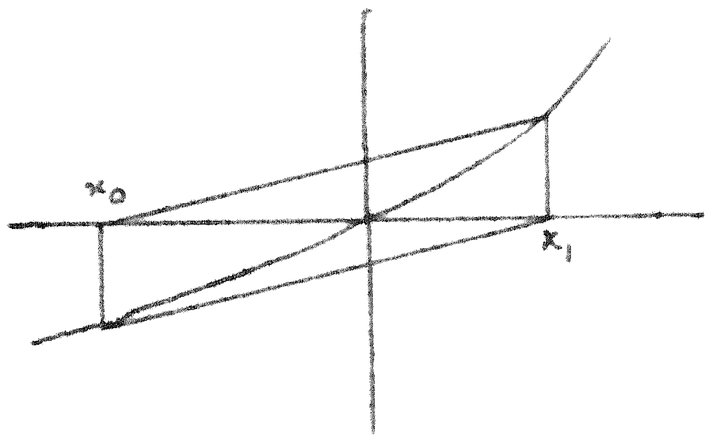
Om maar eenvoudig te beginnen zullen wij $c_n = c$ een constante voor elke n kiezen. Dit proces als beschreven door von Mises (1929) heeft enkele voor rekenmachines plezierige eigenschappen, ook nadelen.

Zij te bepalen alle reele nulpunten van $f(x)$. Indien het mogelijk is $|c|$ te bepalen kleiner dan het omgekeerde van $\text{Max}\,|f'(x)|$ dan vindt men startend in een bepaald punt x_0 een voorgeschreven nulpunt. De waarde van x_n nadert monotoon tot de wortel. Heeft men een keer een wortel gevonden, dan wijzigt men het teken van c, verhoogt of verlaagt de waarde der wortel een klein beetje en itereert door.



Een wortel van even multipliciteit wordt twee keer gevonden, één van oneven multipliciteit slechts één keer.

Kiest men c te groot in absolute waarde, dan bestaat het gevaar dat men nulpunten overslaat, hoewel er ook cycli kunnen ontstaan.

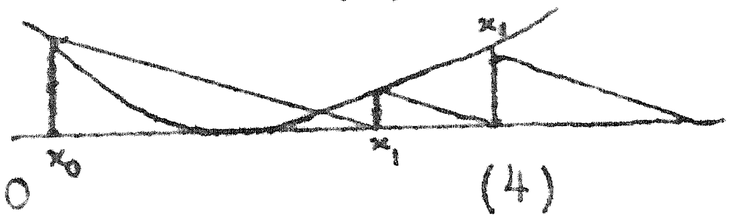


Voorwaarde hiertoe is $f(x_0) + fx_1) = 0$

$f(x_0) + f\left[x_0 - c\, f(x_0)\right] = 0.$

Zelfs meervoudige cycli kunnen ontstaan en al deze cycli kunnen stabiel of instabiel zijn.

Er dient bedacht te worden dat de grootte van c afhankelijk is van f(x) en dus niet voor elke f(x) hetzelfde genomen kan worden.



(4)

We stellen nu dat voor f(x) geldt:

$$f(x) = a_1x + a_2x^2 + O(x^3) \quad \text{met} \quad a_1 \neq 0$$

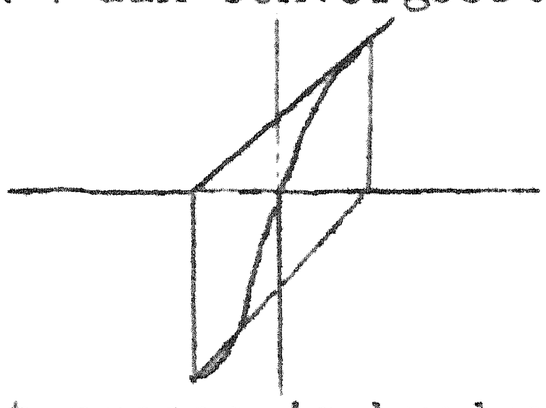
dus $x_{n+1} = (1 - ca_1)x_n + ca_2x_n^2 + O(x_n^3)$

Dit is dus een lineair proces tenzij $ca_1 = 1$. In dit laatste geval is het proces kwadratisch tenzij $a_2 = 0$ enz.

Stel nu $f(x) = a_1x^P + O(x^Q)$ met $Q > 1$,

dus $x_{n+1} = x_n - cf(x_n) = x_n - ca_1x_n^P + O(x_n^Q)$

Is nu $P > 1$ dan is de convergentie lineair maar slecht en indien $P < 1$ dan convergeert het proces niet weer; men krijgt meer cycli



e.d. Er is echter niet meer voldaan aan $|c| < {}^1\!/\!_{\text{Max}}|f'(x)|$

Een tweede voorbeeld is het proces van Newton

$$c_n = {}^1\!/\!_{f'(x_n)}$$

Dit proces is kwadratisch indien f(x) voldoet aan (4), in alle andere gevallen is het lineair. Voldoet f(x) aan

$$f(x) = a_1x^p + O(x^{p+1}) \qquad p > 0 \qquad (5)$$

dan is het proces $x_{n+1} = x_n - pf(x_n)/f'(x_n)$

steeds kwadratisch. Een andere manier om een altijd kwadratisch proces af te leiden is de volgende: De functie $\frac{f(x)}{f'(x)}$ bezit slechts enkelvoudige nulpunten en dus geeft het proces van Newton hierop toegepast een kwadratisch proces:

$$x_{n+1} = x_n - \frac{f(x_n).f'(x_n)}{\{f'(x_n)\}^2 - f(x_n)f''(x_n)} \qquad (6)$$

Een kwadratisch proces heeft altijd een convergentiegebied; trouwens de volgende stelling geldt:

Indien de orde α van een iteratieproces groter is dan 1, d.w.z. er geldt:

$$x_{n+1} = Ax_n^{\alpha} + O(x_n^{\alpha+\beta}) \text{ met } \alpha > 1 \text{ en } \beta > 0$$

dan is er een omgeving aan te geven zodanig dat het proces convergent is:

Er geldt $x_{n+1} = x_n^{\alpha}(A + R_n x_n^{\beta})$

dus

$$\frac{x_{n+1}}{x_n} = x_n^{\alpha-1}(A + R_n x_n^{\beta}).$$

Volgens gegeven blijft R_n beneden een bepaalde waarde en dus is x_n zo klein te bepalen dat $|A + R_n x_n^{\beta}) < 2|A|$ en dan weer is x_n zo klein te bepalen dat

$$|x_n^{\alpha-1}| < \frac{1}{3A}, \text{ wegens } \alpha - 1 > 0.$$

Hieruit volgt $\left|\frac{x_{n+1}}{x_n}\right| < \frac{2}{3}$ en dus convergentie.

Een lineair proces kan convergent, divergent of oscillatorisch zijn. Het proces van Newton is zeker divergent in die punten waarin geldt $f'(x) = 0$. Ook kunnen hierbij cycli optreden waarbij moet gelden

$$\frac{f(x_0)}{f'(x_0)} + \frac{f(x_1)}{f'(x_1)} = 0.$$

Nog een voorbeeld is Lin's methode van "synthetic division" voor het bepalen van wortels van polynomen.

Zij het polynoom $f(x) = \sum_{i=0}^{k} a_i x^i$

en bepaal de coefficienten b_l zodanig dat $b_0 = 0$ en

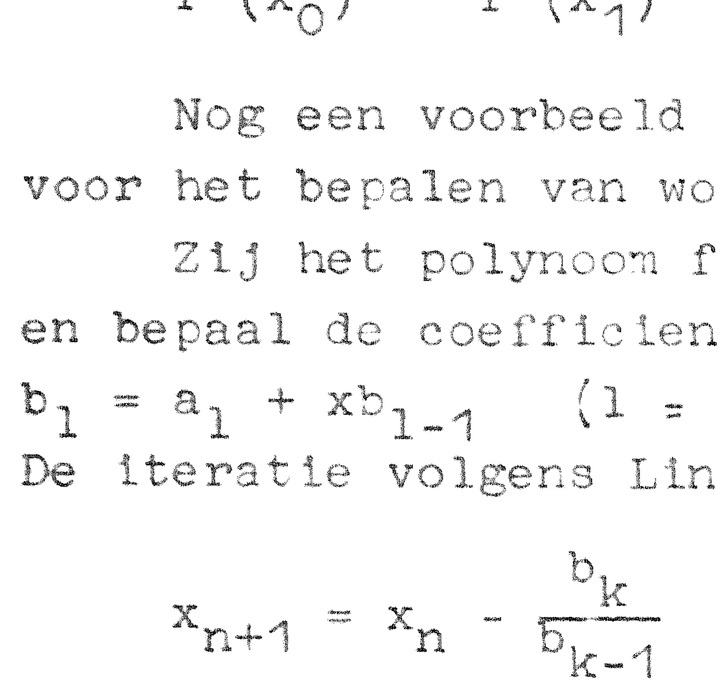
$b_l = a_l + xb_{l-1}$ $(l = 1(1)k)$.

De iteratie volgens Lin is dan

$$x_{n+1} = x_n - \frac{b_k}{b_{k-1}}$$

of anders geschreven $x_{n+1} = x_n - \frac{x_n . f(x_n)}{f(x_n) - f(0)}$ (7)

Divergentie zeker indien $f(x_n) = f(0)$.



Voor de ordebeschouwing van het proces kan men beter de wortel a stellen, $x_n = a + \varepsilon_n$ en dan (4) aannemen. Er volgt

$$\varepsilon_{n+1} = \left[1 + a\frac{f'(a)}{f(0)}\right] . \varepsilon_n + O(\varepsilon_n^2)$$

zodat de divergentie optreedt indien $\left| 1 + a \frac{f'(a)}{f(0)} \right| > 1$ en convergentie als die factor < 1 is. Het proces is lineair tenzij

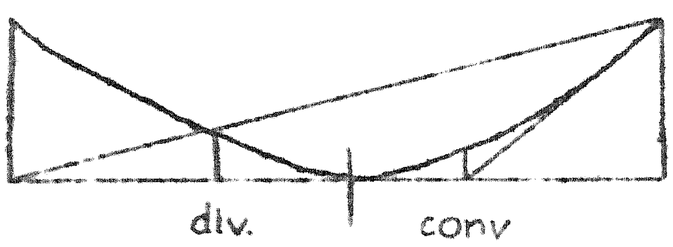
$$a = - \frac{f(0)}{f'(a)} .$$

We schrijven even het (divergente) voorbeeld van Hildebrand (p.454) over:

x	1.3	1.45	0.91	-5.8
1	1	1	1	1
0	1.3	1.45	0.91	
-1	0.69	1.102	-0.172	
-1	-0.103	0.598	-1.157	
Δ x	0.15	-0.54	-6.7	

Treedt divergentie op, dan doet men verstandig het punt x = 0 te verleggen. Neemt men voor f(x) niet (4) maar een andere ontwikkeling aan, dan is de orde van het proces steeds weer te bepalen. De factor $1 + a \frac{f'(a)}{f(0)} = 1$ in het geval dat

$f(x) = (x - a)^P + O((x - a)^Q)$ met $p > 1$ en convergentie hangt af van hogere termen.



De methode van Lin is echter een speciaal geval van "Regula Falsi" die wij nu gaan behandelen.

Wij kiezen eerst twee punten x_0 en x_1 en bepalen dan x_2 enx. volgens

$$x_{n+1} = x_n - \frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)} \cdot f(x_n). \qquad (8)$$

Nu geldt asymptotisch

$$c_n = \frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)} \sim \frac{1}{f'(\alpha)}$$

dus indien f(x) voldoet aan (4) dan is de orde zeker hoger dan 1. Regula Falsi bezit een convergentiegebied, d.w.z. mits men niet willekeurig twee punten x_0 en x_1 in dat gebied kiest, want $f(x_0) = f(x_1)$ betekent steeds divergentie.

Uit (4) en (8) volgt:

$$x_{n+1} = \frac{a_2}{2a_1} x_n x_{n-1} + \text{hogere termen}$$

Stelt men nu $x_{n+1} \sim Ax^{\alpha}$ dan volgt

$$\alpha^2 - \alpha - 1 = 0 \quad \text{of} \quad \alpha = \frac{1 + \sqrt{5}}{2} = 1.618 \text{ enz.} \qquad (9)$$

Newton en Regula Falsi hangen samen: wat bij de een differentiaal-quotient is, is bij de ander differentie quotient. Indien het rekenwerk vereist voor bepaling van $f(x_n)$ groot is, werkt Regula Falsi sneller dan Newton.

De afleiding van hogere orde iteratieprocessen is gegeven door E. Schröder.

Stelt men $y = f(x)$, dan is x als functie van y te ontwikkelen in de omgeving van x_n:

$$x = x_n + \sum_{h=1}^{k-1} \frac{(y - y_n)^h}{h!} \left(\frac{d}{dy}\right)^h x_n + O((y - y_n)^k)$$

De gevraagde wortel x voldoet aan $y = 0$ dus

$$x_{n+1} = x_n + \sum_{h=1}^{k-1} \frac{(-1)^h y_n^h}{h!} \left(\frac{d}{dy}\right)^h x_n$$

is een hogere orde iteratieproces.

Schrijf nu weer $f(x_n)$ i.p.v. y_n en

$$\frac{d}{dy} = \frac{d}{df} = \frac{1}{f'} \frac{d}{dx}$$

verder $\quad \frac{d}{dy} x_n = \left(\frac{dx}{dy}\right)_{x=x_n} = \frac{1}{f'(x_n)}$,

zodat

$$x_{n+1} = x_n + \sum_{h=1}^{k-1} \frac{(-1)^h \{f(x_n)\}^h}{h!} \left[\left\{ \frac{1}{f'(x_n)} \cdot \frac{d}{dx} \right\}^{h-1} \frac{1}{f'(x_n)} \right] \qquad (10)$$

het gevraagde proces is.

Een derde proces (als (4) geldt) is bijv:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{f''(x_n)\{f(x_n)\}^2}{2\{f'(x_n)\}^3} \qquad (11)$$

maar ook

$$x_{n+1} = x_n - \frac{2f(x_n)\, f'(x_n)}{2[f'(x_n)]^2 - f(x_n) f''(x_n)} \qquad (12)$$

Dit laatste proces door P.Wynn toegeschreven aan Richmond wordt gebruikt voor een aardige toepassing (MTAC, April 1956) Hij elimineert n.l. de tweede afgeleide indien de functie f(x) voldoet aan een differentiaalvergelijking van de tweede orde.

Keren we nu weer terug naar Regula Falsi en het proces van Newton. Met behulp van differentie-rekening hebben we uit Newton Regula Falsi gekregen. Ditzelfde proberen wij bij het proces gedefinieerd door (11).

Zij gegeven de drie punten x_n, x_{n-1} en x_{n-2}, daarbij berekend f(x). We leggen een parabool door deze drie waarden en berekenen met behulp daarvan $f'(x_n)$ en $f''(x_n)$.

$$x_{n+1} = x_n - \frac{Ax_n[\,B^2 - AC\,f(x_n)]}{B^3} \qquad (13)$$

met

$$A = (x_{n-2} - x_{n-1})(x_{n-1} - x_n)(x_n - x_{n-2})$$

$$B = (x_{n-1} - x_n)^2 f(x_{n-2}) - (x_{n-2} - x_n)^2 f(x_{n-1}) + (x_{n-2} - x_{n-1})(x_{n-3} - 2x_{n-2} + x_{n-1})\, f(x_n)$$

$$C = (x_n - x_{n-1})f(x_{n-2}) - (x_{n-2} - x_n)f(x_{n-1}) + (x_{n-2} - x_{n-1})f(x_n)$$

Substitueert men (4) teneinde de orde van het proces te bepalen dan vindt men

$$x_{n+1} = -\frac{a_3}{6a_1} x_n x_{n-1} x_{n-2} + \tfrac{1}{2}\left(\frac{a_2}{a_1}\right)^2 x_n^3 + \text{hogere termen} \qquad (14)$$

en de orde α is de wortel van de vergelijking

$$\alpha^3 - \alpha^2 - \alpha - 1 = 0$$

of $\alpha \sim 1,840$

Bij het proces (12) vindt men dezelfde orde. Het resultaat is niet veelbelovend. Uit extrapolerend vindt men bij een proces steunend op k punten

$$x_{n+1} = A\, x_n\, x_{n-1} \cdots x_{n-k+1}$$

en

$$\alpha^k - \alpha^{k-1} - \alpha^{k-2} - \ldots - 1 = 0$$

zodat $\alpha < 2$ maar $\lim_{k \to \infty} \alpha = 2$

Wil men de orde van het proces (13) verbeteren dan zou men de formule zodanig moeten wijzigen dat in (14) de term $x_n\, x_{n-1}\, x_{n-2}$ verdwijnt. We zouden dan bij $x_n\, x_{n-1}^2$ als laagste term $\alpha = 2$ vinden, bij $x_n^2 x_{n-2}$ als laatste geldt $\alpha^3 - 2\alpha^2 - 1 = 0$ of $\alpha \sim 2,2056$

Tenslotte nog iets over het δ^2 proces van Aitken.

Uit de oorspronkelijke reeks geïtereerde waarden x_n vindt men de nieuwe reeks u_n

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

of $\quad u_n = x_{n-2} - (\Delta x_{n-2})^2 / \Delta^2 x_{n-2}$

of $\quad u_n = x_n - (\Delta x_n)^2 / \Delta^2 x_n$ (Steffensen)

Is de orde van het oude proces gelijk aan 1, dan is de orde van de nieuwe reeks 2. Is de orde van de oude reeks gelijk aan $r \neq 1$, dan is die der nieuwe $2r - 1$.
Er volgt dus uit dat u_n slechts in het geval $r=1$ een betere benadering is dan x_n.

Maar van een lineair proces als beschreven door F.B. Hildebrand maakt het δ^2 proces een kwadratisch.
En dat betekent dat het nieuw verworven proces een convergentiegebied heeft:

Lin's methode $x_0 = 1.3 \rightarrow x_1 = 1{,}45$ en $x_2 = 0.91$
$\delta^2 \rightarrow 1.33 \quad x_0 = 1.33 \rightarrow x_1 = 1.3006$ en $x_2 = 1.4460$
$\delta^2 \rightarrow 1.3251$; werkelijke wortel 1.3247180.

Er kan nog opgemerkt, dat men algemener ontwikkelingen dan (4) kan toelaten;
dat als $r > 1$ men iteratieprocessen van willekeurig hoge orde kan bereiken, maar dat men daarmede niet dichter bij het doel komt.

Beschouwt men het proces

$$x_{n+1} = x_n + c\, f(x_n),$$

waarbij voor $f(x_n)$ de ontwikkeling (4) geldt, dan krijgt men door toepassing van Aitken een kwadratisch proces, behalve in de gevallen $c = -1/a_1$ en $c = -2/a_1$.
In het eerste geval $c = 1/a_1$ is het proces zelf kwadratisch en Aitken maakt het van de derde orde, in het tweede is het proces zelf lineair en Aitken geeft weer een derde orde.

Colloquium "Capita uit de Numerieke Wiskunde"

Enige beschouwingen over iteratieve processen en niet lineaire transformaties

door

Dr. J. Berghuis

II

Meer dan een variabele

Onder $\bar{x}$ versta ik een vector met componenten x^i, i = 1(1)n.

Zij de op te lossen vergelijkingen

$$\bar{f}(\bar{x}) = 0 \text{ of } f^i(x',\ldots,x^n) = 0 \text{ voor } i = 1(1)n.$$

We veronderstellen dat er een dergelijke oplossing is.

De iteratieve methoden worden geschreven in een van de twee vormen

$$\text{I} \begin{cases} & \bar{x}(m+1) = \bar{x}(m) + C(m).\bar{f}(\bar{x}(m)) \\ \text{of} & \\ & x^i(m+1) = x^i(m) + c^i_{.k}(m).f^k(x'(m),\ldots,x^n(m)) \quad i=1(1)n \end{cases}$$

$$\text{II} \begin{cases} & \bar{x}(m+1) = \bar{x}(m) + C(m).\bar{f}(\bar{x}) \\ \text{met} & \\ & x^i(m+1) = x^i(m) + c^i_{.k}(m)\, f^k(x'(m+1),\ldots,x^{i-1}(m+1), \\ & \qquad\qquad x^{i+1}(m),\ldots,x^n(m)) \quad i = 1(1)n. \end{cases}$$

Indien onder- en bovenindex dezelfde zijn, dient men hierover te sommeren.

De methoden onder I vallende heten "iterations by total steps" die onder II "iterations by single steps".

Er zij opgemerkt dat de matrix $C(m) = ((C^i_k(m)))$ de rang n moet hebben, anders behoeft men de oplossing niet te vinden. Deze eis komt overeen met de (natuurlijke) eis $c_n \neq 0$ in het enkelvoudige geval.

Bij een veelgebruikte vorm van iteratie bevat de matrix C(m) alleen maar van nul verschillende elementen op de hoofddiagonaal; daarbij moet wel op de volgorde gelet worden van de vergelijkingen f^x.
Bij deze vorm van iteratie is wel door deze van nul verschillende elementen kleiner dan $^1/\text{Max}\left|\frac{\partial f^k_i}{\partial x^i}\right|$ te kiezen te zorgen dat het proces altijd convergent is. De monotonie blijft echter niet gehandhaafd.

Men kan nu weer veronderstellingen maken over de functies $f^1(\bar{x})$ in de omgeving van de oplossing. Zonder de algemeenheid te schaden mag men weer $\bar{x} = \bar{0}$ als oplossing aannemen. Een mogelijke veronderstelling is dan

$$f^i(\bar{x}) = f^i_{.k} x^k + \tfrac{1}{2} f^i_{.kl} . x^k x^l + O(\bar{x}^3) \qquad (3)$$

waarbij de ordeterm aangeeft dat de fout kleiner is dan een constante onafhankelijk van $\bar{x}$ vermenigvuldigd met een homogeen polynoom van de graad 3 in de variabelen

$x^i (i = 1,2,....,n)$.

Een iteratief proces is van de orde α indien

$$\bar{x}\ (m+1) = O(\bar{x}^{\alpha}(m)). \qquad (4)$$

Uit I en (3) volgt

$$x^i(m+1) = x^i(m) + c^i_h\, f^h_k . x^k(m) + \tfrac{1}{2} c^i_h\, f^h_{kl}\, x^k(m) . x^l(m) + O(\bar{x}^3(m))$$

Een proces dat in het algemeen lineair is. Het is kwadratisch, indien

$$1 + c^i_h . f^h_k = 0$$

of $c^i_h . f^h_k = -\delta^i_k$

dus als $C(m) = -((F'))^{-1}$

Dit laatste proces is beschreven door Milne; als eis geldt dat $|F'| \neq 0$. Het vereist echter de kennis van de coëfficiënten f^h_k, welke onbekend zijn. Zij worden evenals bij het proces van Newton benaderd door

te berekenen.

Het analogon van Regula Falsi is ook hier te geven, de differentiaalquotienten moeten dan vervangen worden door differentiequotienten. Het aantal malen dat de functie berekend dient te worden is echter groot bij "total step"-iteratie. Bij de "single step" procedure is een aardige variant, waarvan de orde nog onbekend is.

<u>Frequentie</u>

Laten wij eens het geval van een iteratief proces van eerste orde van een vergelijking met een onbekende beschouwen; de fout x_{n+1} voldoet dan aan de relatie

$$x_{n+1} = a\, x_n + ... \qquad (1)$$

Al sinds 1870 kent men het begrip geitereerde functies: Passen wij de operator (1) n-maal op x_0 toe, dan krijgen wij

$$x_n = a^n x_0 + ...$$

waarin x_0 als een soort amplitude kan worden opgevat en a als een frequentie. De andere termen voorkomende in de operator (1) verstoren dit eenvoudige beeld wel, maar wellicht is iets te

bereiken met het toevoegen van andere frequenties.
Iets dergelijks is ons trouwens bekend uit de theorie van het iteratieproces ter bepaling van eigenwaarden en eigenvectoren bij een symmetrische matrix: Is deze matrix van de orde K, dan krijgen wij iets als

$$x_n = B + \sum_{k=1}^{K} A_k q_k^n \qquad (A)$$

Bij het itereren van de oplossing van meerdere vergelijkingen voldoen de geitereerde waarden in het algemeen niet aan de voorwaarde (A), omdat wij in feite te maken hebben met matrixvermenigvuldigingen van de foutvector.

Met uitzondering van bepaalde nog nader aan te duiden rij, voldoet elke rij x_n aan voorwaarde A op een bepaald moment.

Men kan n.l. bij x_r voor r = ((n - K) (1) (n + K)) in het algemeen 2K + 1 getallen $B_{K,n}$, $A_{k,n}$, $q_{k,n}$, k = 1(1)K aangeven zodanig dat

$$x_r = B_{K,n} + \sum_{k=1}^{K} A_{k,n}\, q_{k,n}^r \qquad r = ((n - K) (1) (n + K) (2)$$

Het getal $B_{K,n}$ kunnen wij de momentane limiet(basis) noemen van de Ke orde.

Gaan wij even terug naar rijen, welke voldoen aan voorwaarde (A): zij kunnen voldoen aan de voorwaarde $|q_k| < 1$ en dus convergent zijn en een limietwaarde hebben n.l. B of zij kunnen divergent zijn en een antilimiet n.l. B hebben. Bovendien kan de convergentie of divergentie nog oscillerend zijn.

Laat nu Δx_n de gewone de gewone voorwaartse differentie van x_n zijn, dan voeren wij in:

$$B_{K,n} = \frac{\begin{vmatrix} x_{n-K} & \cdots & x_n \\ \Delta x_{n-K} & \cdots & \Delta x_n \\ \Delta x_{n-K+1} & \cdots & \Delta x_{n+1} \\ \hline \Delta x_{n-1} & \cdots & \Delta x_{n+K-1} \end{vmatrix}}{\begin{vmatrix} 1 & \cdots & 1 \\ \Delta x_{n-K} & \cdots & \Delta x_n \\ \hline \Delta x_{n-1} & \cdots & \Delta x_{n+K-1} \end{vmatrix}} \qquad (3)$$

met als voorwaarden: $B_{0,n} = x_n$
als de noemer gelijk nul is en de teller niet: $B_{K,n} = \infty$
" " " " " " " " " " ook: $B_{K,n} = B_{K-1,n}$

Opgemerkt dient te worden dat

$$B_{1,n} = \frac{x_{n+1}x_{n-1} - x_n^2}{x_{n+1} + x_{n-1} - 2x_n} \quad . \tag{4}$$

Geschreven in de vorm $B_{K,n} = e_K(x_n)$, zien wij dat de operator e_K geen lineaire is, alhoewel men direct door ontwikkeling van de determinanten naar hun eerste rijen ziet uit (3) dat

$$e_K(x_n) = \frac{c_{n-K}x_{n-K} + \ldots + c_n x_r}{c_{n-K} + \ldots + c_n} \quad . \tag{5}$$

waarin c_i de bijbehorende minoren voorstellen.

Als C een constante is gelden wel de volgende regels

$e_K(Cx_n) = Ce_K(x_n)$

en $e_K(x_n + C) = e_K(x_n) + e_K(C) = e_K(x_n) + C.$

Door optelling van de eerste, tweede tot en met r + eerste rij bij de eerste rij in de determinant van de teller van (3) vindt men

$$e_K(x_n) = \frac{c_{n-K}x_{n-K+r} + \ldots + c_n x_{n+r}}{c_{n-K} + \ldots + c_n}$$

Tot nu toe hebben wij stilzwijgend de $B_{K,n}$ gebruikt in de formule (2) en (3); wij gaan bewijzen dat dit toegstaan is.

Stel eens

$$\varepsilon_n = \sum_{k=1}^{K} A_k q_k^n \qquad (n = 0(1)(2K - 1))$$

Nu zijn de q_k's de wortels van de vergelijking

$$\begin{vmatrix} 1 & q & \ldots & q^K \\ \varepsilon_0 & \varepsilon_1 & \ldots & \varepsilon_K \\ \ldots & \ldots & \ldots & \ldots \\ \varepsilon_{K-1} & \varepsilon_K & \ldots & \varepsilon_{2K-1} \end{vmatrix} = 0$$

Substitueren wij nu voor ε de waarde Δx_{n-K+r} en voor A_k de waarde $A_{k,n}\left[q_{k,n}^{n-K+1} - q_{k,n}^{n-K}\right]$ dan zijn de $q_{k,n}$ de wortels van

$$\begin{vmatrix} 1 & q & \ldots & q^K \\ \Delta x_{n-K} & \Delta x_{n-K+1} & \ldots & \Delta x_n \\ \ldots & \ldots & \ldots & \ldots \\ \Delta x_{n-1} & \ldots & & \Delta x_{n+K-1} \end{vmatrix} = 0$$

of in de notatie van (5)

$$c_{n-K} + \ldots + c_n q^K = 0 \qquad (6)$$

Substitueert men (2) in (5) en maakt men gebruik van bovenstaande vergelijking dan krijgt men

$$e_K(x_n) = B_{K,n}, \qquad (7)$$

indien althans niet aan de vergelijking

$$c_{n-K} + c_{n-K+1} + \ldots + c_n = 0$$

voldaan is. In dit laatste geval voldoet $q = 1$ aan de vergelijking (6) en dan is de propositie (2) niet mogelijk.

Met behulp van de operator e_K zijn weer nieuwe operatoren te definieren, b.v. door herhaald toepassen van e_K, uit de verkregen waarden een bepaalde rij te nemen en daarop weer een e_K toe te passen. Voor hun eigenschappen nader bestudeerd worden geven wij eens een voorbeeld.

Zie Shanks Journal Math. & Phys. XXXIV no. 1, April, 1955

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

n	x_n	e_1	(e_1^2)	$(e_1)^3$	$(e_1)^4$
0	4.0000000				
1	2.6666667	3.1666667			
2	3.4666667	3.1333333	3.1421053		
3	2.8952381	3.1452381	3.1414502	3.1415993	
4	3.3396825	3.1396825	3.1416433	3.1415909	3.1415928
5	2.9760462	3.1427129	3.1415713	3.1415933	3.1415927
6	3.2837385	3.1408814	3.1416029	3.1415925	
7	3.0170718	3.1420718	3.1415873		
8	3.2523659	3.1412548			
9	3.0418396				

Nemen wij hetzelfde voorbeeld en passen hierop de lineaire middelingsoperator toe:

n	A_n	1	2	3	4	5	6	7	8	9
0	4.0000000									
		3.3333333								
1	2.6666667		3.2							
		3.0666667		3.1619048						
2	3.4666667		3.1238096		3.1492064					
		3.1809524		3.1365080		3.1445888				
3	2.8952381		3.1492064		3.1399712		3.1428128			
		3.1174603		3.1434344		3.1410368		3.1421024		
4	3.3396825		3.1376624		3.1421024		3.1413920		3.1418099	
		3.1578644		3.1407704		3.1417472		3.1415174		3.1416868
5	2.9760462		3.1438784		3.1413920		3.1416428		3.1415636	
		3.1298924		3.1420136		3.1415383		3.1416098		
6	3.2837385		3.1401488		3.1416845		3.1415768			
		3.1504052		3.1413554		3.1416152				
7	3.0170718		3.1425620		3.1415459					
		3.1347188		3.1417364						
8	3.2523659		3.1409108							
		3.1471028								
9	3.0418396									

Bij het bestuderen der eigenschappen van de niet lineaire operatoren $B_{k,n}$ zij allereerst opgemerkt dat zij <u>niet regulier</u> zijn:

Wij noemen een operator T regulier indien als de rij x_n convergeert ook $T(x_n)$ convergeert en tevens

$$\lim_{n \to \infty} T(x_n) = \lim_{n \to \infty} x_n .$$

De operator e_1 is niet regulier

$$3 = 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \ldots .$$

De oneven $B_{1,n}$ zijn alle oneindig, de even $B_{1,n}$ alle gelijk aan 3 dus $\lim_{n \to \infty} B_{1,n}$ bestaat niet.

Wel geldt de stelling van S. Lubkin:
Als x_n convergeert en $e_1(x_n)$ convergeert dan
$\lim_{n \to \infty} e_1(x_n) = \lim_{n \to \infty} x_n$.

<u>Bewijs</u>: Stel $e_1(x_\infty) \neq x_\infty$. Nu gelden

$$e_1(x_n) = x_{n-1} + \frac{\Delta x_{n-1}}{1 - R_{n+1}} = x_n + \frac{\Delta x_n}{1 - R_n} \qquad (8)$$

$$\text{met } R_n = \frac{\Delta x_{n-1}}{\Delta x_n} \quad \text{en} \quad n > 0.$$

$$\text{Dus } \frac{\Delta x_n}{1 - R_n} = e_1(x_n) - x_{n-1} \to e_1(x_\infty) - x_\infty \neq 0.$$

Wegens x_n convergent volgt $\Delta x_n \to 0$ dus $1 - R_n \to 0$ of $R_n \to 1$. Er bestaat dus een N zodanig dat voor $n > N$ alle Δx_n hetzelfde teken hebben. Verder houdt ook $1 - R_n$ hetzelfde teken voor voldoend grote n wegens $e_1(x_\infty) - x_\infty \neq 0$.

$R_n > 1$ is niet mogelijk wegens de convergentie van x_n.
Als $R_n < 1$ dan bestaat of Δx_n of $-\Delta x_n$ vanaf $n > N$ uit positieve afnemende termen en convergentie vereist $n . \Delta x_n \to 0$. Maar

$$\frac{n \Delta x_n}{n(1 - R_n)} \longrightarrow e_1(x_\infty) - x_\infty \neq 0$$

dus ook $n.(1 - R_n) \longrightarrow 0$

Volgens een criterium uit K. Knopp, Theorie und Anwendung der Unendliche Reihen p. 286, no. 170 kan dan x_n niet convergeren.

Er volgt dus $e_1(x_\infty) = x_\infty$.

Enige voorbeelden:

I $\begin{cases} x_n & = 1,0,1,0,1,0 \text{ enz.} \\ e_1(x_n) & = \frac{1}{2} \end{cases}$

II $\begin{cases} x_n = 1, 1-2, 1-2+4, \text{ enz. of } \Delta x_n = 1,-2,4,-8,16,-32, \text{ enz.} \\ e_1(x_n) = \frac{1}{3} \end{cases}$

III $\begin{cases} \Delta x_n = 1,2,4,8,16 \\ e_1(x_n) = -1 \end{cases}$

IV $\begin{cases} \Delta x_n = 1, -\frac{1}{2}, +\frac{1}{3}, \ldots \quad \frac{(-1)^{n+1}}{n} \\ \Delta e_1(x_n) = \frac{(-1)^n}{n(4x^2-1)} \quad \text{eerste term } \frac{2}{3} \end{cases}$

D. Shanks, Theorema I. Zij $f(m) = \sum_{i=0}^{M_1} f_i m^i$ en $g(m) = \sum_{i=0}^{M_2} g_i m^i$

met $f_{M_1} \neq 0$, $g_{M_2} \neq 0$ en $g(m) \neq 0$ $(m = 0,1,2,\ldots .)$ en zij

$$A_n = \sum_{m=0}^{n} (-1)^m f(m)/g(m)$$

dan geldt:

a: als $M_2 > M_1$, A_n convergeert en elke afgeleide rij $B_n = e_1(A$ $C_n = e_1(B_n)$ convergeert sneller dan zijn voorganger naar dezelfde limiet.

b: als $M_1 > M_2$, A_n divergeert en elke afgeleide rij divergeert minder sterk dan zijn voorganger tot $e_1^M(A_n)$ met $M \neq \frac{1}{2}[(M_1-M$ De rij $e_1^{M+1}(A_n)$ convergeert en elke verdeelde rij convergeert sneller naar dezelfde limiet.

Bewijs: Uit (8) volgt

$$\Delta B_n = \Delta A_n \frac{\Delta A_{n+1} \Delta A_{n-1} - (\Delta A_n)^2}{(\Delta A_n - \Delta A_{n+1})(\Delta A_{n-1} - \Delta A_n)} =$$

$$= \Delta A_n \frac{f^2(n+1)g(n)g(n+2) - g^2(n+1)f(n)f(n+2)}{[f(n)g(n+1) - f(n+1)g(n)][f(n+1)g(n+2) - f(n+2)g(n+1)]} \qquad (10)$$

Voor voldoend grote waarden van n geldt dan

$$\Delta B_n = \Delta A_n \left\{ \frac{M_1 - M_2}{4n^2} + O\left(\frac{1}{n^3}\right) \right\} \qquad (9)$$

en wegens $$\frac{\Delta A_n}{\Delta A_{n-1}} = -\frac{f(n+1)g(n)}{g(n+1)f(n)}$$

is $\frac{\Delta A_n}{\Delta A_{n-1}}$ negatief.

De reeks $A_0 + \sum_{n=0}^{\infty} \Delta A_n$ is alternerend en indien a), geldt $\Delta A_n \to 0$ dus convergeert de reeks. Wegens (9) convergeert B_n ook en zelfs convergeert B_n sneller. Uit (10) volgt dat geschreven mag worden $\Delta B_n = (-1)^n \frac{f'(n)}{g'(n)}$, waarbij de graad van f' gelijk aan M_1' en die van g' gelijk aan M_2' wordt gesteld. Wegens (9) geldt echter

$$M_2' - M_1' = M_2 - M_1 + 2 > 0$$

en dus is a) bewezen.

Het bewijs van b) is nu duidelijk.

S. Lubkin is verder gegaan dan deze stelling: Hij bewees de stelling voor het geval

$$\frac{\Delta A_n}{\Delta A_{n-1}} = \alpha_0 + \frac{\alpha_1}{n} + \frac{\alpha_2}{n^2} + \ldots.$$

met $|\alpha_0| < 1$ of $\alpha_0 = -1$.

Zij x_n de rij voortgebracht door de kettingbreuk

$$1 + \cfrac{1}{2 + \cfrac{1}{2 + 1}}$$

n	x_n	$e_1(x_n)$	$e_1^2(x_n)$
1	1		
2	3/2	17/12	
3	7/5	99/70	19601/13860
4	17/12	577/408	
5	41/29		

N.B. 19601/13860 = 1.414213564

Het blijkt dat e_1^m toegepast op x_n identiek ik met het itereren van $\sqrt{2}$ met behulp van Newton.

Schrijft men $x_n = \frac{T_n}{N_n}$ dan blijkt uit de recursievergelijkingen

$$T_n = 2T_{n-1} + T_{n-2}, \qquad N_n = 2N_{n-1} + N_{n-2}$$
$$T_1 = N_1 = 1, \qquad T_2 = 3, \qquad N_2 = 2$$

dat $x_n = \sqrt{2}\left[\frac{1 + (2\sqrt{2} - 3)^n}{1 - (2\sqrt{2} - 3)^n}\right]$.

maar uit

$$x_n = \sqrt{2}\,\frac{[1 + x^n]}{[1 - x^n]}$$

volgt $e_1(x_n) = \sqrt{2}\,\dfrac{1 + (x^2)n}{1 - (x^2)n} = x_{2n}$

en eveneens $\frac{1}{2}(x_n + \dfrac{2}{x_n}) = x_{2n} = e_1(x_n)$

zodat bij herhaald toepassen het gevraagde volgt.
Zij x_n de partiële som der reeks

$$c + cx + cx^2 + cx^3 + \ldots. \qquad (11)$$

dus $x_n = \dfrac{c}{1-x} - \dfrac{cx^n}{1-x}$

dan is dus $e_1(x_n) = \dfrac{c}{1-x}$, x_n bevat n.l. één frequentie.
Dit antwoord is onafhankelijk van de grootte van x, zelfs mag de reeks (11) divergent zijn. In het geval $x = 1$ krijgt men $e_1(x_n) = \infty$ maar dit is ook het goede antwoord.

Als volgende voorbeeld beschouwen wij

$$f(z) = \frac{2}{(1 - z)(2 - z)} = 1 + \frac{3}{2}z + \frac{7}{4}z^2 + \frac{15}{8}z^3 + \frac{31}{16}z^4 + \ldots.$$

De partiële sommen der laatste reeks zijn te schrijven in de vorm:

$$x_n = f(z) - \frac{2z}{1-z}\,z^n + \frac{z}{z-2}\left(\frac{z}{2}\right)^n \qquad (12)$$

zoals direct blijkt door f(z) te splitsen in partiële breuken.
Uit (12) volgt nu dat x_n twee frequenties bevat dus

$e_2(x_n) = f(z)$,

behalve de waarden $z = 1$ of $z = 2$.
De vraag rijst wat er gebeurt indien wij e_1^m op de rij (12) toepassen.

Bij het bewezen verband tussen e_1 toegepast op de afgebroken kettingbreuk van $\sqrt{2}$ en het proces van Newton zij nog opgemerkt dat e_1 slechts uit de rationale getallen der kettingbreuk weer rationale getallen levert. Aangezien $\sqrt{2}$ irrationaal is volgt daaruit dat een eindig aantal malen e_1 toepassen slechts een rational benadering voor $\sqrt{2}$ levert.

Vroeger hebben wij al gezien dat toepassing van e_1 op een kwadratisch convergerende rij geen verbetering meer levert, zodat we e_1 niet meer kunnen toepassen op A_1, B_2, C_3 enz.
Hoe het resultaat is wanneer men e_2, e_3 enz. toepast op een kwadratisch convergerende reeks is niet bekend.

Toepassing van e_K met $K > 1$ levert ook slechts rationale getallen, waaruit weer volgt dat de rij der benaderende kettingbreuken tenminste ∞ veel frequenties bevat.

In het nu volgende gaan wij e_1^m toepassen op de partiële sommen der reeksontwikkeling van de functie

$$f(z) = 1/(z - z_0)(z - z_1)$$

met $0 < |z_0| < |z_1|$.

Door $f(z)$ als verschil van twee breuken $C_1(z - z_0)^{-1}$ en $C_2(z - z_1)^{-1}$ te schrijven vindt men voor de te behandelen rij partiaalsommen

$$x_n = f(z) + f(z)\,\frac{z-z_1}{z_1-z_0}\,\left(\frac{z}{z_0}\right)^{n+1} - f(z)\,\frac{z-z_0}{z_1-z_0}\,\left(\frac{z}{z_1}\right)^{n+1} \qquad (13)$$

waaruit blijkt dat x_n precies twee frequenties bevat. Toepassing van $B_{2,n}$ geeft dus direct het goede resultaat.

Maar nu e_1:

We ~~gebruiken het Theorema~~ IV van D. Shanks (loc.cit.).

Uit (13) kan door middel van rekenen worden afgeleid dat

$$B_n = e_1(x_n) = f(z) - \left(\frac{z}{z_1}\right)^{n+2} \frac{f(z)(z_1-z_0)}{(z-z_0)-(z-z_1)(z_0/z_1)^{n+2}} \cdot \qquad (14)$$

D.w.z. dat B_n convergeert binnen de cirkel $|z| = |z_1|$, terwijl x_n convergeert binnen de cirkel $|z| = |z_0|$. Verder is de grootste frequentie $(\frac{z}{z_0})$ geëlimineerd; hiervoor zijn er oneindig veel in de plaats gekomen, zoals men ziet door de noemer in (14) te ontwikkelen

$$B_n = f(z) - f(z)\,\frac{z_1-z_0}{z-z_0} \sum_{i=0}^{\infty} \left[\frac{z-z_1}{z-z_0}\right]^i \left[\frac{z}{z_1}\,\left(\frac{z_0}{z_1}\right)^i\right]^{n+2} . \qquad (15)$$

Door de operator e_1 op (14) toe te passen krijgen we

$$C_n = f(z) - \frac{f(z)(z-z_1)(z_1-z_0)^3 z_0^{-1} z_1^{-1} (zz_0/z_1^2)^{n+3}}{(z-z_0)^2(z-z_1)^2 z_0 z_1^{-3} + \sum_{i=1}^{2} a_i (z_0/z_1)^{in}} \qquad (16)$$

met a_i bepaalde coëfficiënten afhankelijk van z_0, z_1.

Volgens gegeven is $\left|\frac{z_0}{z_1}\right| < 1$ en C_n is dus convergent indien

$$\left|\frac{zz_0}{z_1^2}\right| < 1.$$

En dus weer is het convergentiegebied vergroot tot $\left|\frac{z_1^2}{z_0}\right|$.

Maar als $z = \frac{z_1^2}{z_0}$ dan wordt de frequentie $\frac{z_0 z}{z_1^2} = 1$ en C_n convergeert naar $f(z)\left\{1 - \frac{z_0 z_1}{(z_0+z_1)^2}\right\}$ (17)

een foutief antwoord.

Voor verdere toepassing van e_1 op C_n in het punt $z = z_1^2/z_0$ kan men C_n opvatten als een limiet (17) met frequenties en dat betekent dat bij verdere toepassing van e_1 men steeds het foutieve antwoord krijgt. Verder krijgt men een foutief antwoord in de punten $z = (\frac{z_1}{z_0})^r . z_0$, want deze leveren bij r maal toepassen van e_1 een z_0 frequentie gelijk aan 1.

Indien wij uitgaan van meer dan twee frequenties, dan blijft gelden dat de grootste frequentie geëlimineerd wordt door toepassing van e_1; er worden echter oneindig veel andere ingevoerd. Het convergentiegebied wordt wel vergroot door eliminatie van de grootste frequentie, maar het blijft zeer de vraag of herhaalde toepassing van e_1 het resultaat aanzienlijk zal verbeteren.

Een ieder is bekend met het itereren van de grootste eigenwaarde van symmetrische matrices en weet ook dat toepassing van het δ^2 proces vaak eer een verslechtering dan een verbetering geeft. Het zij opgemerkt dat de rij getereerde waarden in dit geval convergent is en dat herhaald toepassen van e_1 geen foutief antwoord kan geven. Versnelt nl. e_1 het iteratieproces, dus bestaat $e_1(x_\infty)$ dan is $e_1(x_\infty) = x_\infty$ volgens Lubkin. Maar het δ^2 proces kan minder snel convergeren, zoals al gebleken is in het geval dat het iteratieve proces niet van de eerste orde is. Lubkin heeft criteria aangegeven, waarvoor toepassing van e_1 minder snel convergente rijen geeft: o.a. het geval $\frac{\Delta x_n}{\Delta x_{n+1}} \to 1$ is zeer gevaarlijk. In het laatste geval dient e_1 gewijzigd te worden.

Men kan gemakkelijk nagaan dat toepassing van e_n bij itereren van eigenwaarden van een matrix van de orde n direct het antwoord geeft; dat bij het gelijk zijn van twee eigenwaarden e_n hetzelfde antwoord geeft als langdurige iteratie op de gewone wijze. In dit verband is het van belang het gedrag van bijv. e_2 en e_3 te onderzoeken op rijen, die meer dan drie frequenties bevatten. Waarschijnlijk worden de grootste twee of drie frequenties geëlimineerd en zouden deze operatoren van groot belang bij matrixiteraties zijn.

<u>Andereoperatoren</u>

Wij hebben gedefiniëerd de waarden $B_{k,n}$.
Allereerst zij gedefiniëerd de diagonale transformatie,

$$e_d(x_n) = B_{n,n} .$$

Deze is nl. de diagonaal van de getallen

$$\begin{matrix} B_{0,0}, & B_{0,1} & B_{0,2} & B_{0,3} & B_{0,4} \\ & B_{1,1} & B_{1,2} & B_{1,3} & B_{1,4} \\ & & B_{2,2} & B_{2,3} & B_{2,4} \\ & & & B_{3,3} & B_{3,4} \end{matrix}$$

Voeren wij verder in $B_{k,n} = e_k(x_n)$

$C_{k,n} = e_k(B_{k,n})$

$D_{k,n} = e_k(C_{k,n})$

dan wordt $\tilde{e}_k$ gedefinieerd door

$$\tilde{e}_k(x_n) = x_0;\ B_{k,k};\ C_{k,2k};\ D_{k,3k} \text{ enz.}$$

Het is begrijpelijk, dat men ook kan toepassen $e_d^2 = e_d.e_d$ of $\tilde{e}_k.e_d$ of $\tilde{e}_k^2$ enz. De operator e_1 blijkt voor de sommatie van een divergente reeks zeer plezierig: D. Shanks geeft de volgende voorbeelden

$$\ln 3 = 0 + 2 - \tfrac{1}{2}2^2 + \tfrac{1}{3}2^3 - \tfrac{1}{4}2^4 + \ldots$$

$$\int_0^\infty \frac{e^{-t}}{1+t}\,dt = 0{,}596347 = 0! - 1! + 2! - 3! + 4! - 5!$$

$$C = \tfrac{1}{2} + \tfrac{1}{2}B_2 + \tfrac{1}{4}B_4 + \tfrac{1}{6}B_6$$

$$G = \frac{\pi}{2}(3 \ln 3 - 5 \ln 5 + 7 \ln 7 \ldots)$$

$$f(x) = \frac{x^2}{2!} - \frac{x^5}{5!} + \frac{11x^8}{8!} - \frac{375x^{11}}{11!} + \ldots$$

de laatste voor $x > 3.12735$.

Direct kan weer opgemerkt worden dat $\tilde{e}_k$ en e_d slechts rationale getallen leveren wanneer de input x_n rationaal is.

Kiezen wij dus de reeks voor de logarithme

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} \ldots$$

en stellen wij $x = -2$, dan kunnen wij nooit de goede waarde $\log(-1) = \pi i$ of $-\pi i$ vinden (naar gelang het Riemannse oppervlak waarop wij ons bevinden). De operator e_d blijkt goede waarden te leveren mits x zich niet bevindt op de snede lopende langs de negatief reële as. Men noemt dit convergentie behalve in de schaduw van de vertakkingspunten (d.w.z. de willekeurig aan te brengen snede).

Colloquium "Capita uit de Numerieke Wiskunde"

Multi-lengte programmering.

door

E.W. Dijkstra

De wijze, waarop multi-lengte rekenwerk door een automatische rekenmachine het meest efficiënt verricht kan worden, zal in hoge mate beinvloed worden door de specifieke eigenschappen van de betroffen machine.

Zonder te streven naar volledigheid laten wij ter inleiding enkele mogelijkheden volgen.

Wij zullen ons beperken tot machines met vaste komma. (Bij een machine met ingebouwde drijvende komma zal men van de drijvende komma waarschijnlijk geen enkel voordeel hebben, mogelijkerwijs zelfs nadeel, b.v. als er in de code niet voorzien is in een soepele methode ter isolatie van het minst significante gedeelte van een product.)

Het multi-lengte getal wordt opgesplitst in "blokken"; voor elk blok wordt een woord van de standaardlengte ter beschikking gesteld.

Bij een binaire machine rijst allereerst de vraag, of de eenheid van blok 10^n of 2^m maal zo significant moet zijn als de eenheid van het volgende blok.

Om het lange getal in decimale blokken te splitsen, biedt eigenlijk alleen voordeel bij de invoer en de uitvoer; bij administratieve processen via ponskaarten valt het te overwegen, mits de deling door de machine voldoende snel kan worden uitgevoerd. Dit laatste is minder urgent, als het administratieve proces hoofdzakelijk uit additieve bewerkingen bestaat; zij worden vergemakkelijkt als de woordcapaciteit meer dan twee maal zo groot is dan de betroffen tienmacht, die met de bloklengte overeenkomt.

Voor de ARMAC, zonder ingebouwde deling, en voor wetenschappelijke berekeningen, was het opsplitsen in decimale blokken kennelijk niet het overwegen waard.

Van vrij veel invloed is de in de machine gangbare representatie van negatieve getallen (annex de werking van het arimetisch orgaan).

1. Het "complementen-systeem".

Is de woordlengte m+1 bits, (tekencijfer, gevolgd door m binaire cijfers) en denken wij de komma onmiddellijk achter het tekencijfer, dan zijn in dit systeem de capaciteitsgrenzen

$$-1 \leq x \leq 1-2^{-m}$$

0 is hier "het kleinste positieve getal". Een getal wordt van teken gewisseld, door alle cijfers te inverteren (0 door 1 vervangen en omgekeerd) en bovendien 2^{-m} (een 1 op de minst significante plaats) op te tellen. Tekenwisseling geschiedt dan wel m.b.v. de opteller; additieve bewerkingen worden zonder end around carry uitgevoerd.

(Een andere versie van het complementensysteem zou zijn met de capaciteitsgrenzen

$$-(1-2^{-m}) \leq x \leq 1 \ ;$$

0 zou dan "het grootste negatieve getal" zijn. De voorstelling van positieve getallen wordt dan minder vanzelfsprekend; m.i. wegen de - aanvechtbare! - bezwaren van een negatieve nul ruimschoots op tegen het bezit van de +1.)

2. Het inversen-systeem.

Hier zijn de capaciteitsgrenzen

$$-(1-2^{m}) \leq x \leq 1-2^{-m}$$

Er zijn twee representaties voor het getal nul: +0 (allemaal nullen) en -0 (allemaal enen). Een getal wordt van teken gewisseld door alle cijfers te inverteren; dit kan dus geschieden zonder gebruik te maken van de optelller. Additieve bewerkingen moeten echter uitgevoerd worden met end around carry.

Voor de representatie van een getal van q-voudige lengte nemen we aan dat het tekencijfer van elk woord als tekencijfer van het blok fungeert, dat dus met een woordlengte van m+1 bits de opeenvolgende blokken een factor 2^{m} in significantie verschillen.
Als we de blokken als geheel getal opgevat met B_i aanduiden, is het multi-lengte getal gelijk aan

$$B_1.2^{-m} + B_2.2^{-2m} + ... + B_q 2^{-qm}$$

In het complementensysteem nu is het aantrekkelijkste om het meest-significante blok het teken van het getal mee te geven, en alle volgende blokken het +-teken (resp. het teken van de nul) te laten hebben.
(Analoog aan de voorstelling: log 0.2 = -1 + .30103).

Dat dit voordelen heeft wordt het beste geillustreerd met het feit, dat met deze conventies de capaciteitsrestricties (voor positieve nul) luiden:

$$-1 \leqslant x \leqslant 1-2^{-mq} \ .$$

In het inversensysteem verdient het de voorkeur, om elk blok het teken van het lange getal te geven. Dan luiden n.l. de capaciteitsrestricties analoog aan het enkele woord

$$-(1-2^{-mq}) \leqslant x \leqslant 1-2^{-mq}$$

Met betrekking tot de additieve bewerkingen valt - ongeacht of de machine uitgerust is met een echte dubbel-lengte accumulator - de vergelijking tussen beide systemen uit ten gunste van het complementen-systeem. Weinig verschil maakt het bij additie van twee getallen van hetzelfde teken: de optelling wordt uitgevoerd te beginnen bij het minst significante blok, de overdracht, die ontstaat wordt bij de optelling der volgende blokken in rekening gebracht. Bij het complementen-systeem is deze overdracht 0 of 1, bij het inversen-systeem is deze overdracht +1 of +0 bij additie van positieve getallen, -1 of -0 bij die van negatieve getallen. Bij het complementen-systeem gaat de additie van twee getallen van verschillend teken even gemakkelijk; bij het inversen-systeem echter is de inhoud van het minst significante blok van het antwoord afhankelijk van het uiteindelijke teken van de som, dus van iets wat pas na afloop van de optelling bekend is. De optelling van twee getallen van verschillend teken geschiedt nu in twee fasen: eerst worden de getallen bloksgewijs opgeteld (overdracht is uitgesloten); als aan het meest significante blok $\neq 0$ het teken van het antwoord gedetecteerd is, wordt het "teken consistent" gemaakt.

De ARMAC heeft geen dubbel-lengte accumulator; evenwel is dankzij de schuifopdrachten, welker verrichtingen, zoals in een machine werkend met het inversen-systeem voor de hand

ligt, volslagen symmetrisch zijn in de enen en de nullen, een simpele mogelijkheid geboden om de "getekende overdracht" te isoleren, zoals deze voorkomt bij de additie van getallen van hetzelfde teken.

Bij een vermenigvuldiging van twee teken consistente getallen kan hiervan uiteraard alle vruchten geplukt worden.

De multi-lengte routines van de ARMAC zijn bestemd voor het rekenen met breuken, en niet net heel grote gehele getallen. Wat gewoonlijk als deling wordt aangeduid is hier dus vanwege de veronachtzaming van de rest een quotientberekening. Hiervoor is een proces gekozen, waarbij multi-lengte optellingen slechts hoeven worden toegepast op getallen van hetzelfde teken. Er is gebruik gemaakt van een variant van de relatie:

$$\frac{a}{1-c} = (1+c)(1+c^2)(1+c^4)(1+c^8) \ldots\ldots\ldots \qquad |c|<1$$

Evenals bij toepassing van bovenstaande formule nodig, resp. gewenst is, worden teller en noemer van teken gewisseld, als de noemer negatief is en worden ze beide met 2^n vermenigvuldigd, n zo gekozen, dat de nieuwe noemer b voldoet aan de ongelijkheden

$$\tfrac{1}{2} \leq b < 1$$

Als we de zo verkregen teller en noemer met a resp. b aanduiden zou het boven gegeven proces aanleiding geven tot het rekenschema:

Inleiding: $(q_0 = a)$ en $c_0 = 1-b$
Repetitie: $q_{i+1} = q_i + q_i c_i$ en $c_{i+1} = c_i^2$
Resultaat: " als $c_i = 0$, dan geldt $q_i = \frac{a}{b}$ ".

Omdat dit proces per stap twee vermenigvuldigingen van multi-lengte getallen vergt is de volgende variant gekozen:

Inleiding: $(q_0 = a$ en $b_0 = b)$
Repetitie: $c_i \approx 1-b_i$ $q_{i+1} = q_i + q_i c_i$ en
$b_{i+1} = b_i + b_i c_i$
Resultaat: "als $b_i = c$, dan geldt $q_i = \frac{a}{b}$ "

Als in de repetitie exact $c_i = 1-b_i$ gekozen wordt, zijn de processen identiek; we kiezen nu voor c_i een getal, dat slechts in één blok van nul afwijkt (en wel zó, dat

$c_i \leq 1-b_i$, m.a.w. de q_i blijft het quotient van onderen naderen, de c_i blijven dus positief).

De twee vermenigvuldigingen per stap zijn nu teruggebracht tot vermenigvuldigingen van een multi-lengte getal met het enkele-lengte getal c_i. Een en ander gaat - tenslotte - ten koste van de quadratische convergentie. Desondanks is het proces in deze vorm in tijd, en zeker in programmaruimte, aanmerkelijk voordeliger.

Hoewel we het niet hebben toegepast, zie ik geen enkel bezwaar om de quadraat-wortel op een analoge wijze te berekenen: nl.

Inleiding: $w_0 = b.2^{\frac{1}{2}n}$ en $b_0 = b.2^n$ zodat $\frac{1}{2} \leq b_0 \leq 1$
(Ter versnelling van de convergentie):

Repetitie: $c_i \approx \frac{1}{2}(1-b_i)$ $w_{i+1} = w_i(1+c_i)$ en $b_{i+1} = b_i(1+c_i)^2$

Resultaat: als $b_i = 1$, dan is $w_i = b^{\frac{1}{2}}$.

De dubbel-lengte worteltrekking in de ARMAC b.v. maakt er expliciet gebruik van, dat het getal maar in dubbele lengte gegeven is: het argument wordt tussen $\frac{1}{4}$ en 1 genormeerd: dan wordt de (enkele lengte) wortel met de normale wortel-subroutine getrokken, deze wordt met één naslag m.b.v. Newton's iteratie tot tweevoudige lengte verbeterd, en over het halve aantal binaire plaatsen teruggeschreven.

Resumerend moet worden opgemerkt, dat de multi-lengte routines voor de ARMAC nog al ernstig zijn beinvloed door de beperkingen van de machine. Het inversen-systeem is hiervan de minste. Meer sporen hebben achtergelaten de afwezigheid van de ingebouwde deling van het grotere snelle geheugen en wellicht van de "volle" dubbel-lengte accumulator.