TECHNISCHE HOGESCHOOL EINDHOVEN
Afdeling Algemene Wetenschappen
Onderafdeling der Wiskunde

NUMERIEKE WISKUNDE II

Prof. Dr. E.W. Dijkstra

Voorjaarssemester 1962/1963

TECHNISCHE HOGESCHOOL EINDHOVEN

ONDERAFDELING DER WISKUNDE

Voorjaarssemester 1962/1963

NUMERIEKE WISKUNDE II

door

Prof.dr. E.W. Dijkstra

Inhoudsbeschrijving

NUMERIEKE WISKUNDE II

Voorjaarssemester 1962/1963

JdG, 28 Juli 2005

1. APPROXIMATIES VOLGENS DE METHODE DER KLEINSTE KWADRATEN

1.1 Inleiding

Stel, dat wij de positie van een punt in een plat vlak proberen te meten, maar dat de uitkomst van de meting door een meetfout verstoord wordt, waarvoor slechts een waarschijnlijkheidsverdeling gegeven is.

We nemen aan, dat we een willekeurig loodrecht assenkruis met coördinaten x en y gekozen hebben. De kans, dat er in de richting van de x-as een meetfout gemaakt is tussen x en $x + \Delta x$, evenzo in de y-richting een fout tussen y en $y + \Delta y$ zij

$$f(x,y)\,\Delta x \Delta y.$$

Het enige, wat we van de waarschijnlijkheidsdichtheid dan f(x,y) weten is

a) dat f(x,y) nergens negatief is

b) dat $\int_{-\infty}^{+\infty}\int_{\infty}^{+\infty} f(x,y)\,dx\,dy = 1$ is.

Weten wij echter,

a) dat de storingen in twee onderling loodrechte richtingen ongecorrelleerd zijn, dwz. dat f(x,y) te schrijven is als

$$f(x,y) = g(x) * g(y)$$

b) dat het storend verschijnsel "rotatie-symmetrisch" was, dwz. dat de keuze van de richting van het assenkruis willekeurig was,

dan ligt de analytische gedaante van de functie g (en daarmee van de functie f) vast.

Immers:
als wij het assenkruis draaien over een hoek H en overgaan op een nieuw coördinatenstelsel (x1,y1), gegeven door

$$x1 = x * \cos(H) - y * \sin(H)$$
$$y1 = x * \sin(H) + y * \cos(H)$$

dan moet volgens de laatste veronderstellingen gelden

$$g(x) * g(y) = g(x1) * g(y1).$$

Dit moet gelden voor elke hoek H, we kunnen beide zijden dus naar H differentiëren. Omdat de linkerkant niet van H afhangt en voorts geldt

$$\frac{dx1}{dH} = -y1 \quad \text{en} \quad \frac{dy1}{dH} = x1 \quad \text{volgt}$$

$$0 = -y1 * g'(x1) * g(y1) + x1 * g(x1) * g'(y1)$$

of: $$g'(x1)/(g(x1) * x1) = g'(y1)/(g(y1) * y1) .$$

Links staat dezelfde functie van x1 als rechts van y1, x1 en y1 waren echter willekeurig, de beide leden zijn dus gelijk aan een constante. Noem die constante -2 * c, dan volgt uit

$$g'(x1)/g(x1) = \frac{d\ ln(g(x1))}{d\,x1} = - 2 * c * x1$$

onmiddellijk

$$ln(g(x1)) = - c * x1\uparrow 2 + ln\, b$$

en

$$g(x) = b \quad exp(- c * x\uparrow 2)$$

waarbij b uit c volgt door de nevenvoorwaarde

$$\int_{-\infty}^{+\infty} g(x)\, dx = 1 .$$

Uit de eis van convergentie volgt, dat de constante c positief moet zijn. Men spreekt in dit geval over een "normaal verdeelde fout".

(Opm. In de kinetische gastheorie kan men de snelheidsverdeling van Boltzmann "afleiden" op een geheel analoge wijze, als waarop wij nu de normale verdeling hebben "afgeleid".)

Stel nu, dat we voor de plaatsbepaling van een zeker punt in een bepaald, vast coördinatenstelsel een N-tal metingen gedaan hebben. Noem deze coördinatenparen (X1,Y1), (X2,Y2),....., (XN,YN).

Als deze coördinatenparen niet alle gelijk zijn - terwijl ze het wel hadden behoren te zijn - dan kunnen we ons voorstellen, dat een stochastische storing de metingen bedorven heeft. We kunnen ons afvragen, welke positie (X,Y) het aan moet nemen, opdat de waarschijnlijkheid van het optreden van de gemeten coördinatenreeksen zo groot mogelijk zij. Dit betekent, dat men zoekt naar die waarden X en Y zodat

$$f(X1 - X,Y1 - Y) * f(X2 - X,Y2 - Y) * ..... * f(XN - X,YN - Y)$$

zo groot mogelijk is. (De aldus gevonden oplossing heet om begrijpelijke reden: "the solution of maximum likelyhood".) Nemen wij aan, dat f(x,y) van de vorm

$f(x,y) = g(x) * g(y)$ is,

dan valt de bepaling van de meest waarschijnlijke waarden X en Y uiteen in twee separate problemen: X moet bepaald worden zodat

$$g(X1 - X) * g(X2 - X) * \ldots\ldots * g(XN - X)$$

maximaal is en Y volgt uit een geheel analoge voorwaarde. (We zullen daarom verder alleen over X spreken.)

Als wij aannemen, dat de storing normaal verdeeld is, volgt uit

$$g(x) = b * \exp(- c * x\uparrow 2)$$

zonder dat we over de numerieke waarde van c veronderstellingen hoeven te doen al, dat X bepaald is door de voorwaarde, dat

$$(X1 - X)\uparrow 2 + (X2 - X)\uparrow 2 + \ldots\ldots + (XN - X)\uparrow 2$$

minimaal moet zijn. Dit is een quadratische vorm in X met positieve coëfficient van de term X2 en heeft daarom een uniek minimum. Dit wordt aangenomen bij

$$X = (X1 + X2 + \ldots\ldots + XN)/N ,$$

het gemiddelde der observaties.

Omdat de veronderstelde normale verdeling aanleiding geeft tot het streven de som van de kwadraten der discrepanties te minimaliseren, heet dit proces in de wandeling "de methode der kleinste kwadraten". Omdat na differentiatie van de quadratische vorm - om het extremum te beperken - er lineaire verbanden plegen te ontstaan, mag de methode zich verheugen in een grote populariteit, een populariteit, die aanmerkelijk groter is dan de praemissen, die er aan ten grondslag liggen, rechtvaardigen.

1.2 Discrete geval in meer dimensies

Bij passende definitie van de absorptie van licht van speciale golflengte geldt, dat de absorptie van een oplossing van een bepaalde stof evenredig is met de concentratie. Sterker: als wij een gemengde oplossing beschouwen van N opgeloste stoffen met specifieke absorptie

$$a[0],\ a[1],\ldots\ldots,\ a[N - 1]$$

en deze stoffen voorkomen in concentraties

$$c[0],\ c[1],\ldots\ldots,\ c[N - 1],$$

dan is de totale absorptie van de gemengde oplossing gegeven door

$$a[0] * c[0] + a[1] * c[1] + ... + a[N-1] * c[N-1] =$$

$$SIGMA(j,0,N-1,a[j] * c[j]) .$$

Dit geldt bij een vaste golflengte. Als men nu bij een aantal (minstens N) golflengten, waarbij de absorptie van de N zuivere stoffen bekend is, de absorptie van een bepaalde oplossing meet, dan kan men op deze wijze een kwantitatieve analyse van het mengsel uitvoeren. Immers:

als $c[j]$ de - vooralsnog onbekende - concentratie van stof j is;
als $A[i,j]$ bij de i-de golflengte de absorptie is van een oplossing van eenheidsconcentratie van de j-de stof is,
als $y[i]$ de absorptie van het mengsel bij de i-de golflengte is,
dan geldt volgens het bovenstaande

$$A[i,0] * c[0] + A[i,1] * c[1] + ... + A[i,N-1] * c[N-1] = y[i] .$$

Als wij i laten lopen van 0 tot N-1 - dwz. bij N verschillende golflengten absorpties meten - dan hebben we dus N lineaire vergelijkingen met N onbekenden en hieruit kunnen in principe de onbekenden $c[i]$ worden opgelost. In vectornotatie hebben we dan het vergelijkingenstelsel

$$A . \underline{c} = \underline{y}$$

met als oplossing

$$\underline{c} = A^{-1} . \underline{y} .$$

Er is echter alle reden, om naar een betere bepaling van de concentraties uit te zien. De methode is nl. juist aantrekkelijk, waar andere chemische analyses op moeilijkheden stuiten, nl. waar het mengsels van gelijksoortige stoffen betreft. Maar deze hebben vaak de niet verrassende eigenschap, dat hun absorptiespectra niet zoveel verschillen: de kolommen van de matrix A zullen dan neiging vertonen tot onderlinge afhankelijkheid. (Het begint er al mee, dat alle elementen van de matrix A in dit geval positief zijn.) Een voor de hand liggende methode om de soms niet al te grote verschillen in de absorptiespectra der "zuivere componenten" in rekening te brengen, is om de meting uit te breiden tot een aanmerkelijk groter aantal golflengten, zeg $M \gg N$.

Deze absorptiespectra van onze zuivere componenten zijn dan gekarakteriseerd door een matrix M * N matrix A. (Elke kolom van deze matrix is M elementen lang en representeert - voor discrete waarden van de golflengte - het absorptiespectrum van een van de zuivere stoffen.) Onze meetresultaten zijn een kolomvector $\underline{y}$ met M elementen en we zoeken naar een kolomvector $\underline{c}$ van N elementen, zodat

$$A . \underline{c} = \underline{y} .$$

Nu hebben we echter meer vergelijkingen dan onbekenden en we mogen niet verwachten, dat dit stelsel een oplossing heeft. Een manier, om hier onderuit te komen is, om dit te wijten aan meetfouten, die gemaakt zijn bij de bepaling van de elementen van y. Als wij aannemen, dat deze meetfouten normaal verdeeld zijn, dan vinden we de meest waarschijnlijke waarden voor de concentratie c[i] door oplossing van het stelsel

$$A \cdot \underline{c} = \underline{y1}$$

waarbij $\underline{y1}$ zo gekozen worde dat

a) dit stelsel een oplossing heeft, en

b) SIGMA(i,0,M - 1,(y1[i] - y[i])↑2)

- dwz. de som van de kwadraten van de veronderstelde meetfouten - minimaal is.

Wil deze uitdrukking minimaal zijn, dan moet de partiële afgeleide naar c[k] (voor k = 0,1,...,N-1) = 0 zijn. Als we voor y1[i] de uitdrukking in de c's substitueren, dan ziet de te minimaliseren som er als volgt uit:

$$\sum_{i=0}^{M-1} ((A[i,0] * c[0] + A[i,1] * c[1] + ... + A[i,N-1]*c[N-1]-y[i])\uparrow 2)$$

en de (halve) partiële afgeleide naar c[k] geeft

$$\sum_{i=0}^{M-1} (A[i,k] * (A[i,0] * c[0] + ... + A[i,N-1] * c[N-1] - y[i])) =$$

$$\sum_{i=0}^{M-1} (A[i,k] * (y1[i] - y[i])) = 0.$$

Hier staat het scalair product van de k-de kolom van A met de "foutvector" $\underline{y1} - \underline{y}$. De foutvector staat dus loodrecht op de kolommen van A (wij komen hier later op terug). De k-de kolom van A is echter de k-de rij van de getransponeerde A^T van A, en als wij k laten lopen van 0 tot N-1 dan vinden we N vergelijkingen, die in matrixvorm luiden

$$A^T \cdot (\underline{y1} - \underline{y}) = 0$$

of, met $A \cdot \underline{c} = \underline{y1}$

$$(A^T \cdot A) \cdot \underline{c} = A^T \cdot \underline{y}.$$

Hier staan N lineaire vergelijkingen met N onbekenden en als de N × N coëfficiëntenmatrix $A^T \cdot A$ niet singulier is - dwz. als de

kolommen van A lineair onafhankelijk zijn - dan kan men dit stelsel oplossen.

Moet men frequent een dergelijke kleinste-kwadraten-aanpassing uitvoeren met dezelfde matrix A en verschillende "meetseries" y, dan verdient het aanbeveling om te schrijven

$$\underline{c} = B \cdot \underline{y} \quad \text{met} \quad B = (A^T \cdot A)^{-1} \cdot A .$$

Hieruit volgt, dat de elementen van $\underline{c}$ lineair en homogeen afhangen van de elementen van $\underline{y}$. Dat is ook niet zo verwonderlijk, als we ons proces geometrisch interpreteren in een M-dimensionale ruimte, waarin we aan elke meting - golflengte in het geval van kwantitatieve analyse met de absorptiespectrometer - een coördinaatrichting toekennen. De N kolommen van A zijn dan N vectoren, voor willekeurige $\underline{c}$ is de vector $A \cdot \underline{c}$ een vector in de N-dimensionale ruimte, die door de kolommen van A wordt opgespannen. De eis is, om $\underline{y}1 = A \cdot \underline{c}$ zo te kiezen, dat de lengte van $\underline{y}1 - \underline{y}$ minimaal is. We hebben gezien, dat deze eis er toe leidde, dat de vector $\underline{y}1 - \underline{y}$ loodrecht op deze N-dimensionale onderruimte moest staan. Met andere woorden: $\underline{y}1$ is de projectie van $\underline{y}$ op de door de kolommen van A opgespannen onderruimte. Maar projectie is, zoals bekend, een homogene, lineaire operator.

Opm. 1. De methode der kleinste kwadraten is een manier om bij strijdige vergelijkingen - ontstaan door meer vergelijkingen dan onbekenden - er in zekere zin het beste van te maken, wat er van te maken is. Het wordt wel eens vergeten, dat bij afhankelijkheid de methode geen soulaas geeft: als de kolommen van A bijna afhankelijk zijn, dan is $A^T \cdot A$ bijna singulier. Als dit het geval is, dan is dat het teken van onoordeelkundig gebruik van de methode der kleinste kwadraten.

Opm. 2. De matrix A^TA is - bij reële elementen van A - positief definiet. Het iteratie-proces van Gauss-Seidel is dus in principe bruikbaar. Of men het ook gebruiken wil, hangt van vele andere factoren af: ik noem de mate, waarin de elementen op de hoofddiagonaal overheersen en de omstandigheid of men - op andere gronden - over een goede schatting voor de onbekende c's beschikt. Dit laatste zou zich voor kunnen doen, als men een van de metingen verwerpt.

Als de veronderstelling, dat de strijdigheid der vergelijkingen te wijten is aan "meetruis" bij de bepaling van de y[i]'s juist is, dan moeten alle verschillen y1[i] - y[i] van een orde van grootte zijn, die niet onredelijk is vergeleken bij de meetnauwkeurigheid. Als een van deze verschillen abnormaal groot is, dan kan men de veronderstelling maken, dat deze meting suspect is - bv. omdat er een afleesfout gemaakt is. Men kan dan besluiten, deze meting als niet gedaan te beschouwen en deze y-waarde van de overeenkomstige rij uit A te schrappen. Noemen we de nieuwe matrix A1, dan is $A1^T \cdot A1$ gemakkelijk uit $A^T \cdot A$ te berekenen. Als het aantal metingen groot is, mag men veronderstellen een redelijk goede start

voor een iteratieproces te hebben. (En aan dat grote aantal is natuurlijk voldaan, want anders heeft men te weinig rechtvaardiging om een meting te schrappen.)

Opm. 3. Uit de geometrische interpretatie, dat het uiteinde van y1 het voetpunt is van de loodlijn, uit het uiteinde van y neergelaten op de onderruimte, die wordt opgespannen door de kolomvectoren van de matrix A volgt onmiddellijk

a) dat het gevonden extremum voor (y1 - y,y1 - y) inderdaad het minimum is;
b) dat voorts geldt - stelling van Pythagoras -

$$(y,y) = (y1,y1) + (y1 - y,y1 - y).$$

Beide beweringen zijn desgewenst natuurlijk ook rekenenderwijs te verifiëren.

1.3 Het continue geval in meer dimensies

In de vorige paragraaf beschouwden we - geïnspireerd door de verwerking van meetgegevens - een matrix A en een kolomvector met een eindig aantal rijen, respectievelijk elementen, genummerd door een index i.

We krijgen een analoog probleem als we, in plaats van kolomvectoren, functies van een continue onafhankelijk veranderlijke beschouwen, zeg van x.

Gegeven zijn een N-tal functies

$$f[i](x) \qquad (i = 0(1)N - 1; \quad a \leqslant x \leqslant b)$$

die op het interval [a,b] lineair onafhankelijk zijn. Voorts is gegeven een functie y(x), tevens op het interval $a \leqslant x \leqslant b$. Gevraagd de coëfficiënten c[i] zodanig te bepalen, dat, als

$$y1(x) = \sum_{i=0}^{N-1} c[i] * f[i](x) \quad \text{is,}$$

de waarde van de integraal

$$\int_a^b (y(x) - y1(x))\uparrow 2 \; . \; dx$$

minimaal is.

De oplossing van dit vraagstuk geschiedt geheel analoog aan de oplossingsmethode van de vorige paragraaf. We krijgen weer N lineaire vergelijkingen voor de N onbekende c's, maar de coëffi-

ciënten zijn nu niet scalaire producten van vectoren, doch bepaalde integralen van producten van functies.

Opm.

Mits de beschouwde functies zo zijn, dat de voorkomende integralen blijven convergeren, is er niets tegen, als de integralen worden uitgestrekt van $-\infty$ tot $+\infty$.

1.4 Orthogonale polynomen

De voorafgaande probleemstelling gaan we nu deels specialiseren, deels veralgemenen.

De specialisatie zal daaruit bestaan, dat we ons voor de functies f[i](x) zullen beperken tot i-de graadspolynomen in x. Om dit tot uiting te laten komen zullen we ze noteren als "p[i](x)".

De uitbreiding is tweeërlei.
Ten eerste zullen we ons niet meer beperken tot een vast, eindig aantal van deze functies, dwz. we leggen aan de graad der polynomen geen bovengrens op.
Ten tweede voeren we een zg. "gewichtsfunctie" w(x) in. (In de vorige paragraaf beperkten we ons, zoals blijken zal, tot een gewichtsfunctie w(x) = 1.)

De rol van de gewichtsfunctie komt tot uiting in de formulering van de minimaliseringseis: geminimaliseerd moet worden

$$\int_a^b w(x) * (y1(x) - y(x))\uparrow 2 * dx .$$

De functie w(x) dient hierbij non-negatief te zijn. De naam "gewichtsfunctie" is duidelijk: als w(x1) > w(x2) dan betekent dit, dat in x1 een discrepantie tussen y1 en y zwaarder telt dan in x2. Omgekeerd betekent dit, dat men door geschikte keuze van de gewichtsfunctie bereiken kan, dat op bepaalde trajecten van het gebied - nl. door daar w(x) groot te kiezen - de approximatie's beter zullen uitvallen dan op de rest van het gebied $a \leqslant x \leqslant b$.

Opm.

De functie w(x) moet non-negatief zijn, dwz. mag voor sommige waarden van x ook = 0 zijn. De probleemstelling, waarbij een integraal van a tot b geminimaliseerd moet worden, betekent, dat we ons er niet voor interesseren, hoe slecht de benadering uitpakt voor $x < a$ of voor $x > b$. Dit kunnen we ook uitdrukken door de definitie van w(x) uit te breiden met

$$w(x) = 0 \quad \text{voor} \quad x < a \vee x > b$$

en de integratiegrenzen uit te strekken van $-\infty$ tot $+\infty$. Het is duidelijk, dat w(x) voldoende van 0 moet verschillen, dwz. niet zodanig mag zijn, dat de te minimaliseren integraal voor practisch elke functie y1(x) zou blijken = 0 te zijn. (Let wel, dat we bv. geen continuïteit van w(x) verondersteld hebben.)

Ten opzichte van een gekozen gewichtsfunctie w(x) introduceren we voor twee willekeurige functies f(x) en g(x) een inwendig product, genoteerd en gegeven door

$$(f,g) = \int_a^b w(x) * f(x) * g(x) * dx .$$

Opm.

Soms zullen we in plaats van (f,g) noteren "(f(x), g(x))".

We bewijzen nu de volgende

Stelling. Er bestaat een rij polynomen p[i](x) met i = 0,1,2,3,..., waarbij p[i](x) een polynoom van de i-de graad is, voldaan is aan

$$(p[i], p[j]) = \delta[i,j] .$$

Door deze voorwaarde zijn de polynomen op het teken na bepaald; stellen wij bovendien de eis, dat de coëfficiënt van de hoogste macht van het polynoom p[i](x) positief is, dan zijn de polynomen uniek bepaald.

(In de formulering is gebruik gemaakt van het Kronecker-symbool

$$\delta[i,j] = \begin{cases} 0 \text{ als } i \neq j \\ 1 \text{ als } i = j . \end{cases}$$

We zullen deze stelling in twee gedeelten bewijzen. De existentie zullen we constructief bewijzen; daarna zullen we de eenduidigheid bewijzen.

Het constructieve bewijs wordt geleverd door eerst p[0](x) en p[1](x) te constructueren en van dat punt af met volledige inductie alle volgende polynomen.

Het polynoom p[0](x) moet een positieve constante zijn, bepaald door de voorwaarde dat het inwendig product

$$(p[0], p[0])$$

moet zijn. Hieruit volgt, dat

$$p[0](x) = 1 \,/\, \mathrm{sqrt}\left(\int_a^b w(x) * dx \right)$$

is. (Met de notatie "sqrt(x)" wordt de vierkantswortel van x bedoeld.)

De constructie van p[1](x) geschiedt in twee stappen. Eerst construeren we een 1-ste-graads polynoom q[1](x) dat voldoet aan de orthogonaliteitsvoorwaarde

$$(q[1],\ p[0]) = 0 .$$

Stellen we q[1](x) = x + A dan vinden we voor de bekende term A

$$A = -\left(\int_a^b w(x) * x * dx\right) \Big/ \left(\int_a^b w(x) * d(x)\right)$$

$$-\left(\int_a^b w(x) * x * p[0](x) * dx\right) * p[0](x) .$$

We stellen nu dat p[1](x) = C * q[1](x) is, waarbij C een geschikt gekozen positieve constante zal zijn. Aan de orthogonaliteitsvoorwaarde is kennelijk voldaan, immers

$$(p[1],\ p[0]) = (C * q[1](x),\ p[0](x)) =$$

$$C * (q[1],\ p[0]) = C * 0 = 0 .$$

We bepalen C zodanig, dat

$$1 = (p[1],\ p[1]) = C\uparrow 2 * (q[1],\ q[1]).$$

Wat leidt tot de waarde $C = 1 \Big/ \mathrm{sqrt}((q[1],\ q[1]))$.

Na deze voorbereidingen starten we de inductie. Stel, dat we voor een zekere waarde van k polynomen p[i](x) gevonden hebben, die aan

$$(p[i],\ p[j]) = \delta[i,j] \quad \text{voor} \quad i,j \leq k$$

voldoen. We construeren nu een p[k + 1] in twee stappen, nl. als veelvoud van q[k + 1], waarvan we in eerste instantie slechts eisen, dat aan

$$(q[k + 1],\ p[j]) = 0 \quad \text{voor} \quad j \leq k$$

voldaan is.

We stellen

$$q[k + 1](x) = (x + A[k + 1]) * p[k](x) - B[k + 1] * p[k - 1](x).$$

De reden, dat we dit doen is, dat

a) dan $(q[k+1] \cdot p[j]) = 0$ voor $j < k-1$ altijd bevredigd is,
b) de twee "kiesbare" constantes $A[k+1]$ en $B[k+1]$ zo gekozen kunnen worden, dat $(q[k+1], p[j]) = C$ ook voor $j = k-1$ en $j = k$ bevredigend is.

We beginnen met het laatste.

We vinden, met behulp van de inductieveronderstelling

$$0 = (q[k+1], p[k]) = (x * p[k](x), p[k](x)) + A[k+1]$$

waaruit de waarde van $A[k+1]$ onmiddellijk volgt; over het teken van $A[k+1]$ valt niets te zeggen ($A[k+1]$ kan $= 0$ zijn).

De constante $B[k+1]$ vinden we uit de orthogonaliteitseis:

$$0 = (q[k+1], p[k-1]) = (x * p[k-1](x), p[k](x)) - B[k+1].$$

Het k-de graadspolynoom "$x * p[k-1](x)$" kan - op één manier - geschreven worden als lineair compositum van $p[0]$ t/m $p[k]$

$$x * p[k-1](x) = \sum_{j=0}^{k} (c[j] * p[j](x))$$

omdat in $p[k-1](x)$ de coëfficiënt van de hoogste macht van x positief is, geldt dat eveneens in $x * p[k-1](x)$. De enige bijdrage tot de hoogste macht rechts wordt geleverd door de term

$$c[k] * p[k](x)$$

en omdat de hoogste macht van x in $p[k](x)$ eveneens een positieve coëfficiënt heeft, volgt, dat $c[k]$ positief moet zijn. Onze vergelijking voor $B[k+1]$ levert dus - wegens veronderstelde orthogonaliteit -

$$B[k+1] = (x * p[k-1](x), p[k](x))$$

$$= \left(\sum_{j=0}^{k} (c[j] * p[j](x)), p[k](x)\right)$$

$$= c[k] .$$

Hieruit volgt, dat $B[k+1]$ positief is.

Het is nu gemakkelijk in te zien, dat de aldus gedefiniëerde $q[k+1]$ ook voor $j < k-1$ voldoët aan

$$(q[k+1], p[j]) = 0 .$$

Immers:

$$(q[k+1],\ p[j]) =$$

$$(x * p[j](x), p[k](x)) + A[k+1] * (p[j], p[k]) - B[k+1] * (p[j], p[h-1]).$$

De laatste twee inwendige producten zijn onmiddellijk als = 0 te herkennen vanwege de orthogonaliteitsveronderstelling, maar ook het eerste inwendige product is = 0, want omdat $j < k-1$ is het polynoom "$x * p[j](x)$" hoogstens van de graad $k-1$. We hebben dus bewezen, dat het polynoom $q[k+1]$ voor alle $j \leq k$ voldoet aan $(q[k+1],\ p[j]) = 0$. Dit geldt dus eveneens voor het polynoom $p[k+1] = C * q[k+1]$, waarbij we C bepalen uit de voorwaarde, dat

$$(p[k+1],\ p[k+1]) = 1 \quad \text{moet zijn,}$$

waaruit volgt

$$C = 1 \Big/ \mathrm{sqrt}((q[k+1],\ q[k+1])).$$

Omdat de hoogste macht van x in $q[k+1]$ een positieve coëfficiënt had, geldt dit, omdat $C > 0$ is, eveneens voor $p[k+1]$. Resumerend vinden we een recurrente betrekking van de gedaante

$$p[k+1](x) = (P[k+1] * x + Q[k+1]) * p[k](x) - R[k+1] * p[k-1](x)$$

waarbij $P[k+1]$ en $R[k+1]$ beide > 0 zijn.

Door boven beschreven constructie is onder meer het existentiebewijs der orthogonale polynomen geleverd, en daarmee is het eerste gedeelte van onze stelling bewezen. Rest ons nog slechts het bewijs van de eenduidigheid.

Aangenomen, dat we voor $i \leq k$ onze onderling orthogonale polynomen hebben, maar dat voor $i = k+1$ er twee zouden zijn, zeg $p'[k+1]$ en $p''[k+1]$, voldoend aan

$$(p'[k+1[,\ p[i]) = 0 \quad \text{resp.} \quad (p''[k+1],\ p[i]) = 0.$$

Dan geldt voor elk lineair compositum

$$q(x) = a * p'[k+1] + b * p''[k+1]$$

eveneens

$$(q, p[i]) = 0 \quad \text{voor} \quad i \leq k$$

in het bijzonder voor het lineair compositum $q(x)$ waarbij a en b zo gekozen zijn, dat de coëfficiënt van de $k+1$-ste macht = 0 is; voor die waarden van a en b in $q(x)$ hoogstens van de macht k en kan $q(x)$ dus geschreven worden als

$$q(x) = \sum_{j=0}^{k} (c[j] * p[j](x)) \ ;$$

hieruit volgt

$$0 = (q,p[i]) = \left(\sum_{j=0}^{k} (c[j] * p[j](x)),\ p[i](x) \right) = c[i]$$

voor $i \leqslant k$, dwz. het polynoom $q(x)$ is identiek = 0.

De polynomen $p'[k+1]$ en $p''[k+1]$ kunnen dus slechts een factor schelen; uit de voorwaarden

$$(p'[k+1],\ p'[k+1]) = (p''[k+1],\ p''[k+1]) = 1$$

volgt, dat $p'[k+1]$ en $p''[k+1]$ op het teken na gelijk moeten zijn; doordat aan beide de eis is opgelegd van positieve coëfficiënt van de hoogste macht moeten tenslotte de polynomen $p'[k+1]$ en $p''[k+1]$ aan elkaar gelijk zijn. En hiermee is ook de eenduidigheid bewezen.

Omdat elk willekeurig polynoom van de graad k geschreven kan worden als lineair compositum van de polynomen

$$p[0](x) \quad \text{t/m} \quad p[k](x) \quad \text{geldt}$$

voor elk willekeurig k-de graads polynoom $Q[k](x)$

$$(Q[k],\ p[j]) = 0 \quad \text{voor} \quad k < j\ .$$

Hieruit volgt onmiddellijk, dat alle nulpunten van $p[j]$ reëel en enkelvoudig zijn en in het interval (a,b) liggen. Immers:

als $x[1],\ x[2],\ \ldots\ x[k]$ $(x[i] \neq x[j]$ voor $i \neq j)$

de reële wortels van $p[j] = 0$ van oneven multipliciteit in het inwendige van het interval (a,b) zijn, (waarbij meervoudige wortels slechts eenmaal in de reeks zijn opgenomen) dan kiezen we

$$Q[k](x) = (x - x[1]) * (x - x[2]) * \ldots * (x - x[k])\ .$$

Hieruit volgt dat de integrand

$$(Q[k],\ p[j]) = \int_a^b w(x) * Q[k](x) * p[j](x) * dx$$

òf over het hele gebied non-negatief, òf over het hele gebied non-positief is. Hieruit volgt

$$(Q[k],\ p[j]) \neq 0\ ,$$

zodat aan $k < j$ niet voldaan kan zijn. Dus geldt $k = j$.

Tenslotte bewijzen we:
als

$$z[1] < z[2] < \ldots\ldots\ldots < z[k]$$

de naar opklimmende grootte gerangschikte nulpunten van het polynoom $p[k]$ zijn en als evenzo

$$Z[1] < Z[2] < \ldots\ldots\ldots < Z[k] < Z[k+1]$$

de nulpunten van $p[k+1]$ zijn, dan geldt

$$a < Z[1] < z[1] < Z[2] < z[2] < \ldots\ldots < Z[k] < z[k] < Z[k+1] < b$$

maw: de nulpunten van twee opeenvolgende orthogonale polynomen liggen "om en om".

Dit volgt uit de tijdens het constructieve bewijs gevonden recurrente betrekking:

$$p[k+1](x) = (P[k+1] * x + Q[k+1]) * p[k](x) - R[k+1] * p[k-1](x)$$

waarbij $P[k+1] > 0$ en $R[k+1] > 0$ zijn.

Ten eerste merken we op, dat $p[k+1]$ en $p[k]$ geen gemeenschappelijk nulpunt kunnen hebben. Zou dit immers het geval zijn, dan was dit ook een nulpunt van $p[k-1]$, dus ook van $p[k-2]$, etc. en met inductie zou men concluderen, dat het ook een nulpunt van $p[0]$ zou zijn, maar $p[0]$ is een constante > 0.

Uit het feit, dat $p[0] > 0$ is, volgt, dat $p[2]$ in het enige nulpunt van $p[1]$ negatief is, omdat de coëfficiënt van de tweede macht in $p[2]$ positief is, heeft $p[2]$ rechts in links van het nulpunt van $p[1]$ nog een nulpunt, en daarmee is de stelling van de alternerende nulpunten bewezen voor $k = 1$. Met inductie bewijzen we het voor de overige polynomen.

Stel, dat de stelling bewezen is voor $k \leq K - 1$; we moeten bewijzen, dat de nulpunten van $p[K+1]$ onderling door een nulpunt van $p[K]$ gescheiden worden.

Tussen elk tweetal opeenvolgende nulpunten van $p[K]$ ligt, blijkens de inductieveronderstelling, één nulpunt van $p[K-1]$; in de opeenvolgende nulpunten van $p[K]$ heeft $p[K-1]$ dus waarden van alternerend teken; dank zij de recurrente betrekking heeft dus ook $p[K+1]$ in de nulpunten van $p[K]$ waarden van alternerend teken. Tussen elk tweetal opeenvolgende nulpunten van $p[K]$ ligt dus een oneven aantal nulpunten van $p[K+1]$.

Uit het feit, dat de eerste constante uit de recurrente betrekking positief is, volgt, dat alle polynomen $p[i]$ een positieve coëfficiënt van de hoogste macht hebben, dus positief zijn rechts van hun meest rechtse nulpunt. Het meest rechtse nulpunt van $p[K]$ ligt rechts van het meest rechtse van $p[K-1]$, $p[K-1]$ is daar

dus positief en omdat de laatste coëfficiënt van de recurrente betrekking negatief is, volgt daaruit, dat p[K + 1] bij het meest rechtse nulpunt van p[K] dus negatief is. Rechts van het meest rechtse nulpunt van p[K] moet p[K + 1] dus nog een oneven aantal nulpunten hebben. Evenzo kan men aantonen, dat p[K + 1] links van het meest linkse nulpunt van p[K] nog een oneven aantal nulpunten moet hebben.

Op deze manier hebben we K + 1 intervallen aangegeven, die alle een oneven aantal nulpunten van p[K + 1] moeten bevatten. Er zijn echter in totaal precies K + 1 nulpunten van p[K + 1] en in alle intervallen ligt er dus precies ééntje, waarmee het gestelde bewezen is.

(De recurrente betrekking tussen de orthogonale polynomen behelst, dat deze polynomen een zg. "Sturm-rij" vormen. Op de wijze, waarop men deze recurrente betrekking arithmetisch kan benutten, komen wij later nog terug.)

1.5 Functiebenadering met behulp van orthogonale polynomen

We beschouwen nu in het interval [a,b] een N-de graads benadering van de functie f(x). We schrijven deze N-de graadsbenadering

$$\sum_{i=0}^{N} (c[i] * p[i](x))$$

en bewijzen, dat de benadering

$$\sum_{i=0}^{N} (a[i] * p[i](x))$$

waarbij

$$a[i] = (f,p[i])$$

ten opzichte van de betroffen gewichtsfunctie de beste kleinste quadratenbenadering is. (De aldus bepaalde coëfficiënten heten de "Fourier-coëfficiënten".) Immers, voor willekeurige c[i] geldt:

$$0 \leq ((f - \sum_{i=0}^{N} (c[i] * p[i])),(f - \sum_{i=0}^{N} (c[i] * p[i]))) =$$

$$(f,f) - 2 * \sum_{i=0}^{N} (c[i] * (f,p[i])) + \sum_{i=0}^{N} (\sum_{j=0}^{N} (c[i] * (c[j] * (p[i],p[j])))) =$$

$$(f,f) - 2 * \sum_{i=0}^{N} (c[i] * a[i]) + \sum_{i=0}^{N} (c[i]\uparrow 2) =$$

$$(f,f) - \sum_{i=0}^{N} (a[i]\uparrow 2) + \sum_{i=0}^{N} ((a[i] - c[i])\uparrow 2) .$$

Voor het minimum van het rechterlid geldt

$$0 \leqslant (f,f) - \sum_{i=0}^{N} (a[i]\uparrow 2) =$$

$$((f - \sum_{i=0}^{N} (a[i] * p[i])),(f - \sum_{i=0}^{N} (a[i] * p[i]))) \quad ;$$

deze minimumwaarde wordt slechts aangenomen mits

$$c[i] = a[i] \quad \text{voor alle} \quad i \leqslant N .$$

Wij zullen nu bewijzen, dat bij <u>eindig</u> interval [a,b] voor <u>continue</u> functies f(x) dit minimum naar nul gaat als $N \to \infty$.

Wij haken aan bij de stelling van Weiersbass, die zegt, dat voor elke continue functie f(x) op een eindig interval een polynoom pW(x) bestaat, zodat

$$\text{abs}(f(x) - pW(x)) < \varepsilon \quad \text{als} \quad a \leqslant x \leqslant b .$$

Kies nu N minstens gelijk aan de graad van het polynoom pW(x); met f[N](x) noteren we

$$f[n](x) = \sum_{i=0}^{N} (a[i] * p[i](x)).$$

Dan geldt volgens het zojuist bewezene

$$0 \leqslant (f - fN,\ f - fN) \leqslant (f - pW,\ f - pW) .$$

Het middelste lid is gelijk aan

$$(f,f) - \sum_{i=0}^{N} (a[i]\uparrow 2)$$

het laatste lid is $\leqslant \varepsilon\uparrow 2 * \int_a^b w(x) * dx$

als wij N naar oneindig laten groeien vinden we

$$0 \leq (f,f) - \sum_{i=0}^{\infty} (a[i]\uparrow 2) \leq \varepsilon\uparrow 2 * \int_a^b w(x) * dx .$$

De grootheid ε was willekeurig klein, en dus geldt

$$(f,f) = \sum_{i=0}^{\infty} (a[i]\uparrow 2) .$$

(Dit is de stelling van Parseval.)

Beschouw nu de verschilfunctie

$$d(x) = f(x) - \sum_{i=0}^{\infty} a[i] * p[i](x) .$$

Het is gemakkelijk in te zien, dat de Fourier-coëfficiënten van d(x) alle = 0 zijn; hieruit volgt dus

$$(d,d) = \sum_{i=0}^{\infty} (0\uparrow 2) = 0 .$$

Hieruit volgt, dat "overal waar $w(x) \neq 0$ is"

$$d(x) = 0$$

moet zijn.

1.6 Chebyshef polynomen

Uitgangspunt van het volgende is de bekende goniometrische relatie:

$$2 * \cos(a) * \cos(b) = \cos(a + b) + \cos(a - b) . \qquad (1)$$

Substitueren wij $a = n * \varphi$ en $b = \varphi$, dan volgt:

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

Hiermee is een recurrente betrekking gegeven tussen drie opeenvolgende termen van de reeks $\cos(i * \varphi)$ - voor $i = 0,1,2,3....$ -; uit het feit, dat de nulde term van deze reeks = 1 is en de eerste term van deze reeks = $\cos(\varphi)$ is, volgt samen met de gedaante van de recurrente betrekking, dat ook alle volgende termen van de reeks polynomen zijn in $\cos(\varphi)$.

Het ligt daarom voor de hand om cos(φ) als nieuwe onafhankelijke in te voeren; in plaats van "cos(φ)" schrijven we "x", in plaats van "cos(n * φ)" schrijven we T[n](x), nu gegeven door de startwaarden:

$$\left.\begin{aligned} &T[0](x) = 1 \qquad \text{en} \\ &T[1](x) = x \\ &\text{en verder door de recurrente betrekking} \\ &T[n+1](x) = 2 * x * T[n](x) - T[n-1](x) . \end{aligned}\right\} \qquad (2)$$

Opm. 1

Deze polynomen in x worden aangeduid met de letter "T" als eerbetoon aan de Russische wiskundige Chebyshef, wiens naam onder andere ook wel als "Tchebyshef" geschreven wordt.

Opm. 2

Ik prefereer de definitie der Chebyshef-polynomen volgens (2) boven de gesloten vorm

$$T[n](x) = \cos(\arccos(x) * n) ;$$

bij deze definitie moet je nl. apart aantonen, dat de meerduidigheid van de arccos geen aanleiding geeft tot dubbelzinnigheden en dat T[n](x) inderdaad een polynoom in x is. Tenslotte spreekt deze alternatieve definitie mij bij complexe x of bij abs(x) > 1 beslist minder aan. Een en ander neemt niet weg, dat wij geregeld op de interpretatie van een cosinus zullen terugvallen.

Wij bewijzen nu de volgende

Stelling 1. De Chebyshef polynomen zijn onderling loodrecht ten aanzien van de gewichtsfunctie

$$1 \,/\, \mathrm{sqrt}(1 - x\uparrow 2)$$

op het interval $-1 \leqslant x \leqslant +1$.

$$(T[n], T[m]) =$$

$$\int_{-1}^{+1} T[n](x) * T[m](x) \,/\, \mathrm{sqrt}(1 - x\uparrow 2) * dx =$$

$$\int_{-1}^{+1} \cos(n * \varphi) * \cos(m * \varphi) \,/\, \sin(\varphi) * d(\cos(\varphi)) =$$

$$\int_0^{\pi} \cos(n * \varphi) * \cos(m * \varphi) * d\varphi =$$

$$\tfrac{1}{2} \int_0^{\pi} (\cos((n + m) * \varphi) + \cos((n - m) * \varphi))\, d\varphi$$

Aangezien voor gehele M geldt

$$\int_0^{\pi} \cos(M * \varphi) * d\varphi = \begin{cases} \pi \text{ voor } M = 0 \\ 0 \text{ voor } M \neq \end{cases}$$

volgt hieruit dat

$$(T[n], T[m]) = \begin{cases} = 0 & \text{voor } n \neq m \\ = \pi/2 & \text{voor } n = m \neq 0 \\ = \pi & \text{voor } n = m = 0 \end{cases}$$

Opm.

De polynomen T[n] zijn een orthogonaal stelsel, ze zijn niet orthonormaal, omdat

$$(T[n], T[n]) \neq 1$$

is. Als dit inwendige product voor alle n gelijk $\pi/2$ geweest was, had men ongetwijfeld alle polynomen met sqrt(2/π) vermenigvuldigd (de recurrente betrekking was dan immers blijven gelden). Nu echter (T[0], T[0]) = π is, heeft men de recurrente betrekking laten prevaleren. Het gevolg is, dat in veel formules over Chebyshef-polynomen homogeen een factor π ergens optreedt, en dat de term met T[0] wel eens een extra factor 2 meekrijgt. Wij zullen hiervan straks voorbeelden tegenkomen.

Uit de recurrente betrekking (2) volgt dat

1) de coëfficiënt van de hoogste macht van x in T[n](x) gelijk is aan $2\uparrow(n - 1)$.
2) als n even (oneven) is, T[n](x) een even (oneven) functie van x is.

Uit de interpretatie van T[n] als cosinus volgt

voor abs(x) $\leq$ 1 geldt: abs(T[n](x)) $\leq$ 1

waarbij het gelijkteken voor n > 0 geldt in de n + 1 punten

$$x = \cos(i * \pi/n) \quad \text{voor} \quad i = 0,1,\ldots,n\,.$$

Stelling 2: Als Q[n](x) een n-de graadspolynoom is met coëfficient van de hoogste macht = 2↑(n - 1) dan geldt:

$$\max_{abs(x)\leq 1} (abs(Q[n](x)) \geq 1$$

waarbij het gelijkteken slechts geldt, als Q[n] = T[n].

Bewijs: Als in het interval $-1 \leq x \leq 1$ het maximum van de absolute waarden van Q[n] overal <1 is, dan neemt het polynoom van de graad n - 1

$$P = T[n] - Q[n]$$

in de punten $x = \cos(i * \pi/n)$, waar T[n] zijn extreme waarden $\pm$ 1 aanneemt, waarden aan van hetzelfde teken, als T[n]. In deze n + 1 punten heeft T[n] echter alternerend teken, dus ook P zou in deze n + 1 punten alternerend teken hebben. P zou dus minstens n nulpunten moeten hebben, wat voor een polynoom van de graad n - 1 dat niet identiek nul is, een beetje te veel is, waarmee het eerste gedeelte van de stelling bewezen is.

Het bewijs, dat het gelijkteken slechts geldt, als Q[n] = T[n] stellen we gedeeltelijk uit. De redenering is dat als voor alle

$$x[i] = \cos(i * \pi/n) \qquad \text{voor} \quad 0 \leq i \leq n$$

aan de gelijkheid

$$Q[n](x[i]) = T[n](x[i]) \qquad \text{voor} \quad 0 \leq i \leq n \qquad (4)$$

voldaan is, dat dan (volgens Lagrange!) de polynomen Q[n] en T[n] identiek gelijk zijn, terwijl we later zullen laten zien, dat als niet aan de n + 1 gelijkheden (4) voldaan is, we een polynoom P(x) van de graad n kunnen construeren, dat

a) coëfficiënt van de hoogste macht = 2↑(n - 1) heeft

b) voldoet aan $\max_{abs(x)<1} (P(x)) < 1$

wat volgens het eerste gedeelte van deze stelling onmogelijk is, zodat dus aan de n + 1 gelijkheden (4) voldaan moet zijn.

We zien dus, dat van alle n-de graads polynomen met constante coëfficiënt van de n-de macht (een veelvoud van) het Chebyshef-polynoom zich op het gebied [-1,+1] het minst ver van de x-as verwijdert.

Van een op het interval [-1,+1] gegeven continue functie f(x) kunnen we een zg. "Chebyshef-ontwikkeling" maken, die we schrijven

$$f(x) = \sum_{i=0}^{\infty}{}' a[i] * T[i](x)$$

(waarbij met het symbool $\sum'$ bedoeld is, dat de nulde term van de reeks gehalveerd moet worden, dwz.:

$$\sum_{i=0}^{\infty}{}' t[i] = \tfrac{1}{2} * t[0] + \sum_{i=1}^{\infty} t[i] \qquad .)$$

Deze conventie is zo gekozen, omdat nu de coëfficiënten a[i] voor alle i gegeven zijn door

$$a[i] = \frac{2}{\pi}\,(f(x),\ T[i](x))$$

$$= \frac{2}{\pi}\int_{-1}^{+1} f(x) * T[i](x) \Big/ \mathrm{sqrt}(1 - x\uparrow 2)\,dx$$

$$= \frac{2}{\pi}\int_{0}^{\pi} f(\cos(\theta)) * \cos(r * \theta) * d\theta\ .$$

Deze relaties volgen onmiddellijk uit de betrekkingen (3).

Wij gaan er even aan voorbij, dat deze wijze van berekening van de a[i]'s, voor het geval, dat de integralen niet analytisch uitgerekend kunnen worden, misschien niet de meest practische is, om eerst even te kijken naar het convergentiegebied van dergelijke Chebyshefreeksen.

Van de machtreeks

$$\sum_{i=0}^{\infty} c[k] * z\uparrow k$$

is bekend, dat hij in het complexe vlak convergeert binnen een cirkel met straal R, wanneer de reeks

$$\sum_{k=0}^{\infty} c[k] * R\uparrow k$$

absoluut convergent is.
Wij vragen ons af, wanneer de reeks

$$\sum_{i=0}^{\infty}{}' a[i] * T[i](z) \qquad (5)$$

convergent is, als van de a[i]'s gegeven is, dat de reeks:

$$\sum_{i=0}^{\infty}{}' a[i] * R\uparrow i$$

absoluut convergent is. Wij beperken ons tot het geval, dat $R > 1$ is. (Immers: op het interval $[-1,+1]$ blijven de polynomen $T[i](x)$ voor $i \to \infty$ steeds geregeld de waarden ± 1 aannemen. Wil de reeks convergeren, dan moeten de $a[i] \to 0$ gaan. Wij kiezen het geval, dat de $a[i]$ minstens als $R\uparrow(-i)$ naar nul gaan.)

Onder de veronderstelling (6) convergeert de reeks (5) voor die waarden van z, waarvoor

$$abs(T[i](z)) < \text{constante} * R\uparrow i \ .$$

Het gedrag van $T[i](z)$ voor grote waarde van i volgt onmiddellijk uit de recurrente definitie (2), een homogene recurrente betrekking, die als algemene oplossing heeft

$$T[i](z) = c1 * x1\uparrow i + c2 * x2\uparrow i$$

waarbij x1 en x2 de wortels zijn van de vierkantsvergelijking

$$x\uparrow 2 - 2 * z * x + 1 = 0 . \qquad (7)$$

We moeten, om te onderzoeken, in welk gebied de reeks 5 convergent is, vaststellen, voor welke waarden van z de grootste wortel van (7) een modulus $\leq R$ heeft.

We onderzoeken eerst het gebied, waar de grootste wortel - noem deze x1 - in absolute waarde exact = R is.

Noem $x1 = \cos(\varphi) * R + i * \sin(\varphi) * R$;

dan is (het product van de wortels is immers = 1)

$$x2 = \cos(\varphi) / R - i * \sin(\varphi) / R \ .$$

De bijbehorende waarde van $z = u + i * v$ is echter het gemiddelde van deze twee wortels, zodat

$$2 * z = \cos(\varphi) * (R + 1/R) + i * \sin(\varphi) * (R - 1/R)$$

maw. $\cos(\varphi) = u / (.5 * R + .5/R)$

$$\sin(\varphi) = v / (.5 * R - .5/R)$$

waaruit volgt, dat

$$(u/(.5 * R + .5/R))\uparrow 2 + (v/(.5 * R - .5/R))\uparrow 2 = 1.$$

Maw.: z ligt op een ellips met de lange as van lengte $R + 1/R$

langs de reële as, en de korte as van lengte R - 1/R langs de imaginaire as. Het is niet moeilijk, om in te zien, dat de Chebyshefreeks convergeert in het inwendige van de ellips en divergeert daarbuiten.

Opm.

Als R zakt tot 1, ontaardt de ellips in het stuk van de reële as tussen -1 en +1.

Voorbeeld

Ter illustratie van de macht van Chebyshef-ontwikkelingen zullen we

$$f(x) = 1 / (1 + x\uparrow 2)$$

benaderen over het gebied $-2 \leqslant x \leqslant 2$. Dit is met een machtreeks helemaal niet mogelijk, omdat deze vanwege de polen bij $\pm i$ een convergentiecirkel met straal 1 heeft. In ons geval geven deze polen de vorm van de convergentie-ellips aan, nl. lange as 2 * zo lang als de korte as:

$$R + 1/R = 2 * (R - 1/R)$$

$$R\uparrow 2 = 3 \rightarrow R = \sqrt{3}.$$

Om te zorgen dat de convergentie-ellips behalve de goede vorm, ook het goede formaat heeft, gaan we over op een andere veranderlijke

$$y = \alpha * x$$

zodat

$$y = \sqrt{3} + 1/\sqrt{3}$$

met x = 4 (de lengte van de lange as, in y, resp. x uitgedrukt) overeenkomt. Dit geeft

$$\alpha * 4 = \sqrt{3} + 1/\sqrt{3} = 4/\sqrt{3} \qquad \text{en}$$

$$\alpha = 1/\sqrt{3}.$$

We gaan dus

$$F(y) = f(x) = f(y * \sqrt{3}) = 1 / (1 + 3 * y\uparrow 2)$$

ontwikkelen als Chebyshefreeks

$$\sum_{k=0}^{\infty}{}' a[k] * T[k](y).$$

Hierbij zijn de coëfficiënten gegeven door

$$a[k] = \frac{2}{\pi} \int_0^{\pi} \frac{\cos(k*\theta)}{1+3*(\cos(\theta))\uparrow 2}\, d\theta$$

$$= \frac{2}{\pi} \int_0^{\pi} \frac{\cos(k*\theta)}{2.5+1.5*\cos(2*\theta)}\, d\theta$$

noemen we k/2 = n, en 2 * θ = φ, dan vinden we

$$a[k] = \frac{1}{\pi} \int_0^{2*\pi} \frac{\cos(n*\varphi)}{2.5+1.5*\cos(\varphi)}\, d\varphi$$

overgang op de variabele z = exp(i * φ) geeft ons

$$a[k] = \frac{1}{i*\pi} \oint \frac{.5*(z\uparrow n+z\uparrow(-n))}{2.5+.75*(z+1/z)} * \frac{dz}{z}$$

waarbij de kringintegraal genomen moet worden over de eenheidscirkel.

Als wij in de oorspronkelijke uitdrukking voor a[k] van de variabele θ overgaan op ψ = π - θ, dan volgt onmiddellijk, dat

$$a[k] = 0 \qquad \text{voor oneven } k.$$

(Dit was te verwachten, want onze oorspronkelijke functie was even.) Wij kunnen ons dus bij n = k/2 beperken tot gehele n.

We kunnen nu a[k] schrijven als de som van twee integralen

$$a[k] = I1 + I2 =$$

$$\frac{2}{3} * \frac{1}{i\pi} \oint \frac{z\uparrow n}{(z+1/3)*(z+3)}\, dz +$$

$$\frac{2}{3} * \frac{1}{i\pi} \oint \frac{z\uparrow(-n)}{(z+\frac{1}{3})*(z+3)}\, dz \quad ,$$

waarbij beide integralen over de eenheidscirkel genomen moeten worden. Als wij in de tweede integraal overgaan op de nieuwe variabele

$$u = 1/z$$

dan vinden we I2 = I1. (Deze substitutie leidt tot

$$dz = - u\uparrow(-2) * du \ ,$$

maar dit minteken wordt opgesoupeerd door het feit dat als z de eenheidscirkel in de ene richting doorloopt, u de eenheidscirkel in de andere richting doorloopt.)

De integrand van I1 heeft binnen de eenheidscirkel slechts een enkelvoudige pool bij z = - 1/3, het residu is

$$2\pi i * \frac{(-1/3)\uparrow n}{8/3} ,$$

I1 is gelijk aan $\frac{2}{3} * \frac{1}{i\pi}$ maal dit residu en we vinden

$$a[k] = a[2 * n] = 2 * I1 = (- 1/3)\uparrow n .$$

Maw.: de a[k]'s gaan, zoals voorspeld, met $R\uparrow(-k)$ naar nul. (R was in dit voorbeeld immers = $\sqrt{3}$.)

Onze uiteindelijke ontwikkeling is dus

$$\sum_{n=0}^{\infty}{}' ((- 1/3)\uparrow n * T[2 * n](x/\sqrt{3})) .$$

Chebyshef-approximaties worden eigenlijk haast altijd slechts op de reële as gebruikt. De stelling over het elliptisch convergentiegebied in het complexe vlak is dan ook minder bedoeld om aan te geven, voor welke waarden van z een Chebyshef-ontwikkeling convergeert, als wel om ons een onmiddellijk inzicht te geven in hoe, op grond van de ligging van de polen, de coëfficiënten van Chebyshef-ontwikkeling naar nul gaan.

De reeks uit het laatste voorbeeld convergeert voor reële x mits abs(x) < 2; in de praktijk zal men deze reeks echter uitsluitend gebruiken voor abs(x) $\leq \sqrt{3}$ = 1.73205, dwz. voor die waarden van x, waarvoor het argument van de Chebyshef-polynomen absoluut ≤ 1 is. Omdat voor dat gebied de Chebyshef-polynomen absoluut ≤ 1 zijn, is hiermee nl. onmiddellijk een majorant van de restterm gegeven, als we van de Chebyshefreeks slechts een eindig aantal termen meenemen.

Het is om die reden gewenst, dat de coëfficiënten "behoorlijk snel" naar nul gaan; we prefereren dus benaderingen met grote waarden voor R. Hoe langer gerekt de ellips, hoe dichter R tot 1 zakt. Is dus wiskundig wel waar, dat we voor de functie $1/(1 + x\uparrow 2)$ een convergente Chebyshef-ontwikkeling kunnen opschrijven, die geldt op een willekeurig eindig interval abs(x) $\leq$ X, voor grote waarden van X is de convergentie zo langzaam, dat de reeks dan aanzienlijk aan bruikbaarheid inboet.

Een tweede voordeel van een Chebyshef-ontwikkeling, waarvan de coëfficiënten behoorlijk snel naar nul gaan, is dat we dan meer gerechtvaardigd zijn om in het geval van een afgebroken Chebyshefreeks de eerst-verwaarloosde term als hoofdcomponent van de gemaakte "truncation error" te beschouwen. Dit betekent, zoals wij straks zullen laten zien, dat

$$\sum_{i=0}^{n}{}' a[i] * T[i](x)$$

een zeer goede benadering is van het n-de graads polynoom, dat op het interval [-1,1] de functie f(x) met minimale maximumfout benadert. (Zie onder.)

Opm. 1

Voor $-1 \leq x \leq 1$ convergeert een convergente Chebyshefreeks uniform.

Opm. 2

Als de afgekapte reeks slechts gebruikt wordt ter benadering van een functie op het interval [-1,+1], kunnen we ermee volstaan, de coëfficiënten op te geven in een vaste, absolute precisie.

1.7 De sommatie van Chebyshefreeksen

Zij

$$f(x) = \sum_{i=0}^{n} (a[i] * p[i](x))$$

waarbij p[i](x) de orthogonale polynomen ten opzichte van een of andere gewichtsfunctie zijn. Wij veronderstellen de numerieke waarde der coëfficiënten gegeven; gevraagd zij om voor een bepaalde waarde van x, zeg x0, de functiewaarde f(x0) numeriek te bepalen.

Als de polynomen p[i](x) door hun coëfficiënten gegeven zijn, vergt de berekening van p[i](x0) ongeveer i stappen (elk bestaande uit een vermenigvuldiging en een optelling), de uiteindelijke sommatie vergt dan nog n + 1 dergelijke stappen, het totaal vergt circa n * (n + 1)/2 van dergelijke stappen.

Als wij ons echter herinneren, dat elk stel orthogonale polynomen, ongeacht de vorm van de gewichtsfunctie aan een 2-de orde recurrente betrekking voldoet, dan kunnen we een aanmerkelijk goedkopere methode afleiden. Nadat we p[0](x0) en p[1](x0) hebben berekend, kost het ons hoogstens 3 vermenigvuldigingen en hoogstens 2 optellingen per volgende term uit de rij p[i](x0). Op deze wijze komen wij op een hoeveelheid rekenwerk van circa 4 * (n + 1) stappen als boven geïntroduceerd. (Door eens en vooral passende schaalfactoren c[i] in te voeren, kan men bereiken, dat het aantal benodigde vermenigvuldigingen slechts met 3 * n oploopt.)

In het geval van sommatie van een Chebyshef-reeks kan men nog zuiniger uit, dank zij het feit, dat de coëfficiënten in de re-

currente betrekking tussen T[k + 1], T[k] en T[k - 1] niet van k afhangen.

De bewering is dat na uitvoering van het rekenschema:

```
b[n + 2]: = b[n + 1]: = 0 ;
for i: = n step -1 until 0 do
b[i]: = 2 * x * b[i + 1] - b[i + 2] + a[i]
```

(8)

geldt dat:

$$(b[0] - b[2]) / 2 = \sum_{i=0}^{n}{}' (a[i] * T[i](x)) ,$$

een zo elegant rekenschema, dat het ons zonder al te veel moeite verzoent met de conventie der halvering van de eerste term, als bij $\sum'$ was afgesproken.

Om dit te bewijzen, redeneren we als volgt. Een bepaalde a[j] (voor $j \leq n$) wordt niet in de berekening meegenomen, zolang nog $i > j$ is. In b[j] levert deze coëfficiënt een bijdrage a[j], in de term b[j - 1] levert deze coëfficiënt een bijdrage 2 * x * a[j]. Hieruit volgt, dat de b's lineair van de a's afhangen, en als we dus stellen, dat

$$b[0] = \sum_{i=0}^{n} (a[i] * Q[i])$$

(waar we niet expliciet hebben aangegeven, dat de Q's van x afhangen), dan volgt onmiddellijk uit het rekenschema, dat de Q's bepaald zijn door

$$Q[0] = 1 \; ; \; Q[1] = 2 * x ;$$

en verder door de recurrente betrekking

$$Q[i + 1] = 2 * x * Q[i] - Q[i - 1] .$$

Dit is dezelfde recurrente betrekking, als waaraan de Chebyshefpolynomen T[i](x) voldoen, het verschil tussen beide is slechts de (tweede) startwaarde (Q[1] = 2 * T[1](x)).

Voorts geldt (ten duidelijkste)

$$b[2] = \sum_{i=2}^{n} (a[i] * Q[i - 2]) ,$$

$$= \sum_{i=1}^{n} (a[i] * Q[i - 2])$$

(dit laatste, omdat blijkens de recurrente betrekking $Q[-1] = 0$ is).

Hieruit volgt, dat - omdat $Q[0] = 1 = T[0](x)$ -

$$(b[0] - b[2]) = a[0] * T[0](x) + \sum_{i=1}^{n} (a[i] * (Q[i] - Q[i-2])) .$$

Rest ons dus slechts te bewijzen, dat

$$Q[i] - Q[i-2] = 2 * T[i](x) . \qquad (9)$$

Omdat $Q[1] - Q[-1] = 2 * x - 0 = 2 * T[1](x)$ en

$$Q[2] - Q[0] = 4 * x\uparrow 2 - 1 - 1 = 2 * T[2](x) ,$$

is aan (9) voor i = 1 en i = 2 voldaan, en omdat linker- en rechterlid van (9) aan dezelfde recurrente betrekking voldoen, is daarmee het gestelde bewezen.

Opm.

Hoewel de b[i] 's uit het rekenschema (8) door voortplanting (cumulatie) van afrondingsfouten danig verstoord kunnen worden, vallen deze geïntroduceerde fouten bij de verschilvorming b[0] - b[2] grotendeels weg.

Afrondingsfouten worden gemaakt in de n + 1 maal herhaalde laatste regel van rekenschema (8); in plaats van het effect van deze afronding te interpreteren als een "bedorven arithmetiek" kunnen we de verkregen b[i] iedere keer beschouwen als het resultaat van feilloze arithmetiek, behalve dat we in de berekening in plaats van de gegeven coëfficiënt a[i] een een beetje andere coëfficiënt a'[i] hebben meegenomen.

Kennelijk geldt $abs(a[i] - a'[i]) < \varepsilon$ als ε de rekennauwkeurigheid voorstelt.

$$(b[0] - b[2]) / 2 = \sum_{i=0}^{n}{}' (a'[i] * T[i](x))$$

en omdat op het gebied [-1,+1] $abs(T[i](x)) \leq 1$ is geldt dus zeker, dat

$$abs(f(x) - (b[0] - b[2]) / 2) < (n + 1) * \varepsilon . \quad \text{QED.}$$

(Met de substitutie $x = \cos(\varphi)$ kunnen we een alternatieve vorm voor de Q's opschrijven, nl.

$$Q[i] = \cos(i * \varphi) + \sin(i * \varphi) / tg(\varphi) .$$

Als $tg(\varphi)$ klein is, kan Q[i] dus groot worden;

$b[0] = \sum_{i=0}^{n} a[i] * Q[i]$ kan bij kleine fluctuaties van de a[i]'s dus drastisch van waarde veranderen.)

-.-.-.-.-

Behalve de continue orthogonaliteitsrelaties (3) voldoen de Chebyshef-polynomen ook aan discrete orthogonaliteitsrelaties.

Als voor $0 \leq i \leq K$ de punten x[i] gegeven zijn door

$$x[i] = \cos(i * \varphi) \qquad \text{met } \varphi = \pi/K$$

dan geldt voor $m \leq K$ en $n \leq K$:

$$\sum_{i=0}^{K}{}'' T[m](x[i]) * T[n](x[i]) = \begin{cases} 0 & \text{als } m \neq n \\ K/2 & \text{als } m = n \neq 0 \text{ of } K \\ K & \text{als } m = n = 0 \text{ of } K \end{cases} \qquad (10)$$

waar met het symbool $\sum''$ aangegeven wordt, dat de eerste èn de laatste term gehalveerd moeten worden.

Het linkerlid van 10 laat zich herschrijven tot

$$\sum_{i=0}^{K}{}''(\cos(m * \varphi) * \cos(n * \varphi)) =$$

$$\tfrac{1}{2} * \sum_{i=0}^{K}{}'' \cos((n + m) * \varphi) + \tfrac{1}{2} * \sum_{i=0}^{K} ((n - m) * \varphi) .$$

Het gestelde volgt onmiddellijk uit het feit dat

$$\sum_{i=0}^{K}{}'' \cos(h * \varphi) = \begin{cases} K \text{ als h een veelvoud van 2K is} \\ 0 \text{ als h niet een veelvoud van 2K is,} \end{cases}$$

hetwelk, zowel voor h even als oneven, op grond van gepaste symmetrie-overwegingen evident is. (Opm. We hebben hier, voor algemene gehele m en n bewezen, dat het linkerlid = K/2 is, als som of verschil van m en n een veelvoud van 2 * K is, en K als aan beide condities voldaan is.

Op grond van de orthogonaliteitsrelaties (10) mogen we verwachten, dat althans voor $m \ll K$

$$A[m] = \frac{2}{K} * \sum_{i=0}^{K}{}''(f(x[i]) * T[m](x[i])) \qquad (11)$$

een goede benadering is voor a[m]. (A[m] = a[m] geldt exact, als f(x) een polynoom van hoogstens de graad K is.)

We kunnen A[m] - die als som, althans numeriek, makkelijker te berekenen is dan de integraaluitdrukking voor a[m] - een willekeurig goede benadering laten zijn voor a[m] door K maar groot genoeg te kiezen. Richtsnoer voor de keuze van K is de volgende uitdrukking voor A[m] in de a's.

$$A[m] = \frac{2}{K} * {\sum_{i=0}^{K}}'' (f(x[i]) * T[m](x[i])) =$$

$$= \frac{2}{K} * \sum_{i=0}^{K}{}'' \left(\sum_{j=0}^{\infty}{}' (a[j] * T[j](x[i]) * T[m](x[i]))\right) =$$

$$= \frac{2}{K} * \sum_{j=0}^{\infty}{}' \left(a[j] * \sum_{i=0}^{K}{}'' (T[j](x[i]) * T[m](x[i]))\right).$$

De binnensommatie is alleen $\neq 0$ voor die waarden van j zodat verschil en/of som van j en m een veelvoud van 2 * K is. Zo komen we tot:

$$A[m] = a[m] + a[2 * K - m] + a[2 * K + m] + a[4 * K - m] + \ldots \text{ etc.}$$

Conclusie is, dat de benadering (11) veilig is, zolang a[2 * K - m] bij a[m] in het niet valt.

Opm.

De coëfficiënt van de Chebyshef-ontwikkeling zijn niet anders dan de Fouriercoëfficiënten van een even periodieke functie met periode 2π. Bovenstaande uitdrukking van de A[m] in de a's is niet anders dan een precisering van de bewering, dat de integraal over een periode van een periodieke functie met periode L passend benaderd wordt door

$$\int_0^L f(x) * dx = (L/K) * \sum_{i=0}^{k-1} f(i * L/K).$$

1.8 Integraalrelatie's van Chebyshefpolynomen

Voor de onbepaalde integralen geldt

$$\int T[r](x) * dx = \int \cos(n * \varphi) * \sin(\varphi) * d\varphi =$$

$$= \tfrac{1}{2} * \int (\sin((n + 1) * \varphi) - \sin((n - 1) * \varphi)) * d\varphi .$$

Hieruit volgt onmiddellijk

$$T[r](x) * dx = \begin{cases} T[1](x) & \text{voor } r = 0 \\ T[2](x) / 4 & \text{voor } r = 1 \\ (T[r+1](x)/(r+1) - T[r-1](x)/(r-1))/2 & \text{voor } r \geq 2. \end{cases}$$

Is nu van de functie f(x) de Chebyshef-ontwikkeling gegeven, nl.

$$f(x) = \sum_{r=0}^{\infty}{}' (a[r] * T[r](x))$$

dan volgt

$$\int f(x)dx = \text{const} + \tfrac{1}{2} * a[0] * T[1] + \tfrac{1}{4} * a[1] * T[2] +$$

$$+ \tfrac{1}{2} * \sum_{r=2}^{\infty} (a[r] * (T[r+1]/(r+1) - T[r-1]/(r-1)))$$

$$= \sum_{r=0}^{\infty}{}' (A[r] * T[r](x))$$

waarbij A[0] door de ondergrens van de integratie gegeven is, terwijl

voor $r \geq 1$ $\quad A[r] = (a[r-1] - a[r+1]) / (2 * r)$.

Op het gebied [-1,+1] was de termsgewijze integratie beslist toelaatbaar; we zien dat door de coëfficiënt "2 * r" in de noemer de reeks voor de integraal wat beter convergeert; dit in overeenstemming met het feit, dat numerieke integratie een proces pleegt te zijn, waarbij men precisie wint.

Het differentiatie-probleem is het omgekeerde, nl. om uit gegeven A's de a's te vinden. Wij kunnen de laatste relatie herschrijven tot

$$a[r-1] = a[r+1] + 2 * r * A[r].$$

Als A[n] de hoogste coëfficiënt is, die niet verwaarloosbaar is, (dwz. we stellen A[r] = 0 voor r > n) dan benaderen we de functie door een n-de graads polynoom, de afgeleide waarvan een polynoom van de graad n-1 is, dwz. a[r] = 0 voor $r \geq n$. We vinden dan het rekenschema:

```
a[n] := a[n + 1] := 0
for r := n step -1 until 1 do
a[r - 1] := a[r + 1] + 2 * r * A[r].
```

De coëfficiënten A[r] worden nu met de factor "2 * r" vermenigvuldigd en de a[r]'s zijn daarom minder goed bepaald. We zien hier het verlies aan precisie, dat onveranderlijk het proces van numerieke differentiatie begeleidt.

Opm.

In gesloten vorm luidt de exacte uitdrukking voor de a's:

$$a[r] = 2 * \sum_{k=0}^{\infty} ((r + 2 * k + 1) * A[r + 2 * k + 1]) .$$

1.9 Approximaties "in de Chebyshefse zin"

Gegeven een continue functie f(x) op het interval [a,b]; zij p(x) een polynoom van hoogstens de n-de graad, dat op dit interval de functie f moet benaderen. Noem de "fout" van de benadering

$$e(x) = p(x) - f(x) .$$

We spreken van benadering in de Chebyshefse zin, als het polynoom zo gekozen wordt, dat de grootheid

$$E = \max_{a \leq x \leq b} (\mathrm{abs}(e(x)))$$

geminimaliseerd wordt. (Het principe van "minimale maximale fout".)

Wij zullen nu bewijzen, dat door genoemde eisen het polynoom p(x) eenduidig bepaald is en dat als f(x) zelf niet een polynoom van graad $\leq n$ is de maximale fout minstens n + 2 maal wordt aangenomen, preciéser: dat er minstens n + 2 punten x[i] zijn zodat

$$a \leq x[0] < x[1] < \ldots\ldots < x[n] < x[n + 1] \leq b ,$$

en voor alle i:

$$e(x[i]) = + E * (-1)\uparrow i$$

of voor alle i

$$e(x[i]) = - E * (-1)\uparrow i \qquad \text{geldt.}$$

In andere woorden: in minstens n + 2 opeenvolgende punten neemt de fout e(x) met alternerend teken zijn extreme waarde aan. Wij noemen dit de "eigenschap van alternerende extremen".

Laat p(x) een polynoom zijn, dat niet voldoet aan de eigenschap der alternerende extremen. Dan kunnen we een polynoom c(x), eveneens hoogstens van de n-de graad, construeren, zodat in alle punten y, waarvoor geldt

$$\text{abs}(e(y)) = \max_{a \leqslant x \leqslant b} (\text{abs}(e(x)))$$

gelden zal dat

$$\text{sgn}(c(y)) = \text{sgn}(e(y)) .$$

Het polynoom c(x) heeft in alle punten, waar e(x) zijn maximale absolute waarde aanneemt, dus hetzelfde teken als e(x) en dit betekent, dat we een zodanige positieve λ kunnen kiezen, dat

$$p(x) - \lambda * c(x)$$

een betere benadering is. Immers, als we langzaam λ van 0 af laten groeien, dan begint de grootheid

$$\max_{a \leqslant x \leqslant b} (\text{abs}(p(x) - \lambda * c(x) - f(x))$$

af te nemen. We laten λ zolang groeien, totdat ergens voor een of andere waarde van x een nieuw maximum opduikt. Hiermee is dus bewezen, dat de beste benadering in de Chebyshefse zin voldoet aan de eigenschap der alternerende extremen.

De eenduidigheid is nu gemakkelijk bewezen; stel, dat er twee polynomen p'(x) en p"(x) zijn (met fouten e'(x) resp. e"(x)), zodat voor beide geldt

$$\max_{a \leqslant x \leqslant b} (\text{abs}(e'(x))) = \max_{a \leqslant x \leqslant b} (\text{abs}(e"(x))) = E .$$

Van het gemiddelde polynoom p"'(x) = (p'(x) + p"(x))/2 geldt kennelijk, dat

$$e"'(x) = (e'(x) + e"(x))/2$$

waaruit volgt, dat

$$\max_{a \leqslant x \leqslant b} (\text{abs}(e"'(x))) \leqslant E ;$$

dat dit maximum $<E$ is, is uitgesloten, want dan waren de polynomen p'(x) en p"(x) geen "beste benaderingen". Hieruit volgt, dat

$$\max_{a \leqslant x \leqslant b} (\text{abs}(e"'(x))) = E ,$$

en dat - blijkens het reeds bewezen gedeelte van deze stelling - abs(e"'(x)) in minstens n + 2 dit maximum aanneemt. In die minstens n + 2 punten moet p'(x) = p"(x) gelden en dus zijn ze identiek gelijk. (Dit zou al volgen, als p'(x) = p"(x) in n + 1 verschillende punten!)

De beste polynoombenadering van de n-de graad in de Chebyshefse zin is dus uniek en voldoet aan de eigenschap der alternerende extremen. Het is dus niet uitgesloten, dat - toevallig - de fout e(x) absoluut zijn extreme waarde vaker dan n + 2 keer aanneemt.

Tot slot bewijzen we de omgekeerde stelling: Als een n-de graads polynoom q(x) voldoet aan de eigenschap der alternerende extremen, dan is q(x) de inderdaad de beste benadering in de Chebyshefse zin.

Immers. Volgens de veronderstelling geldt, dat, als

$$M = \max_{a \leqslant x \leqslant b} (\mathrm{abs}(q(x) - f(x))) ,$$

de functie

$$q(x) - f(x)$$

deze waarde in minstens n + 2 punten met alternerend teken aanneemt.

Zij nu p(x) het volgens de vorige stelling de unieke beste benadering en zij , als boven,

$$E = \max_{a \leqslant x \leqslant b} (\mathrm{abs}(p(x) - f(x))) .$$

Uit het feit, dat p de beste benadering is, volgt dat $M \geqslant E$ moet zijn. We zullen bewijzen, dat M = E is, door te laten zien, dat $M > E$ tot een contradictie aanleiding geeft.

Beschouw dan nl.

$$q(x) - p(x) = (q(x) - f(x)) - (p(x) - f(x)) .$$

In de n + 2 punten, waar q(x) - f(x) zijn extreme waarde $\pm$ M aanneemt, heeft q(x) - p(x) dan hetzelfde teken als q(x) - f(x), dwz. alternerend teken. Daar tussen liggen n + 1 nulpunten, wat voor het n-de graads polynoom q(x) - p(x) zoals bekend te veel is.

Het probleem om voor een gegeven functie op het interval [-1,1] de beste n-de graads benadering te bepalen is over het algemeen vrij lastig. Als de coëfficiënten van de Chebyshefreeks snel convergeren is het in de praktijk meestal voldoende om de Chebyshefreeks na de n-de term af te kappen. Als we de coëfficiënten van de reeks tot onze beschikking hebben, is daarmee het probleem dus min of meer opgelost. Als we die niet tot onze beschikking hebben, maar desalniettemin gronden hebben om aan te nemen, dat ze redelijk snel naar nul zullen gaan, dan weten we dus dat de fout van de benadering in eerste instantie gelijkvormig zal zijn met het n+1-ste Chebyshef-polynoom, maw. de plaatsen, waar het n+1-ste Chebyshefpolynoom zijn extrema aanneemt zijn over het algemeen geen slechte schattingen voor de plaatsen, waar de fout van de n-de graadsbenadering zijn extrema zal aannemen.

Als (de) n+2 punten, waar de fout van de beste benadering zijn extreme waarde aanneemt, bekend zijn, dan is het de beste polynoombenadering in principe als volgt te bepalen. Zij dit de punten x[i].

Men bepaalt dan twee interpolatiepolynomen van de graad n + 1, F en E,

$$F(x[i]) = f(x[i])$$

en

$$E(x[i]) = (-1)\uparrow i$$

en kiest dan het lineair compositum G = F + ε * E zodanig, dat dit een polynoom van de graad n wordt. Dit polynoom maakt in de punten x[i] de fout $\pm$ ε.

Men kiest als punten x[i] aanvankelijk de plaatsen, waar T[n + 1](x) extreem wordt en berekent aldus het polynoom G. Meestal is daarmee de kous af. Het is natuurlijk mogelijk, dat op andere punten de functie G een afwijking van f heeft, die absoluut aanzienlijk groter is dan abs(ε). In dat geval bepaalt men de nieuwe ligging van de extremen van de fout van G en gaat men hiermee het spel herhalen. Dit proces pleegt te convergeren.

2. GAUSS-INTEGRATIE

2.1 Legendre-polynomen

Het Legendre-polynoom P[n](z) is gegeven door

$$P[n](z) = \frac{1}{2\uparrow n * (n!)} * (\frac{d}{dz})^n \, (z^2 - 1)^n \, . \qquad (1)$$

Het is duidelijk, dat p[n], als n-de afgeleide van een polynoom van de graad 2 * n, een n-de graadspolynoom is.

We zullen laten zien, dat deze polynomen onderling orthogonaal zijn op het interval [-1,+1] ten opzichte van een constante gewichtsfunctie w(z) = 1. (Zoals ze boven gedefinieerd zijn, zijn ze niet genormeerd.)

Opm. 1

Aangezien de n-de afgeleide van een even functie dan en slechts dan even is voor even n, volgt dat P[n](x) een even functie is voor n = even en een oneven functie voor n = oneven.

Opm. 2

Als we bedenken, dat $(z^2 - 1) = (z + 1)(z - 1)$, dan is het gemakkelijk in te zien, dat P[n](1) = 1; met de vorige opmerking volgt dan, dat $P[n](-1) = (-1)\uparrow n$.

De orthogonaliteit tonen we als volgt aan. Te bewijzen is, dat

$$(P[m],P[n]) = \int_{-1}^{+1} P[m](z) * P[n](z) * dz = 0 \quad \text{voor } m > n.$$

$$2\uparrow(n + m) * (n!) * (m!) * (P[m],P[n]) =$$

$$\int_{-1}^{+1} (\frac{d}{dz})\uparrow n \, . \, (z^2 - 1)\uparrow n * (\frac{d}{dz})\uparrow m \, . \, (z^2 - 1)\uparrow m * dz =$$

$$\int_{-1}^{+1} (\frac{d}{dz})\uparrow n \, . \, (z^2 - 1)\uparrow n * d((\frac{d}{dz})\uparrow(m - 1) \, . \, (z^2 - 1)\uparrow m) =$$

$$-\int_{-1}^{+1} (\frac{d}{dz})\uparrow(n + 1) \, . \, (z^2 - 1)\uparrow n * (\frac{d}{dz})\uparrow(m - 1) \, . \, (z^2 - 1)\uparrow m * dz$$

(omdat $(z^2 - 1)\uparrow m$ in +1 en -1 een m-voudig nulpunt heeft, zijn de eerste m - 1 afgeleiden in de eindpunten = 0 en valt de stokterm nu - en straks - weg.)

Dit proces herhalen we nu nog een keer, en we vinden

$$(-1)\uparrow(n+1) * \int_{-1}^{+1} (\frac{d}{dz})\uparrow(2*n+1).(z^2-1)\uparrow n * (\frac{d}{dz})\uparrow(m-n-1).(z^2-1)\uparrow m * dz.$$

De eerste factor van de integrand is echter identiek = 0, zodat de hele integraal $\neq 0$ is, waarmee de orthogonaliteit bewezen is.

We weten dus dat alle nulpunten van de polynomen P[n] reëel en enkelvoudig zijn, in het inwendige van het gebied [-1,+1] liggen, "elkaar omvatten" en tenslotte, dat drie opeenvolgende polynomen aan een lineaire recurrente betrekking voldoen. Omdat de polynomen P[n] even/oneven zijn als n = even/oneven, is een van de coëfficiënten = 0 en kunnen we stellen

$$P[n](z) = (a[n] * z) * P[n-1](z) - b[n] * P[n-2](z). \qquad (2)$$

De vraag is, of wij deze coëfficiënten kunnen bepalen. De coëfficiënten a[n] is zeer gemakkelijk te bepalen als we bedenken dat

$$a[n] = C[n] \,/\, C[n-1]$$

als C[i] de coëfficiënt van de hoogste macht van z is P[i](z). Deze is ontstaan door i-voudige differentiatie van de hoogste macht van z in $(z^2-1)\uparrow i$, dwz. van $z\uparrow(2*i)$. Dit levert, met de constante factor in (1) op

$$C[i] = \frac{1}{2\uparrow i*(i!)} \quad \frac{(2*i)!}{i!}$$

zodat

$$a[n] = C[n] \,/\, C[n-1] =$$

$$= \frac{(2*n)!}{2\uparrow n*(n!)\uparrow 2} * \frac{2\uparrow(n-1)*((n-1)!)\uparrow 2}{(2*n-2)!}$$

$$= \frac{(2*n)*(2*n-1)}{2*n\uparrow 2} = \frac{2*n-1}{n},$$

waarmee een van beide coëfficiënten bepaald is. De andere volgt dan onmiddellijk, als we in (2) z = 1 substitueren en bedenken, dat P[i](1) = 1 voor alle i.

Zo vinden we

$$n * P[n](z) = (2*n-1) * z * P[n-1](z) - (n-1) * P[n-2](z) \qquad (3)$$

Opm.

(Uit P[0](z) = 1 en P[1](z) = z en boven gegeven recurrente be-

trekking volgt, dat de polynomen (n!) * P[n](z) gehele coëfficiënten hebben.)

2.2 Integratieformules van Gauss

In het volgende zullen we ons beperken tot benadering van

$$\int_{-1}^{+1} f(x) * dx \ .$$

Als oorspronkelijk gevraagd is, om de integraal van a tot b uit te rekenen, dan kunnen we die met een eenvoudige lineaire transformatie van de integratievariabele tot het standaardgeval herleiden.

Wij stellen ons voor de integraal te benaderen door een lineair compositum van integrandwaarden, dwz. we schrijven

$$\int_{-1}^{+1} f(x) * dx = \sum_{k=1}^{n} (c[k] * f(x[k])) + R \ . \qquad (1)$$

Hier is R de restterm en ons streven zal zijn, om R voor "nette functies" zo klein mogelijk te houden. De punten x[k] - de enige punten, waarin de integrand uitgerekend wordt - heten de "steunpunten", de bijbehorende c[k] de "gewichten". Steunpunten en gewichten zijn onafhankelijk van de integrand.

Stel dat bij zekere keuze van c[k] en x[k] ($1 \leqslant k \leqslant n$) in (1) de restterm $R = 0$ is, als we voor f een willekeurig polynoom van graad m (of lager) kiezen, maar $R \neq 0$ is voor sommige polynomen van de graad $m + 1$, dan zeggen we dat deze integratiemethode de precisiegraad m heeft. We interpreteren de bovengestelde eis dat R voor nette functies klein is, nu (enigszins arbitrair) door te eisen: bepaal de steunpunten x[k] en de gewichten c[k] zodanig, dat de precisiegraad van (1) maximaal is.

De precisiegraad $m = n + 1$ halen we beslist, want die halen we zelfs als de ligging van de punten x[k] voorgeschreven is. Door deze n punten leggen we het interpolatiepolynoom van Lagrange

$$p(x) = \sum_{k=1}^{n} f(x[k]) * L[k](x)$$

waarin

$$L[k](x) = \prod_{\substack{i=1 \\ i \neq k}}^{n} \left(\frac{x - x[i]}{x[k] - x[i]}\right) \qquad (2)$$

is. Uit deze definitie volgt, dat $L[k](x[i]) = \delta[k,i]$,

zodat p(x[k]) = f(x[k]). Als f nu een polynoom van hoogstens de graad n - 1 is, dan is p(x) ≡ f(x) en we halen de precisiegraad n - 1 door te kiezen

$$c[k] = \int_{-1}^{+1} L[k](x) * dx .$$

Een precisiegraad m = 2 * n is <u>niet</u> haalbaar.
Kiezen we nl. voor

$$f(x) = Pzn(x) = \prod_{i=1}^{n} (x - x[i])\uparrow 2 ,$$

dan is Pzn(x[i]) = 0 en overal elders is Pzn(x) > 0, zodat toepassing van (1) geeft

$$R = R + \sum_{i=1}^{n} (Pzn(x[i]) * c[k]) = \int_{-1}^{+1} Pzn(x)\, dx > 0 .$$

We zullen nu bewijzen, dat de precisiegraad m = 2 * n - 1 wel haalbaar is. Een dergelijke integratieformule heet "een integratieformule van Gauss".

Bij de rij, vooralsnog onbekende, steunpunten x[k] voeren we in een polynoom van de graad n, nl.

$$pn(x) = \prod_{k=1}^{n} (x - x[k]) .$$

Zij nu f een polynoom van de graad ⩽ 2 * n - 1; dan is het verschilpolynoom

$$v(x) = f(x) - \sum_{k=1}^{n} (f(x[k]) * L[k](x))$$

eveneens hoogstens van de graad 2 * n - 1. Omdat v(x[i]) = 0 voor alle 1 ⩽ i ⩽ n bevat v(x) het polynoom pn(x) als factor en kunnen we v(x) schrijven als

$$v(x) = g(x) * pn(x)$$

waarbij g(x) een polynoom met graad ⩽ n - 1 is.
We kunnen dus f(x) schrijven als

$$f(x) = \sum_{k=1}^{n} (f(x[k] * L[k](x)) + g(x) * pn(x).$$

Substitueren we dit in (1) dan vinden we voor R de gedaante

$$R = \sum_{k=1}^{n} (f(x[k])(\int_{-1}^{+1} L[k](x) * dx - c[k])) + \int_{-1}^{+1} g(x) * pn(x) * dx .$$

Onder de bijvoorwaarde dat $g(x) = 0$ zijn de waarden $f(x[k])$ nog vrij kiesbaar; onder die voorwaarden (f is dan algemeen een polynoom van de graad n - 1) moet R beslist = 0 zijn, zodat

$$c[k] = \int_{-1}^{+1} L[k](x) * dx . \qquad (3)$$

Met andere woorden: als de x[k]'s bepaald zijn, dan volgen daaruit de c[k]. Uit (3) volgt dat

$$R = \int_{-1}^{+1} g(x) * pn(x) * dx$$

en we zien, dat (1) de precisiegraad dan en slechts dan de precisiegraad $2 * n - 1$ heeft, als de punten x[k] zo gekozen zijn, dat

$$\int_{-1}^{+1} g(x) * pn(x) * dx = 0 \qquad (4)$$

voor elk willekeurig polynoom $g(x)$ van graad $\leq n - 1$.

Maar we kennen het n-de graadspolynoom, dat orthogonaal staat op alle polynomen van lagere graad, nl. het polynoom van Legendre. De conclusie is, dat de precisiegraad $m = 2 * n - 1$ haalbaar is, mits we als steunpunten x[k] de nulpunten van het Legendrepolynoom P[n](x) kiezen. Het is in dit verband een geruststelling bewezen te hebben, dat deze n nulpunten reëel en verschillend zijn en alle liggen in het interval $[-1,+1]$.

Opm.

De bijbehorende coëfficiënten c[k], zoals ze door (3) gegeven zijn, zijn alle positief. Dit bewijzen we met een trucje. Beschouw voor vaste k het polynoom van de graad $2 * n - 2$:

$$L[k](x) * (L[k](x) - 1) .$$

Omdat $L[k](x[i]) = \delta[i,j]$ is dit polynoom in alle punten x[i] gelijk aan nul (voor $i \neq k$ wegens de eerste, voor $i = k$ wegens de laatste factor.) We kunnen dus schrijven

$$L[k](x) * (L[k](x) - 1) = g(x) * pn(x)$$

waar $g(x)$ een polynoom van de graad $n - 2$ is.

Volgens (4) geldt dus

$$\int_{-1}^{+1} L[k](x) * (L[k](x) - 1) * dx = \int_{-1}^{+1} g(x) * pn(x) * dx = 0$$

zodat

$$\int_{-1}^{+1} L[k](x) * dx = \int_{-1}^{+1} (L[k](x))\uparrow 2 * dx .$$

Het linkerlid is echter = c[k], terwijl het rechterlid > 0 is, waarmee het gestelde bewezen is.

Het feit, dat alle c's positief zijn, is heel belangrijk. Het stelt ons nl. in staat te bewijzen, dat we in het geval van een continue integrand, een willekeurig goede benadering van de integraal kunnen krijgen door maar genoeg punten te nemen. Preciéser zij S[n](f) de benadering voor

$$\int_{-1}^{+1} f(x) * dx$$

die men krijgt door gebruik te maken van de n-punts integratieformule van Gauss.

Dan geldt voor iedere continue functie f

$$\lim_{n \to \infty} S[n](f) = \int_{-1}^{+1} f(x)\,dx .$$

Bewijs

Bij elke $\varepsilon > 0$ kan volgens de approximatiestelling van Weierstrass bij een continue functie f een polynoom p gevonden worden, zodat

$$\max_{-1 \leqslant x \leqslant 1} (abs(f(x) - p(x))) < \varepsilon .$$

Stel, dat p de graad N(ε) heeft. Dan geldt exact

$$S[n](p) = \int_{-1}^{+1} p(x)\,dx$$

mits $N(\varepsilon) \leqslant 2 * n - 1$.

Verder is

$$abs(S[n](f)-S[n](p)) = abs(\sum_{i=1}^{n} (c[i]*(f(x[i])- p(x[i])))) \leqslant$$

$$\leqslant \sum_{i=1}^{n} \text{abs}(c[i] * (f(x[i]) - p(x[i]))) \leqslant \varepsilon * \sum_{i=1}^{n} \text{abs}(c[i]) =$$

$$= \varepsilon * \sum_{i=1}^{n} c[i] = 2 * \varepsilon .$$

(Opm. Het eerste ongelijkteken rust op het feit, dat de absolute waarde van een som niet groter kan zijn dan de som van de absolute waarde van de termen. Bij het tweede ongelijkteken is een beroep gedaan op Weierstrass; bij het volgende gelijkteken is een beroep gedaan op het feit, dat alle $c[i] > 0$ zijn. Het laatste gelijkteken berust op de relatie

$$\sum_{i=1}^{n} c[i] = 2$$

welke onmiddellijk volgt, als we bedenken, dat de integratieformule van Gauss exact moet zijn voor het 0-de graadspolynoom $f(x) = 1$.)

Anderzijds geldt

$$\text{abs}(\int_{-1}^{+1} f(x) * dx - S[n](p)) =$$

$$= \text{abs}(\int_{-1}^{+1} f(x) * dx - \int_{-1}^{+1} p(x) * dx) =$$

$$= \text{abs}(\int_{-1}^{+1} (f(x) - p(x)) * dx) \leqslant \int_{-1}^{+1} \varepsilon * dx = 2 * \varepsilon$$

Combinatie van deze twee resultaten geeft ons

$$\text{abs}(\int_{-1}^{+1} f(x) * dx - S[n](f)) \leqslant 4 * \varepsilon \ ;$$

aangezien ε willekeurig klein gekozen kan worden, is hiermee het gestelde bewezen.

Opm.

Het positief zijn van de $c[i]$'s en het feit, dat $\sum c[i]$ bekend is, garandeert, dat ook $\sum \text{abs}(c[i])$ begrensd is. Omgekeerd geldt, dat als ook negatieve $c[i]$'s voor mogen komen, - wat bij sommige andere integratieformules het geval is - dat dan $\sum \text{abs}(c[i])$ in-

derdaad niet meer begrensd hoeft te zijn. Dergelijke formules hebben het nadeel, dat men niet gegarandeerd een betere benadering hoeft te krijgen, als men meer punten neemt.

2.3 Een kwalitatieve schatting van de restterm

We zullen niet een algemene schatting van de restterm geven. In plaats daarvan zullen we laten zien hoe de restterm, in het simpelste geval, dat deze niet nul is, van de interval-lengte afhangt.

We beschouwen daartoe een willekeurige integratie-formule met precisiegraad m = M - 1 (de volgende beschouwing is niet gebonden aan Gauss-integratie) voor het standaard-interval [-1,+1] en bekijken de fout, die gemaakt wordt als we met deze integratieformule x↑M integreren, dwz. het eerste polynoom, dat door deze formule niet meer exact geïntegreerd wordt.

Als ℓ gedefinieerd is als $\ell = (b - a) / 2$, dan wordt de integraal

$$\int_a^b f(x) * dx$$

door de substitutie $y = (x - a - \ell) / \ell$, dwz.

$$x = a + \ell + \ell * y$$

herleid tot een integraal op het standaardinterval [-1,+1] en wel met

$$g(y) = f(a + \ell + \ell * y) / \ell\uparrow M$$

$$\ell\uparrow(M + 1) * \int_{-1}^{+1} g(y) * dy .$$

De definitie van g(y) is van een extra factor $\ell\uparrow(-M)$ voorzien, om ervoor te zorgen, dat als f(x) = x↑M, dat dan g(y), dat dan ook een M-de graads polynoom is, ook een 1 als coëfficiënt van de hoogste macht heeft.

Als wij nu de integraal over y met een formule van de precisiegraad m = M - 1 benaderen, dan wordt de fout alleen bepaald door de fout, die gemaakt wordt bij de integratie van

$$\int_{-1}^{+1} y\uparrow m \, dy .$$

Noem deze fout c; dan volgt hieruit, dat de fout, die gemaakt wordt bij integratie van a naar b van een M-de graads polynoom met hoogste macht coëfficiënt = 1 met $\ell = (b - a) / 2$ gelijk is aan

$$c * \ell\uparrow(M + 1) .$$

Omdat we ons hier toch niet in de grootte van c zullen verdiepen, kunnen we ook de restrictie van de gegeven coëfficiënt van de M-de macht laten varen, en concluderen dat de fout, die bij integratie van een M-de machts polynoom met behulp van een formule met precisiegraad m = M - 1 gemaakt wordt, evenredig is met de (M + 1)-ste macht van de integratieweg, ongeacht zijn ligging!

Hiermee is ons een strategie gegeven om de integraal

$$\int_a^b f(x)\,dx$$

te berekenen met behulp van bv. de 3-punt Gauss-formule. Deze heeft een precisiegraad 5. Als wij nu de Gauss-formule een keer over het hele gebied toepassen, en vervolgens een keer op het gebied van $[a, (a + b)/2]$ en het gebied $[(a + b)/2, b]$, dan weten we, dat de fout in de som der beide laatste integralen in eerste benadering $2 * 2\uparrow(-7) = 2\uparrow(-6)$ maal de fout in de eerste benadering is. Op deze manier krijgen we althans een indruk van de orde van grootte van de gemaakte fouten. Dergelijke overwegingen kunnen ook gebruikt worden om de onderverdeling van het interval aan de vereiste precisie aan te passen; wij zullen hier later op terugkomen.

3. VOORTGEZETTE HALVERING

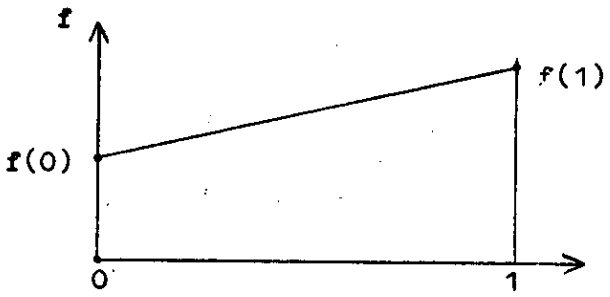
In het volgende zullen we ons, om schrijfwerk te besparen, beperken tot de benadering van de integraal

$$\int_0^1 f(x)\,dx.$$

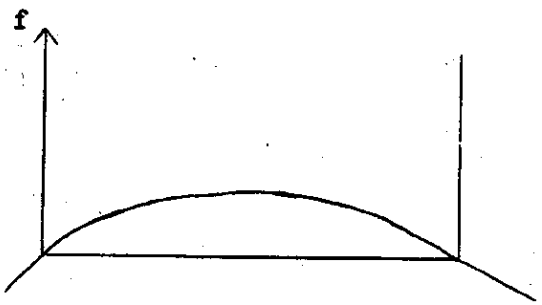
De grofste benadering vinden we als we op het hele interval de zg. "trapeziumregel" toepassen,

$$T[1] = (f(0) + f(1)) / 2$$

dwz. (de lengte van het interval maal) het gemiddelde van de waarden in begin en eindpunt. De precisiegraad van deze integratieformule is = 1, de formule is exact, als f een eerstegraadspolynoom is.



(Opm. De precisiegraad kan niet 2 zijn, want als f een parabool is



van deze gedaante, dan is T[1] = 0 terwijl de integraal dat pertinent niet is.)

Een volgende stap in de benadering krijgen we, wanneer we het interval in twee gelijke helften delen, en op beide helften de trapezium-regel toepassen.

Deze deelbijdragen noteren we als t(1/4) en t(3/4) omdat 1/4 resp. 3/4 de middelpunten van de gebieden zijn, waarover we de trapeziumregel toepassen (in deze notatie is dus T[1] = t(1/2)), dan vinden we

$$t(1/4) = (f(0) + f(1/2))/4$$

en

$$t(3/4) = (f(1/2) + f(1))/4 .$$

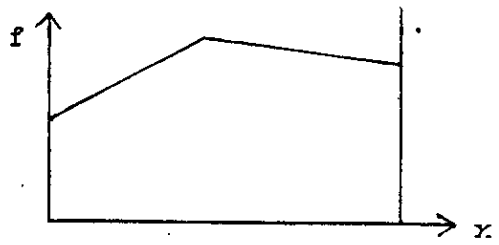
Onze nieuwe benadering voor de gehele integraal is dus de som van deze twee, dwz.

$$T[2] = t(1/4) + t(3/4) =$$

$$f(0)/4 + f(1/2)/2 + f(1)/4$$

$$= (T[1] + f(1/2))/2 .$$

We vinden T[2] dus als het gemiddelde van T[1] en de functiewaarde in x = 1/2; onze benadering is nu exact, wanneer f(x) door twee rechte lijnen voorgesteld kan worden, bv.



Wij kunnen dit herhalen, door het totale interval in vieren te verdelen, dwz. de boven beschreven halveringsmethode op beide helften toe te passen. We vinden dan

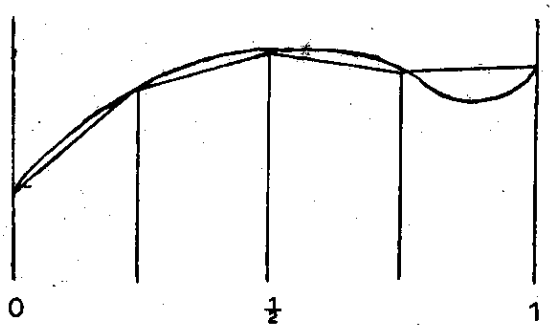
$$T[3] = \left(T[2] + \frac{f(1/4)+f(3/4)}{2}\right)/2 .$$

De volgende benadering in deze reeks is

$$T[4] = \left(T[3] + \frac{f(1/8)+f(3/8)+f(5/8)+f(7/8)}{4}\right)/2$$

etc. Men vindt T[i + 1] uit T[i] door

a) de functie f te berekenen in alle punten, die midden tussen twee opeenvolgende steunpunten van T[i] liggen;
b) het gemiddelde t van deze functiewaarden te nemen;
c) het gemiddelde van t en T[i] te nemen.



T[3] is dus de integraal die we krijgen als we de functie benaderen door een koordenpolynoom. Als de functie f(x) continu is, is f(x) gelijkmatig continu en is het gemakkelijk om te bewijzen, dat

$$\lim_{n \to \infty} T[n] = \int_0^1 f(x) * dx .$$

Alle T[i] zijn natuurlijk exact gelijk aan de gezochte integraal, wanneer f een lineaire functie is. Als wij zouden weten, dat f een 2-de graads-polynoom is, dan kunnen wij ons de vraag stellen, of we de limiet van deze reeks op grond van zijn convergentiegedrag niet kunnen extrapoleren. Dit kan inderdaad. Schrijven we

$$T[i] = \int_0^1 f(x)\,dx + R[i]$$

dan weten we dat op grond van het feit, dat de precisiegraad van T[i] gelijk m = 1 is en de beschouwingen van § 2.3, dat

$$R[i + 1] = R[i] \;/\; 2\uparrow(m + 1) = R[i] \;/\; 4$$

is.

Als we deze relatie gebruiken, om tussen

$$T[i] = \int_0^1 f(x) * dx + R[i]$$

en

$$T[i + 1] = \int_0^1 f(x) * dx + R[i + 1]$$

de restterm te elimineren, vinden we als uitdrukking voor de integraal

$$\int_0^1 f(x) * dx = \frac{4 * T[i+1] - T[i]}{3} .$$

Dit zou exact zijn, als f(x) een tweedegraads polynoom was, als we dit niet weten, dan kunnen we deze lineaire composita slechts als benaderingen voor de integraal beschouwen, en we schrijven dus

$$S[i] = \frac{4 * T[i+1] - T[i]}{3} \qquad \text{voor } i = 1,2,3,4\ldots..$$

De benaderingen S[i], S[2], S[3] ..... heten de benaderingen volgens Simpson.

Opm.

Als wij S[1] = (4 * T[2] - T[1])/3 in functiewaarden uitschrijven, dan vinden we

$$S[i] = (f(0) + 4 * f(1/2) + f(1))/6 .$$

In deze vorm is "de regel van Simpson" het meest bekend. De waarde S[i] kan worden uitgelegd als de integraal van de parabool door de drie steunpunten f(0), f(1/2) en f(1); S[2] is de waarde van de integraal van de benadering, die men krijgt door beide helften van het integratiegebied separaat door een parabool te benaderen, etc.

De regel van Simpson is niet alleen exact voor polynomen van de 2-de graad, maar ook voor polynomen van de 3-de graad. Dit volgt uit de algemene stelling dat een integratieformule waarbij

1) de steunpunten symmetrisch ten opzichte van het midden van het integratie-interval liggen, en
2) de functiewaarden van aldus aan elkaar toegevoegde steunpunten met hetzelfde gewicht in rekening gebracht worden

nooit een even precisiegraad kan hebben.

Dit bewijst men, door te laten zien, dat, als een polynoom van de graad 2 * m exact geïntegreerd wordt, dan ook een polynoom van de graad 2 * m + 1 exact geïntegreerd wordt.

Zij d = (a + b)/2 het midden van het integratie-interval.

Kiezen we voor f een polynoom van graad 2 * m + 1 dan kan dit op één manier geschreven worden als

$$f(x) = C * (x - d)\uparrow(2 * m + 1) + P2m(x)$$

waar P2m(x) een willekeurig polynoom van de graad $\leq$ 2 * m is.

Dan volgt, dat

$$\int_a^b f(x) * dx = \int_a^b P2m(x) * dx .$$

De integratieformule levert echter

$$\sum(c[k] * f(x[k])) =$$
$$\sum(c[k] * P2m(x[k])) + C * \sum(c[k] * (x[k] - d)\uparrow(2 * m - 1)) .$$

De eerste som levert volgens de veronderstelling $\int_a^b P2m(x) * dx$,

maar dat is gelijk aan $\int_a^b f(x) * dx$; de tweede som is altijd = 0,

waarmee het gestelde bewezen is. (De tweede som is nl. = 0 op grond van de volgende overweging: als er een x[k] = d voorkomt, dan is de bijbehorende term = 0 vanwege de factor x[k] - d; voor elke x[k] ≠ d bestaat een x[k'] zodat

$$x[k] - d = -(x[k'] - d)$$

en

$$c[k] = c[k'] .$$

Omdat de exponent 2 * m + 1 oneven is, vallen deze twee termen tegen elkaar weg.)

De constructie van de T's voldoen kennelijk aan de genoemde symmetrie-eisen, de precisiegraad van de T's is dan ook oneven. (nl. = 1) Bij de overgang van de T's naar de S'en zijn de individuele steunpunten niet meer ter sprake geweest en de S'en voldoen dus ook aan genoemde symmetrie-eisen. Het is bovendien gemakkelijk in te zien, dat de precisiegraad van de formule van Simpson geen 4 kan zijn, waarmee het gestelde bewezen is.

Met het gegeven, dat de integratieformules S van de precisiegraad m = 3 zijn, kunnen we analoog aan de stap van T naar S nu van S naar nieuwe integratieformules C overgaan gegeven door

$$C[i] = (4\uparrow 2 * S[i + 1] - S[i]) / (4\uparrow 2 - 1)$$

die voor 4-de graads - en dus ook voor 5-de graads polynomen goed zijn; daaruit construeren we de D's gegeven door

$$D[i] = (4\uparrow 3 * C[i + 1] - C[i]) / (4\uparrow 3 - 1)$$

goed voor 6-de en dus 7-de graads polynomen, en daaruit

$$E[i] = (4\uparrow 4 * D[i + 1] - D[i]) / (4\uparrow 4 - 1)$$

etc. en we krijgen op deze manier een heel schema

T[1]				
T[2]	S[1]			
T[3]	S[2]	C[1]		
T[4]	S[3]	C[2]	D[1]	
T[5]	S[4]	C[3]	D[2]	E[1]
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
precisiegraad: m=1	m=3	m=5	m=7	m=9

We hebben bewezen, dat als we een stap naar rechts gaan, de precisiegraad minstens 1, en op symmetrie-overwegingen dus minstens 2 moet toenemen; we hebben niet bewezen - hoewel het wel zo is - dat de precisiegraad ook niet meer dan 2 kan toenemen.

We hebben gezien, dat bij continue functie f(x)

$$\lim_{n\to\infty} T[n] = F = \int_0^1 f(x) * dx$$

is.
Omdat S[n] het gewogen gemiddelde van T[n] en T[n + 1] is - niet gewichten die onafhankelijk van n zijn - geldt dus dat eveneens

$$\lim_{n\to\infty} S[n] = F.$$

Op dezelfde manier volgt, dat $F = \lim_{n\to\infty} C[n] = \lim_{n\to\infty} D[n] = \lim_{n\to\infty} E[n]$.

In het bovenstaande driehoekige schema zijn alle benaderingen op eenzelfde regel afgeleid van dezelfde steunpunten. In de praktijk zal men dit schema niet voortzetten totdat in de rechter onderhoek de getallen in het gewenste aantal cijfers met elkaar overeenstemmen.

Alle T[i], S[i], C[i], D[i], E[i] etc. zijn lineaire composita van functiewaarden in de steunpunten: zij zijn alle te schrijven in de vorm

$$\sum c[k] * f(x[k]) .$$

Het kan aangetoond worden, dat alle $c[k] \geq 0$ zijn, een conclusie, die ons niet zo hoeft te verbazen, gezien het feit, dat zowel naar onderen als naar rechts de benaderingen in het driehoekige schema allen de integraalwaarde F als limiet hebben.

Tenslotte willen we van de reeks T[1], S[1], C[1] etc. bekijken de precisiegraad m, het aantal steunpunten N en de graad Ni het interpolatiepolynoom door functiewaarden in de steunpunten.

formule	m	N	Ni
T[1]	1	2	1
S[1]	3	3	2
C[1]	5	5	4
D[1]	7	9	8
E[1]	9	17	16
.	.	.	.
.	.	.	.
.	.	.	.

Wij hebben de graad Ni van het interpolerend polynoom erbij geschreven om onze integratieformule te kunnen vergelijken met die van Cotes. Een zg. integratieformule van Cotes heeft betrekking op Ni + 1 aequidistante steunpunten en levert de integraal van het interpolerend polynoom van de graad Ni. (Men leidt bv. de bijbehorende gewichten af, door te eisen dat de polynomen $1,x,x\uparrow 2,\ldots,x\uparrow(Ni)$ alle exact geintegreerd worden; men vindt op deze wijze de Ni + 1 lineaire vergelijkingen voor de Ni + 1 onbekende gewichten.)

T[1], S[1] resp. C[1] zijn identiek aan de 2-punts, 3-punts, resp. 5-punts formule van Cotes. Vanaf D[1] is dit niet meer het geval: hoewel D[1] van 9 steunpunten gebruik maakt, is hij slechts goed voor een 7-graads polynoom en dus niet meer voor het 8-ste graads interpolatiepolynoom van Lagrange. Dat hier de wegen scheiden is maar goed ook, want als we de coëfficiënten van de integratieformules van Cotes berekenen, dan zullen we merken dat bij de 9-punts formule voor het eerst negatieve gewichten gaan optreden.

Opm.

De formules S[1] en C[1] - dus de Cotes-integratieformules voor 3- en 5 punten - zijn erg geliefd, omdat wij zien, dat de precisiegraad van deze integratieformules 1 hoger is dan we op grond van het aantal steunpunten zouden mogen verwachten. Speciaal de formule van Simpson mag zich in grote sympathie verheugen, en wel om twee redenen, ten eerste omdat de winst van de extra graad hier gemarkeerder is, ten tweede, omdat de schatting voor de restterm voor het geval, dat de functie niet een polynoom is, uitgedrukt wordt in een lagere afgeleide dan in het geval van de vijf-punts formule. Om met vertrouwen van de formule van Simpson gebruik te kunnen maken, hoeft men dus minder drastische veronderstellingen over de differentieerbaarheid van de functie te maken.

4. FORMELE MACHTREEKSEN

Als wij definiëren (voor $n \geq 0$)

$$K[n] = \int_0^1 t\uparrow n * \exp((1 - t) * \theta) * dt$$

dan vinden we door partiële integratie

$$K[n] = \frac{1}{n+1} \int_0^1 \exp((1 - t) * \theta) * d(t\uparrow(n + 1))$$

$$= \frac{1}{n+1} * \exp((1 - t) * \theta) * t\uparrow(n + 1) \Big/_0^1 + \frac{\theta}{n+1} * K[n + 1]$$

$$= \frac{1}{n+1} + \frac{\theta}{n+1} * K[n + 1] ,$$

dwz. een recurrente betrekking voor de K's.

Anderzijds geldt

$$K[0] = \int_0^1 \exp((1 - t) * \theta) * dt =$$

$$- \theta\uparrow(-1) * \exp((1 - t) * \theta) \Big/_0^1$$

$$= \theta\uparrow(-1) * (e\uparrow\theta - 1)$$

dus geldt

$$e\uparrow\theta = 1 + \theta * K[0] \qquad (4.0)$$

$$= 1 + \theta + \theta\uparrow 2 * K[1]$$

$$= 1 + \theta + \tfrac{1}{2}\theta\uparrow 2 + \tfrac{1}{2}\theta\uparrow 3 * K[2] \quad \text{etc.}$$

algemeen:

$$e\uparrow\theta = \sum_{i=0}^{m} (\theta\uparrow i/(i!)) + \frac{1}{i!} * \int_0^1 t\uparrow m * e\uparrow((1 - t) * \theta) * \theta\uparrow(m + 1) * dt . \qquad (4.1)$$

Passen wij deze formule (4.1) toe voor $e\uparrow\theta$ en $e\uparrow(-\theta)$, met $m = 2$ dan vinden we, als we schrijven

$$e\uparrow\theta = E \qquad (4.2)$$

na optelling

$$E + E\uparrow(-1) = 2 + \theta\uparrow 2 + \frac{1}{2!} * \int_0^1 t\uparrow 2 * (E\uparrow(1-t) - E\uparrow(t-1)) * \theta\uparrow 3 * dt \; . \qquad (4.3)$$

Passen we formule 4.1 toe met m = 3, voor $e\uparrow\theta$ en $e\uparrow(-\theta)$, dan vinden we na aftrekking

$$E - E\uparrow(-1) = 2 * \theta + \theta\uparrow 3/3 + 1/6 * \int_0^1 t\uparrow 3 * (E\uparrow(1-t) - E\uparrow(t-1)) * \theta\uparrow 4 * dt \; .$$

Als wij hier een factor θ **rechts** buiten haakjes halen en in de 2-de term voor $\theta\uparrow 2$ substitueren wat we uit (4.3) daarvoor oplossen, dan vinden we

$$E - E\uparrow(-1) =$$

$$(E/3 + 4/3 + E\uparrow(-1)/3 + 1/6 * \int_0^1 t\uparrow 2 * (t-1) * (E\uparrow(1-t) - E\uparrow(t-1)) * \theta\uparrow 3 * dt) * \theta \qquad (4.4)$$

Tot zover hebben we niets formeels gedaan: dit komt nu, doordat we E en θ (en daarmee linker en rechterlid van 4.4) als operatoren gaan opvatten.

We stellen - bij vaste h - dat

$$\theta = h * \frac{d}{dx} \; . \qquad (4.5)$$

De enige zinnige interpretatie, die we dan aan $E = e\uparrow\theta$ geven kunnen is de verplaatsingsoperator - immers $E\uparrow 2$ is dezelfde functie van $2 * h$ als E van h is, maw.

$$\theta f(x) = h * f'(x) \quad \text{en} \quad Ef(x) = f(x + h) \; .$$

Dat deze interpretatie consequent is, zien we als we op deze manier de eerste regel van 4.0 proberen te lezen. De operator K[0] is dan gedefinieerd door

$$\int_0^1 E\uparrow(1-t) * dt \; ;$$

we gaan nu beide zijden van de eerste regel van 4.0 laten opereren op de functie f(x).

De linkerkant levert op E.f(x) = f(x + h). De rechterkant levert ons,

wanneer we de integratie over t tot het laatst uitstellen:

$$f(x) + h * \int_0^1 E\uparrow(1 - t) . \theta . f(x) * dt =$$

$$f(x) + h * \int_0^1 E\uparrow(1 - t) . f'(x) * dt =$$

$$f(x) + h * \int_0^1 f'(x + (1 - t) * h) * dt =$$

$$f(x) + \int_0^h f'(x + u) * du = f(x + h)$$

zodat inderdaad aldus opgevat beide zijden aan elkaar gelijk zijn.

Met deze ervaring gewapend gaan we nu beide zijden van 4.4 loslaten op de functie F(x), waarbij we

$$F'(x) = f(x) \quad \text{of} \quad \theta . F(x) = h * f(x)$$

noemen. (Met andere woorden $F(x) = \int f(x) . dx$.)

De linkerkant is eenvoudig: deze levert

$$(E - E\uparrow(-1)) . F(x) = F(x + h) - F(x - h) = \int_{x-h}^{x+h} f(x) * dx .$$

In de rechterkant beginnen we de factor θ, die buiten haakjes is gehaald te laten opereren op F(x) omdat $\theta.F(x) = h * f(x)$ wordt de rechterkant dus een uitdrukking in f en zijn afgeleiden, nl.

$$(h/3) * (f(x - h) + 4 * f(x) + f(x + h)) + R \qquad (4.6)$$

waarin de restterm R gegeven is door

$$R = h\uparrow 4/6 * \int_0^1 t\uparrow 2 * (t - 1) * \{f'''(x + (1 - t) * h) - f'''(x - (1 - t) * h)\} dt .$$

Nu is de uitdrukking tussen accolades gelijk aan

$$2 * h * (1 - t) * f^{IV}(x + \xi * h)$$

voor zekere ξ waarvoor geldt $abs(\xi) < 1 - t$. Hieruit volgt, dat voor zekere η voldoend aan $abs(\eta) < 1$ geldt

$$R = -f^{IV}(x + \eta * h) * h\uparrow 5/3 * \int_0^1 t\uparrow 2 * (t - 1)\uparrow 2 * dt$$

$$= -f^{IV}(x + \eta * h) * h\uparrow 5/3 * (1/5 - 2/4 + 1/3)$$

$$= -f^{IV}(x + \eta * h) * h\uparrow 5/90 . \qquad (4.7)$$

De kopterm van 4.6 geeft ons de integratieformule van Simpson (over een traject 2 * h), 4.7 geeft ons de restterm uitgedrukt in h en de vierde afgeleide van f ergens in het interval. Hieruit volgt, dat het niet zonder meer veilig is de regel van Simpson te gebruiken, zelfs niet met een heel klein interval, wanneer de vierde afgeleide van de integrand een singulariteit vertoont.

4.1 Integratieprocessen met zg. "zelfzoekend interval"

Stel, dat wij voor een gegeven functie f(x) de bepaalde integraal

$$\int_a^b f(x) * dx$$

willen berekenen met een eindige absolute precisie ε . Om dit te bereiken moeten we een geschikte integratiestap h kiezen: kiezen we h te groot, dan halen we de gewenste precisie niet, kiezen we h te klein, dan doen we te veel werk. Voorts mogen we niet verwachten, dat het verstandig is, om over het hele interval met dezelfde integratiestap te werken: over die trajecten waar de functie f(x) zich rustig gedraagt, kunnen we - en willen we dus graag - ons een grotere stap permitteren. Het volgende is van toepassing in de veronderstelling, dat h maximaal zo klein is, dat over een gebied 2 * h de vierde afgeleide van de functie f(x) als min of meer constant beschouwd mag worden. We beperken ons om schrijfwerk te besparen tot het geval a < b.

Het idee is, dat als de totale integraal met een absolute precisie ε berekend mag worden, we al integrerend een "foutdichtheid" ρ = ε/(b - a) mogen introduceren.

Stel, dat we - op grond van een hint of de voorgeschiedenis van het proces - een idee hebben over een geschikte stapgrootte h, terwijl het integratieproces tot x = x1 gevorderd is. We berekenen nu de bijdrage tot de integraal van x1 tot x1 + 2 * h op twee verschillende manieren, nl. door de formule van Simpson één keer op het totale interval [x1, x1 + 2 * h] toe te passen

$$I2 = h * (f(x1) + 4 * f(x1 + h) + f(x1 + 2 * h))/3$$

en door de formule van Simpson op eerste en tweede helft toe te passen en de bijdragen te sommeren:

$$I1 = h*(f(x1)+4*f(x1+.5*h)+2*f(x1+h)+4*f(x1+1.5*h)+f(x1+h))/6$$

Stellen we

$$I2 = \int_{x1}^{x1+2*h} f(x)*dx + R2 \quad \text{en} \quad I1 = \int_{x1}^{x1+2*h} f(x)*dx + R1$$

dan weten we van de ons nog onbekende fouten R2 en R1 dat onder voorbehoud van niet wild variërende vierde afgeleide geldt

$$R2 \approx 16 * R1 .$$

Een schatting voor R1 is dus:

$$R1 \approx R(h) = (I2 - I1)/15 .$$

Omdat we ons een foutdichtheid ρ mogen permitteren, kunnen we dus I1 als volgende bijdrage van de integraal permitteren als

$$abs(R(h)) \leqslant 2 * h * \rho$$

Is aan deze ongelijkheid wel voldaan, dan accepteren we deze stap en gaan bij x1 + 2 * h verder, anders concluderen we, dat we met de grootte van h te optimistisch zijn geweest en beginnen we opnieuw bij x1 te integreren, maar nu met een nieuwe (kleinere) stapgrootte. In beide gevallen hebben we de plicht een nieuwe stapgrootte h' te beslissen.

We doen dit, door op te merken, dat R in eerste benadering evenredig is met de vijfde macht van de stapgrootte. Als

$$abs(R(h)) = C * h\uparrow 5 \text{ en dus } abs(R(h')) = C * h'\uparrow 5$$

dan zoeken we een h', zodat

$$abs(R(h')) \leqslant 2 * h' * \rho.$$

Mikken we op de gelijkheid, dan vinden we de vergelijking

$$h'\uparrow 4 = 2 * \rho/C = h\uparrow 5 * 2 * \rho/abs(R(h))$$

of $$h' = h * (2 * h * \rho/abs(R(h)))\uparrow .25 .$$

In de praktijk zal men iets voorzichtiger zijn en bv. nemen

$$h' = h * 0.9 * (2 * h * \rho/abs(R(h)))\uparrow .25$$

(een andere manier is, om vóór de vierde machtswortel een approximatie te gebruiken, die een veilige minorant is.)

We zien hieruit, dat als aan de ongelijkheid

$$abs(R(h)) < 2 * h * \rho$$

ruimschoots voldaan is, de stap niet alleen geaccepteerd wordt, maar dat het proces dan vol goede moed met h' > h doorgaat.

Tot slot enkele opmerkingen.

Ten eerste doet men er goed aan om aan h zowel een ondergrens als een bovengrens op te geven. Bovendien kan men het beste een bovengrens aan de vergrotingsfactor opleggen (1.5 of 2 of zo), dit laatste om te verhinderen, dat als de fouten R1 en R2 "toevallig" een keer praktisch gelijk zijn, men plotseling verleid zou worden om met een irreëel grote stap door te gaan.

Ten tweede: als we een stap accepteren, kunnen we I1 als bijdrage tot de integraal in rekening brengen. Een beter antwoord mogen we verwachten als we in plaats daarvan

$$I1 - R(h)$$

in rekening brengen, omdat immers R(h) een schatting is voor de afbreekfout van I1. Doen we dit, dan leveren we een aanmerkelijk beter antwoord af, dan geëist, we weten alleen niet meer hoeveel beter.

Tenslotte: als de geëiste precisie zo hoog is dat we een heel groot aantal kleine stapjes moeten nemen, dan moeten we er bij de opbouw van de integraal voor zorgen, dat we niet door afrondingsfouten bij de optelling der individuële bijdragen storingen introduceren.

5. GEWONE DIFFERENTIAALVERGELIJKINGEN

5.1 Een vergelijking van de eerste orde

Onze differentiaalvergelijking zij gegeven in de vorm

$$y' = f(x,y) \ .$$

of, iets uitvoeriger

$$\frac{d}{dx}\, y(x) = f(x,y(x)) \ ;$$

voorts dient onze oplossing in een gegeven punt $x = x0$ te voldoen aan de randvoorwaarde

$$y(x0) = y0 \ .$$

Deze opgave is een generalisatie van de hierboven beschreven integraalberekening, waar

$$y(b) = \int_a^b f(x) \ * \ dx \ ,$$

wat overeenkomt met de differentiaalvergelijking $y' = f(x)$ met de randvoorwaarde $y(a) = 0$. Als wij in de straks te behandelen processen voor de numerieke integratie van differentiaalvergelijkingen voor het rechterlid beperken tot een functie, die alleen van x en niet van y afhangt, dan zullen we dus een - in 't algemeen al bekende! - methode voor integraalberekening vinden.

Opm.

In het Nederlands wordt de term "integratie" zowel gebruikt voor de berekening van bepaalde integralen als voor het oplossen van algemene differentiaalvergelijkingen. In het Engels spreekt men in het eerste geval van "quadrature", in het tweede van "integration".

Bij de integraalberekening hebben we ons beperkt tot de bepaalde integraal

$$\int_a^b f(x) \ * \ dx$$

en dit herleid tot integratie van $x0$ tot $x1 = x0 + h$ met passend gekozen stapgrootte h. Bij de differentiaalvergelijking

$$y' = f(x,y) \quad \text{met} \quad y(x0) = y0$$

zullen we ons analoog beperken tot de - benaderde - berekening van

$$y(x1) \quad \text{met} \quad x1 = x0 + h ,$$

wederom met passend gekozen stapgrootte h.

Onder gebruikmaking van de analytische gedaante van het rechterlid kunnen we differentiëren en schrijven:

$$y'' = \frac{d}{dx} y' = \frac{d}{dx} f(x,y(x)) = fx + fy * y' = fx + fy * f$$

(met de notatie fx voor $\frac{\partial}{\partial x} f(x,y)$ en fy voor $\frac{\partial}{\partial y} f(x,y)$).

We kunnen dan y(x1) benaderen door y1 door de eerste termen van de Taylor-ontwikkeling, bv.

$$y1 = y0 + y'(x0) * h + y''(x0) * h\uparrow 2/2 .$$

Hier maken we een fout van de orde van grootte $h\uparrow 3$; door nogmaals analytisch te differentiëren kunnen we natuurlijk nog een term van de Taylor-ontwikkeling meenemen. In de praktijk wordt het formularium dan vaak al gauw volslagen onhanteerbaar.

Als de eerste analytische differentiatie al op onoverkomelijke moeilijkheden stuit, kan men een term minder meenemen en y1 benaderen door

$$y1^* = y0 + f(x0,y0) * h . \qquad (1)$$

Hierbij maken we per stap van grootte h een fout van de orde $h\uparrow 2$. We introduceren al integrerend dus een fout met "foutdichtheid per eenheid van x" gelijk aan h, zodat met dit proces willekeurig goede resultaten te krijgen zijn, door h maar klein genoeg te kiezen. Proces (1) is echter wel heel weinig geraffineerd.

Exact is y(x1) gegeven door

$$y(x1) = y(x0) + \int_{x0}^{x1} f(x,y) * dx \qquad (2)$$

wanneer we voor y in de integrand de nog onbekende functie y(x) substitueren. Als we de integraal aan de rechterkant met de trapeziumregel benaderen, dan vinden we voor y1 als benadering van y(x1) de relatie

$$y1 = y0 + (f(x0,y0) + f(x1,y1)) * h/2 . \qquad (3)$$

Als rekenschema is dit niet zonder meer bruikbaar, omdat de

rechterkant via f(x1,y1) nog van de onbekende y1 afhangt. (Slechts als de differentiaalvergelijking homogeen is, dwz. f(x,y) van de gedaante

$$y' = f(x,y) = y * F(x) ,$$

geeft (3) aanleiding tot een lineaire vergelijking in y1, die dan direct oplosbaar is.)

In het algemeen zal men (3) iteratief oplossen, door als schatting voor y1 te kiezen y1*, zoals deze door (1) gegeven is, waarbij men de iteratie één maal uitvoert.

Met de afkortingen

$$f0 = f(x0,y0) \quad \text{en} \quad f1^* = f(x1,y1^*)$$

wordt het rekenschema

$$\text{Predictor:} \qquad y1^* = y0 + f0 * h \tag{4}$$

$$\text{Corrector:} \qquad y1 = y0 + (f0 + f1^*) * h/2 .$$

Dit heet het rekenschema van Heun. Ook dit is weliswaar vrij ruw, maar eenvoudig en zeer stabiel.

Het is niet moeilijk - en wel vervelend - om te verifiëren, dat het proces van Heun per stap h een fout van de orde van grootte $h\uparrow 3$ introduceert. Men schrijft daartoe voor f1* een Taylor-ontwikkeling om het punt (xo,yo) tot en met termen van de orde $h\uparrow 2$; door deze uitdrukking in de corrector van (4) te substitueren vindt men van de Taylor-ontwikkeling van y1 de termen tot en met $h\uparrow 3$ (omdat f1* daarin met h vermenigvuldigd wordt).

Daarnaast beschouwen we de Taylor-ontwikkeling

$$y(x1) = y0 + y' * h + y'' * h\uparrow 2/2 + y''' * h\uparrow 3/6 + \ldots\ldots$$

(zie boven: het totale differentiatieproces naar x moet nog een keer verder worden voortgezet om y''' te krijgen). Termsgewijze vergelijking tussen de Taylor ontwikkelingen van y1 en y(x1) toont aan, dat bij de coëfficiënt van $h\uparrow 3$ de eerste afwijking optreedt.

Wij kunnen proberen, om een nauwkeuriger integratieschema te krijgen door niet uit te gaan van de trapeziumregel, maar te kijken, of we een integratieschema kunnen enten op de formule van Simpson.

Laten wij daartoe eens aannemen, dat wij behalve de startwaarde y(x0) = y0 ook al de waarde y1 = y(x1) aan het einde van de eerste stap h tot onze beschikking hebben; wij stellen ons nu ten doel om door te integreren tot x2 = x0 + 2 * h. In dat geval geldt exact

$$y(x2) = y(x0) + \int_{x0}^{x2} f(x,y) * dx$$

wanneer we voor de y in de integrand weer de nog onbekende functie y(x) substitueren. Als wij de integraal aan de rechterkant met de regel van Simpson benaderen vinden we

$$y(x2) \sim y(x0) + \{f(x0,y0) + 4 * f(x1,y1) + f(x2,y(x2))\} * h/3$$

waaraan wij een vergelijking voor y2 als benaderde waarde van y(x2) ontlenen, nl.

$$y2 = y0 + \{f(x0,y0) + 4 * f(x1,y1) + f(x2,y2)\} * h/3 \quad (5)$$

Over het algemeen is dit weer een niet lineaire vergelijking in y2 en kan deze het beste opgelost worden met behulp van een via een predictor verkregen benadering y2* voor y2; we zullen dan, als bij de methode van Heun, een enkele iteratieslag uitvoeren.

Voordat wij straks de draad van ons verhaal weer bij formule (5) opvatten, gaan wij eerst onderzoeken, hoe wij aan een redelijke schatting y2* kunnen komen.

In de punten x0 en x1 kenden we niet alleen de functiewaarden y0 en y1 maar, omdat y(x) de oplossing pretendeerde te zijn van de differentiaalvergelijking

$$y' = f(x,y)$$

kennen we in die twee punten ook de eerste afgeleiden, nl.

$$y0' = f(x0,y0) \quad \text{en} \quad y1' = f(x1,y1).$$

Als schatting y2* kiezen we nu y2* = P3(x2) wanneer P3 het derdegraadspolynoom is, waarvan functiewaarden en eerste afgeleiden in x0 en x1 met die van y overeenstemmen. We zoeken dus een (soort interpolatie-) polynoom P3, bepaald door

$$P3(x0) = y0, \quad P3(x1) = y1, \quad P3'(x0) = y0' \text{ en } P3'(x1) = y1'.$$

Analoog aan bv. de interpolatieformule van Everett kunnen we P3 het meest symmetrisch uitdrukken als we de twee hulpgrootheden p en q invoeren, die met x samenhangen volgens

$$x = x0 + p * h = x1 - q * h$$

(hieruit volgt dus p + q = 1).

De "vrije parameter" y0 moet vermenigvuldigd worden met een derdegraads polynoom, B(x), dat voor p = 0 (dus q = 1) de waarde = 1 aanneemt en een afgeleide = 0 heeft, terwijl dit polynoom in q = 0 (dus p = 1) zowel functiewaarde als afge-

leide = 0 moet hebben. Uit dit laatste volgt, dat het "factorpolynoom" voor y0 dus een factor q↑2 moet hebben, dus van de gedaante

$$B(x)\ q\uparrow 2 * A(p)$$

waar A(p) een eerstegraads polynoom in p is, dat door de voorwaarden in p = 0 (dus q = 1) gegeven is,

uit B(x0) = 1 volgt A(0) = 1
uit B'(x0) = 0 volgt A'(0) = 2 zodat

$$B(x) = q\uparrow 2 * (2 * p + 1)$$

om redenen van symmetrie vinden we de van y0 en y1 afhankelijke bijdragen:

$$q\uparrow 2 * (2 * p + 1) * y0 + p\uparrow 2 * (2 * q + 1) * y0. \quad (6)$$

Om het factorpolynoom C(x) te vinden, waarmee de vrije parameter y0' vermenigvuldigd wordt, bedenken we dat het een dubbel nulpunt in x1 en een enkel nulpunt in x0 moet hebben; dit geeft meteen de factoren p * q 2 ; de eis, dat C'(x0) = 1 is geeft nog een extra factor h, zodat met de bijbehorende symmetrie-overwegingen we de van y0' en y1' afhankelijke bijdragen vinden

$$h * \{ p * q\uparrow 2 * y0' - q * p\uparrow 2 * y1' \}. \quad (7)$$

Optelling van expressies (6) en (7) geeft ons het gezochte polynoom P3(x) en de waarde van y2* vinden we, door x = x2 te stellen, dwz.: p = 2 en q = -1. Zo vinden we

$$y2^* = 5 * y0 - 4 * y1 + h * (2 * y0' + 4 * y1'). \quad (8)$$

Met de afkortingen

$$f0 = f(x0,y0), \quad f1 = f(x1,y1) \quad \text{en} \quad f2^* = f(x2,y2^*)$$

vinden we dus uit (8) als predictor

$$y2^* = 5 * y0 - 4 * y1 + h * (2 * f0 + 4 * f1) \quad (9)$$

en als corrector uit (5)

$$y2 = y0 + h * (f0 + 4 * f1 + f2^*) / 3. \quad (10)$$

Een lacune in ons betoog is, dat wij vooralsnog in het midden hebben gelaten, hoe wij aan y1 kwamen. Wij substitueren nu voor y1 de uitdrukking die door de methode van Heun geleverd wordt. De predictor 9 wordt daardoor aanmerkelijk eenvoudiger nl. $y2^* = yo + 2 * h * (2 * f1 - f1^*)$ en we vinden als totaal rekenschema

```
y1* := y0 + h * f0 ;

y1  := y0 + h * (f0 + f1*) / 2 ;

y2* := y0 + 2 * h * (2 * f1 - f1*) ;

y2  := y0 + h * (f0 + 4 * f1 + f2*) / 3.
```

Als we nu vergeten, dat dit schema is opgebouwd uit een integratiestap volgens de methode van Heun, gevolgd door een die op de regel van Simpson berust en het totale rekenschema als dat van een stap beschouwen, dan ligt het voor de hand dat we in het bovenstaande schema 2 * h door h en x2 resp. y2 door x1, resp. y1 vervangen.

Het is bovendien gebruikelijk het product h * f met de letter k te benoemen. Wij komen na deze redactie tot de aequivalente formules

```
k1 := h * f(x0,y0) ;
k2 := h * f(x0 + h/2, y0 + k1/2) ;
k3 := h * f(x0 + h/2, y0 + k1/4 + k2/4) ;
k4 := h * f(x0 + h, y0 - k2 + 2 * k3) ;
y1 := y0 + (k1 + 4 * k3 + k4) / 6 .
```

(11)

Een dergelijke integratieformule heet een "integratieformule" van Runge-Kutta. We hebben deze formule niet afgeleid, we hebben hem slechts "waarschijnlijk" gemaakt. Nog minder hebben we bewezen, dat de gemaakte fout per stap van de orde van grootte van $h\uparrow 5$ is.

In de praktijk gebruikt men meestal een ander Runge-Kutta schema, waarvan de fout van dezelfde orde van grootte is, nl.

```
k1 := h * f(x0,y0) ;
k2 := h * f(x0 + h/2, y0 + k1/2) ;
k3 := h * f(x0 + h/2, y0 + k2/2) ;
k4 := h * f(x0 + h, y0 + k3) ;
y1 := y0 + (k1 + 2 * k2 + 2 * k3 + k4) / 6 .
```

Men kan algemeen deze coëfficiënten vinden door enerzijds de aldus berekende y1 te ontwikkelen als machtreeks in h, anderzijds de oplossing y(x + h) op grond van de differentiaalvergelijking naar machten van h te ontwikkelen. (Dit is allebei naar rekenwerk.) Door vervolgens te eisen, dat de coëfficiënten tot en met die van $h\uparrow 4$ in beide reeksen overeenstem-

men ongeacht de specifieke functie f, vindt men een aantal vergelijkingen, die afhankelijk blijken. Er is in de keuze van de coëfficiënten dus nog enige vrijheid, vandaar dat verschillende schema's in omloop zijn. (Deze vrijheid is niet voldoende om ook de coëfficiënten van $h\uparrow 5$ overeen te laten stemmen: zou men dit eisen, dan krijgt men strijdige vergelijkingen.) Het laatste schema is niet alleen gebruikelijker dan (11) het is ook in zoverre iets beter, dat er geen negatieve coëfficiënten in voorkomen en daardoor een bronnetje van cijferverlies vermeden wordt.

5.2 Stelsels van gewone differentiaalvergelijkingen

Het vorige is zonder meer overdraagbaar in het geval van stelsels gewone differentiaalvergelijkingen:

$$y[i]' = f[i](x,y[1],\ldots,y[n]) \quad \text{met } 1 \leqslant i \leqslant n .$$

en de startvoorwaarden

$$y[i](x0) = y0[i] .$$

Het analoge rekenschema van Runge-Kutta, dat dan de waarden y1[i] als benadering voor y[i](x1) uitrekent, luidt in dit geval (elke regel voor i = 1,2,...,n):

```
k1[i] := h * f[i](x0,y0[1],...,y0[n]) ;
k2[i] := h * f[i](x0 + h/2,y0[1] + k1[1]/2,...,y0[n] +k1[n]/2) ;
k3[i] := h * f[i](x0 + h/2,y0[1] + k2[1]/2,...,y0[n] +k2[n]/2) ;
k4[i] := h * f[i](x0 + h/2,y0[1]+ k3[1],...,y0[n] + k3[n]/2) ;
y1[i] := y0[i] + (k1[i] + 2 * k2[i] + 2 * k3[i] + k4[i])/6 .
```

Hiermee is dus tevens een middel geschapen om differentiaalvergelijkingen van hoger orde aan te vatten. Immers een differentiaalvergelijking van het type

$$y'' = f(x,y,y') \quad \text{met gegeven } y(x0) \text{ en } y'(x0)$$

kan men via de substitutie y' = z herleiden tot het stelsel

$$y' = z$$

$$z' = f(x,y,z) \quad \text{met gegeven } y(x0) \text{ en } z(x0).$$

Opm.

Voor het veel voorkomende geval

$$y'' = f(x,y)$$

zijn speciale integratieschema's ontwikkeld, die dan minder rekenwerk vergen dan het algemene geval.

5.3 Stabiliteit

Stel, dat wij de oplossing $y = \exp(-x)$ willen vinden van de differentiaalvergelijking

$$y'' = y$$

met de beginvoorwaarden $y(0) = 1$ en $y'(0) = -1$.

De algemene oplossing van deze differentiaalvergelijking is

$$y(x) = a * \exp(-x) + b * \exp(+x)$$

en men kan onze beginvoorwaarden dus ook zo interpreteren, dat ze met zorg gekozen zijn om de oplossing met $a = 1$ en $b = 0$ te vinden.

Als we nu van 0 af in de richting van de positieve x-as numeriek gaan integreren, dan zullen we, ongeacht de gebruikte integratiemethode, door afrondingsfouten etc. een kleine fout introduceren, dwz. in het punt $x1 > 0$ berekenen we een

$$y1 = a1 * \exp(-x1) + b1 * \exp(+x1)$$

waarbij a1 wel ongeveer = 1 en b1 wel ongeveer = 0 zal zijn, maar waarbij we niet meer mogen hopen, dat b1 exact = 0 is. M.a.w. we hebben terwijl we zochten naar de exponentieel afnemende oplossing een spoortje van de exponentieel groeiende oplossing geïntroduceerd. Gaan wij nu door-integreren, dan zal - zelfs als we verder geen enkele fout meer maken - op den duur de exploderende storing de gezochte oplossing volslagen overwoekeren.

De remedie tegen dit euvel valt uitgesproken buiten het bestek van dit college. Als men bv. zo gelukkig is om kennis te hebben van het asymptotisch gedrag van de gezochte oplossing voor $x \to +\infty$, dan kan men bij een heel grote x-waarde beginnen en terug-integreren.

Wel zullen we aandacht schenken aan het verschijnsel van <u>numerieke instabiliteit</u>. Het kan voorkomen, dat hoewel de oorspronkelijke differentiaalvergelijking geen ongewenste groeiende oplossing heeft, de gediscretiseerde vergelijking wel zulk een oplossing heeft.

Een van de bekendste oorzaken is wel, dat de gediscretiseerde vergelijking een recurrente betrekking van hoger orde is dan de oorspronkelijke differentiaalvergelijking. Een heel simpel voorbeeld moge dit toelichten.

Wij beschouwen de differentiaalvergelijking

$$y' = -\lambda * y \qquad \text{met } \lambda > 0 \text{ en } y(0) = 1$$

die in de richting van de positieve x-as geintegreerd moet worden. De differentiaalvergelijking heeft als enige oplossing

$$y = \exp(-\lambda * x)$$

en als we een numeriek schema toepassen, dat wel een exponentieel groeiende oplossing kan produceren, dan hebben we ten duidelijkste met numerieke instabiliteit te doen.

We beschouwen nu het schema van Euler

$$y1 = y0 + h * f(x0,y0).$$

Voor onze simpele vergelijking geeft dit de recurrente betrekking

$$y1 = y0 * (1 - \lambda * h)$$

algemeen

$$y[k + 1] = y[k] * (1 - \lambda * h),$$

dwz. per stap wordt y vermenigvuldigd met $(1 - \lambda * h)$ in plaats van met $\exp(-\lambda * h)$. Hoe kleiner h, hoe beter deze approximatie; we merken - met het oog op later - op, dat, zolang $abs(1-\lambda * h) \leq 1$ is, onze numerieke oplossing, hoewel misschien al zwaar afwijkend van de echte, in elk geval niet explodeert.

In dit verband is de methode van Euler superieur boven de volgende "verfijning". De methode van Euler berekent het begin van de Taylor-reeks en maakt per stap een fout van de orde van $h\uparrow 2$. Als wij de reeksenontwikkelingen om het punt x van de functies $y(x + h)$ en $y(x - h)$ van elkaar aftrekken, dan vallen de quadratische termen tegen elkaar weg, en we zien, dat de benadering

$$y(x + h) = y(x - h) + 2 * h * y'(x)$$

slechts een fout van de orde $h\uparrow 3$ maakt. Men zou kunnen denken, dat deze methode daardoor superieur zou zijn ten opzichte van die van Euler. Dit is wegens de inhaerente numerieke instabiliteit van dit proces echter niet het geval. Immers, dit rekenschema zou in dit geval van $y' = -\lambda * y$ neerkomen op:

$$y[k + 1] + 2 * h * \lambda * y[k] - y[k - 1] = 0.$$

In tegenstelling tot de oorspronkelijk differentiaalvergelijking, die slechts een (onafhankelijke) oplossing heeft, heeft deze recurrente betrekking, die na discretisatie ontstaan is er twee. De algemene oplossing is nl.

$$y[k] = a * X1\uparrow k + b * X2\uparrow k,$$

waarbij $X = X1$ en $X = X2$ de wortels zijn van de vierkantsvergelijking

$$X\uparrow 2 + (2 * h * \lambda) * X - 1 = 0;$$

het product van deze wortels is = - 1, hun som is niet = 0, en dus heeft een van beide wortels een modulus > 1, zodat de numerieke oplossing vroeg of laat onherroepelijk explodeert.

Tenslotte zullen we een derde integratiemethode behandelen, nl. de methode van Heun in de vereenvoudiging, die mogelijk is, als de differentiaalvergelijking homogeen lineair is. Men benadert dan de integraal in

$$y(x1) = y(x0) + \int_{x0}^{x1} y'(x) * dx$$

met de trapeziumregel en vindt als lineaire vergelijking voor y1

$$y1 = y0 - (y1 + y0) / \lambda * h/2$$

wat leidt tot de algemene recursievergelijking

$$y[k + 1] = y[k] * (1 - \lambda * h/2) / (1 + \lambda * h/2)$$

de factor

$$\frac{1-\lambda * h/2}{1+\lambda * h/2}$$

is bij negatieve λ (wat verondersteld was) voor elke positieve h in absolute waarde kleiner dan 1 en we krijgen op deze manier dus <u>nooit</u> exploderende oplossingen.

We hebben hier dus drie gevallen gezien: het laatste schema geeft nooit instabiliteit, het voorlaatste nooit, terwijl het eerste aanleiding tot instabiliteit kan geven, als we h te groot kiezen.

Van het eerste geval geeft "Modern Computing Methods" (N.P.L. Notes on Applied Science, No 16) de volgende omschrijving (pg 91):

"If the differential equation has some solutions which decrease very rapidly compared with others, then it may happen that only the latter are adequately represented by the finite difference equation, while the former are transformed into rapidly increasing functions."

en in het volgende voorbeeld wordt geïllustreerd, dat dit verschijnsel zich ook bij bv. Runge-Kutta formules kan voordoen.

Beschouw de vergelijkingen

$$y' = - 10 * y + 6 * z \quad \text{en} \quad z' = 13.5 * y - 10 * z$$

met $y(0) = 4 * e/3$ en $z(0) = 0$.

De analytische oplossing is

$$y = (2/3) * e - (\exp(-x) + \exp(-19 * x)) \quad \text{en}$$

$$z = e(\exp(-x) - \exp(-19 * x)).$$

Voor waarden van $x > 1$ is de tweede exponentiële functie in 7 decimalen verwaarloosbaar ten opzichte van de eerste. Als wij echter, startend bij de (exacte) waarden voor $x = 1$ deze vergelijkingen gaan integreren met een stapgrootte $h = 0.2$, dan vinden we - als we de berekening in 2 decimalen uitvoeren om het effect van afrondingen drastischer te illustreren - de volgende waarden:

x	1.0	1.2	1.4	1.6	1.8	2.0
y	0.67	0.55	0.46	0.40	0.41	0.68
z	1.00	0.82	0.66	0.51	0.29	-0.28

De laatste waarden zijn kennelijk incorrect: de oplossing $\exp(-19 * x)$ is getransformeerd in een drastisch stijgende oplossing!

De verklaring is analoog aan die bij de situatie van Euler, waar $\exp(-\lambda * h)$ door $1 - \lambda * h$ benaderd werd. Het proces van Runge-Kutta benadert $\exp(-\lambda * h)$ door

$$E1 = 1 - \lambda * h + (\lambda * h)\uparrow 2/2 - (\lambda * h)\uparrow 3/6 + (\lambda * h)\uparrow /24.$$

Dit is een goede benadering voor de eerste exponentiële functie, waar $\lambda * h = 0.2$ is; in de tweede exponentiële functie is $\lambda * h = 3.8$; terwijl $\exp(-3.8) = 0.02$ ... vinden we daarvoor $E1 = 3.96$ ..... Wil men een dergelijk stelsel met Runge-Kutta oplossen, dan moet men dus een veel kleinere stapgrootte h kiezen. Een integratieformule die berust op het gebruik van de trapeziumregel zou geen moeilijkheden hebben opgeleverd.

Tenslotte een motivering, waarom wij zoveel gewicht er aan hechten, dat uitstervende oplossingen niet overgaan in groeiende. Dit berust op het feit, dat differentiaalvergelijkingen van het type

$$y' = y * f(x) + g(x)$$

veel voorkomen, waarbij $f(x) < 0$ is. (De generalisatie tot het meer-dimensionale geval zullen wij later tegenkomen.)

Als $y1(x)$ een particuliere oplossing is en $y2(x)$ een oplossing van de homogene vergelijking

$$y' = y * f(x)$$

dan is de algemene oplossing

$$y1(x) + a \cdot y2(x).$$

Als $f(x) < 0$, dan betekent dit, dat $y2(x) \to 0$ voor $x \to +\infty$, dwz. alle oplossingen kruipen, ongeacht de startcondities, naar de particuliere oplossing toe naarmate we door-integreren. Het is vaak de particuliere oplossing ten opzichte waarvan het begrip stabiliteit "geijkt" wordt.

5.4 Meerpunts randvoorwaarden bij gewone differentiaalvergelijkingen

In deze paragraaf zullen wij ons essentieel beperken tot lineaire differentiaalvergelijkingen. Wij bekijken nu de opgave, waarbij de randvoorwaarden, betrekking hebben op beide einden van het interval. Als voorbeeld beschouwen we de differentiaalvergelijking

$$y'' + f(x) \cdot y' + g(x) \cdot y = k(x) \qquad (1)$$

terwijl in de punten a en b ($a < b$) de waarden $y(a) = A$ $y(b) = B$ gegeven zijn.

Er zijn nu essentieel twee verschillende methoden van oplossing, waarbij we de stap-voor-stap methode eerst zullen beschrijven.

Dit komt er in wezen op neer, dat we differentiaalvergelijking (door de substitutie $z = y'$ in een stelsel overgevoerd) normaal zouden kunnen integreren van a naar b, als we behalve $y(a)$ ook maar $y'(a)$ zouden weten.

Het komt er op neer, dat we in het x-y-vlak een curve zoeken, die aan de differentiaalvergelijking (1) voldoet en verder door de punten (a,A) en (b,B) gaat. We kunnen proberen, deze curve te vinden, door in het punt (a,A) maar eens een helling $y'(a) = C$ te schatten en te proberen, of we daarmee weg-integrerend het punt (b,B) kunnen raken. Onze oplossing wordt dan een functie van deze starthelling, we vinden een curve $y(x) = Y(C,x)$ en we kunnen proberen om C uit de vergelijking

$$Y(C,b) = B$$

op te lossen. Als we een redelijke beginschatting voor C hebben en $Y(C,b)$ een niet te wilde - bij voorkeur monotone - functie van C is, kunnen we C door insluiting, (halvering of eventueel zelfs Regula Falsi) bepalen. Dit is een techniek niet ongelijk aan degene, waarmee de kanonnier door successieve approximatie zijn doel raakt: op de plaats waar wij een integratie uitvoeren, vuurt hij een schot af, terwijl hij tevens over een waarnemer beschikt, die de comparatie tussen $Y(C,b)$ en B maakt.

In het geval van een homogene vergelijking kunnen we in principe althans, eenvoudiger te werk gaan, omdat, gegeven twee

oplossingen y1(x) en y2(x) het lineair compositum $(\frac{1}{2} + s) * y1(x) + (\frac{1}{2} - s) * y2(x)$ ten duidelijkste weer een oplossing is. Voldoen beide oplossingen y1(x) en y2(x) aan de randvoorwaarde in a, dan geldt dit ook voor het lineair compositum, zodat we s vrijelijk kunnen oplossen uit de vergelijking

$$(\tfrac{1}{2} + s) * y1(b) + (\tfrac{1}{2} - s) * y2(b) = B .$$

Opm.

Als deze wortel voldoet aan abs(s) >> $\frac{1}{2}$, dan impliceert het vormen van het lineair compositum wel een aanzienlijk cijferverlies, zodat dan de methode toch niet zo aantrekkelijk is. Dit is in het algemeen het geval, wanneer de oplossingen van de homogene vergelijking een exponentieel verloop hebben.

Het voordeel van deze methode is, dat de individuele integraties in de x-richting met variabel - dwz. optimum - interval uitgevoerd kunnen worden, een voordeel, dat bij de tweede methode verloren gaat.

Hier beginnen we, het interval langs de x-as onder te verdelen in n gelijke stukken ter lengte h = (b - a)/n. We voeren in de notaties (i = 0,1,...,n):

$$x[i] = a + i * h \qquad \text{en} \qquad (2)$$

$f[i] = f(x[i])$, $g[i] = g(x[i])$, $k[i] = k(x[i])$ en $y[i] = y(x[i])$.

Door nu voor y'(x[i]) en y"(x[i]) differentiebenaderingen op te stellen, kan men het hele probleem herleiden tot een stelsel lineaire vergelijkingen.

Benaderen wij bv. voor i = 1,2,...,n - 1

$$h * y'(x[i]) \sim (y[i + 1] - y[i - 1]) / 2$$

en (3)

$$h\uparrow 2 * y''(x[i]) \sim (y[i + 1] - 2 * y[i] + y[i - 1])$$

dan vinden we voor n - 1 inwendige punten uit (1) de n - 1 vergelijkingen:

$$(1 - \tfrac{1}{2} * h * f[i]) * y[i - 1] - (2 - h\uparrow 2 * g[i]) * y[i]) + (1 + \tfrac{1}{2} * h * f[i + 1]) =$$

$$= h\uparrow 2 * k[i] . \qquad (4)$$

Voor i = 1,2,...,n - 1 staan hier n - 1 vergelijkingen in n - 1 onbekenden, omdat de er tevens in voorkomende y[0] en y[n] door de randvoorwaarden y[0] = A en y[n] = B gegeven zijn.

De orde van het stelsel (4) kan heel groot zijn; de oplossing

wordt echter aanzienlijk vergemakkelijkt door het feit dat de coëfficiëntenmatrix een zg. bandmatrix is, en zelfs een tamelijk smalle: er bevinden zich uitsluitend coëfficiënten $\neq 0$ op de hoofd-diagonaal en de beide daarlangs liggende nevendiagonalen:

$$\begin{array}{cccccc} b[1] & c[1] & & & & d[1] \\ a[2] & b[2] & c[2] & & & d[2] \\ & a[3] & b[3] & c[3] & & d[3] \\ & & & \ddots & & \\ & & a[n-2] & b[n-2] & c[n-2] & d[n-2] \\ & & & a[n-1] & b[n-1] & d[n-1] \end{array} \qquad (5)$$

met

$$a[i] = 1 - \tfrac{1}{2} * h * f[i]$$

$$b[i] = -(2 - h\uparrow 2 * g[i])$$

$$c[i] = 1 + \tfrac{1}{2} * h * f[i]$$

$$d[1] = h\uparrow 2 * k[1] - a[1] * y[0]$$

$$d[i] = h\uparrow 2 * k[i] \quad \text{voor} \quad 2 \leq i \leq n-2$$

$$d[n-1] = h\uparrow 2 * k[n-1] - c[n-1] * y[n]$$

(de uiterste rechterleden zijn even afwijkend omdat $y[0]$ en $y[n]$ niet onder de onbekenden vallen.)

Door de eerste vergelijking vermenigvuldigd met $a[2] / b[1]$ van de tweede af te trekken, kan men in de tweede vergelijking $y[1]$ elimineren en krijgt men een vergelijking van het type

$$\beta[2] * y[2] + c[2] * y[3] = \delta[2]$$

die op zijn beurt weer gebruikt kan worden, om $y[2]$ uit de derde vergelijking te elimineren. Zo krijgt men voor $i = 1,2,\ldots, n-2$:

$$\beta[i] * y[i] + c[i] * y[i+1] = \delta[i] \qquad (6)$$

en als laatste

$$\beta[n-1] * y[n-1] = \delta[n-1] \qquad (6')$$

(hierbij geldt: $\beta[1] = b[1]$, $\delta[1] = d[1]$ en verder de recurrente betrekkingen

$$\beta[i] = b[i] - c[i-1] * a[i] / \beta[i-1] \quad \text{en}$$

$$\delta[i] = d[i] - \delta[i-1] * a[i] / \beta[i-1] \quad).$$

Uit (6') kan $y[n-1]$ berekend worden en vervolgens kan men de recurrente betrekking (6) gebruiken, om "achterwaarts" de overige $y[i]$ te berekenen.

Moeilijkheden kunnen slechts optreden, als een van de β's klein is; in menig wel-gesteld probleem komt dit gelukkig niet voor.

Aangenomen, dat we het lineaire stelsel exact opgelost hebben, dan wil dat nog niet zeggen, dat we ook onze differentiaalvergelijking voldoende nauwkeurig hebben opgelost. Immers, bij onze discretisering in de x-richting hebben we door de substituties (3) in de representatie van de afgeleiden y' en y" een afbreekfout gemaakt.

Er zijn twee methoden om deze afbreekfout te verkleinen. De eerste methode is om h kleiner te kiezen, een maatregel waartoe me voor de komst van rekenautomaten liever niet toe overging, omdat dan het aantal vergelijkingen zo stijgt. De tweede methode is om voor y' en y" minder grove benaderingen in te voeren, nl.

$$h * y'[k] \sim (\mu * \delta - \frac{1}{6} \mu * \delta\uparrow 3 + \frac{1}{30} \mu * \delta\uparrow 5 - \ldots) * y[k]$$

$$h\uparrow 2 * y''[k] \sim (\delta\uparrow 2 - \frac{1}{12} \delta\uparrow 4 + \frac{1}{90} \delta\uparrow 6 \ldots) * y[k] \qquad (7)$$

waarbij van de centrale differentie-operator

$$\delta * y[k] = y[k + \tfrac{1}{2}] - y[k - \tfrac{1}{2}]$$

en de centrale middelingsoperator

$$\mu * y[k] = (y[k + \tfrac{1}{2}] + y[k - \tfrac{1}{2}]) / 2$$

gebruik gemaakt is.

Wij geven een schets van de afleiding van formules (7).

De verplaatsingsoperator E en de differentiaaloperator $\theta = h * \frac{d}{dx}$ zijn zoals bekend verbonden door

$$E = e\uparrow\theta .$$

Uit de definitie van δ volgt, dat

$$\delta = E\uparrow(1/2) - E\uparrow(-1/2) = e\uparrow(\theta/2) - e\uparrow(-\theta/2) = 2 * \sinh(\theta/2)$$

Als wij de inverse functie van de sinh met "arcsinh" aanduiden, volgt hier dus uit

$$\theta = 2 * \text{arcsinh}\,(\delta/2)$$

waaruit volgt

$$h\uparrow 2 * y''[k] = \theta\uparrow 2 * y[k] = (2 * \text{arcsinh}(\delta/2))\uparrow 2 * y[k] =$$

$$(\delta\uparrow 2 - \frac{1}{12}\,\delta\uparrow 4 + \frac{1}{90}\,\delta\uparrow 6 \ldots) * y[k]$$

waarmee de tweede regel van 7 is afgeleid (als we de coëfficiënten van de machtreeks van de arcsinh tenminste voor lief nemen).

De eerste regel is iets moeilijker, omdat θ zelf een oneven functie in δ is en we de afgeleide op deze manier zouden uitdrukken in functiewaarden juist midden tussen de steunpunten, die we niet tot onze beschikking hebben. Hier brengt de middelingsoperator μ uitkomst, gegeven door

$$\mu = (E\uparrow(1/2) + E\uparrow(-1/2)) / 2$$

waaruit onmiddellijk volgt, dat

$$\mu\uparrow 2 = 1 + \delta\uparrow 2/4 .$$

We schrijven nu:

$$h * y'[k] = \theta * y[k] = 2 * \text{arcsinh}\,(\delta/2) * y[k] =$$

$$(1 + \delta\uparrow 2/4)\uparrow(-1/2) * (2 * \text{arcsinh}\,(\delta/2) * \mu) * y[k] .$$

Vermenigvuldiging van de bijbehorende formele machtreeksen geeft de eerste formule van (7).

Bij de benaderingen van (3) hebben we de eerste term van de benaderingen (7) gebruikt. Gebruik van de benaderingen 7 geeft in vergelijkingen (4) een extra term

$$+ C[i] * y[i]$$

waarbij de operator C[i] gegeven is door

$$C[i] = (-\frac{1}{12}\,\delta\uparrow 4 + \frac{1}{90}\,\delta\uparrow 6 \ldots) +$$

$$h * f[i] * (-\frac{1}{6}\,\mu * \delta\uparrow 3 + \frac{1}{30}\,\mu * \delta\uparrow 5 - \ldots) .$$

In plaats van het stelsel (5) wat we verkort in matrixnotatie kunnen opschrijven als

$$A \,.\, \underline{y} = \underline{d}$$

hebben we nu het stelsel wat we kunnen schrijven

$$(A + C) \cdot \underline{y} = \underline{d}$$

wat we echter opschrijven als

$$A \cdot \underline{y} = \underline{d} - C \cdot \underline{y}$$

hierbij de vector $C \cdot \underline{y}$ beschouwend als correctie op het rechterlid. Het feit dat A een (smalle) bandmatrix is kunnen we blijven exploiteren bij gratie van het feit, dat de elementen van $C \cdot y$ klein zijn.

Een goede benadering krijgen we door de correctie $C \cdot y$ te vergeten en op te lossen

$$A \cdot \underline{y}^{(1)} = \underline{d}.$$

We corrigeren $\underline{y}^{(1)}$ dan met $\underline{\eta}^{(1)}$ die we vinden uit

$$A * \underline{\eta}^{(1)} = - C * \underline{y}^{(1)}.$$

Als $C * \underline{\eta}^{(1)}$ significant is, dan kunnen we $\underline{y}^{(1)} + \underline{\eta}^{(1)}$ nogmaals corrigeren met $\underline{\eta}^{(2)}$ die we vinden uit

$$A * \eta^{(2)} = - C * \underline{\eta}^{(1)}$$

waarop we $\underline{y}^{(1)} + \underline{\eta}^{(1)} + \underline{\eta}^{(2)}$ als beste benadering gebruiken. (In de praktijk is het ongebruikelijk om deze iteratie vaker dan 2 maal uit te voeren. Als er dan nog geen convergentie bereikt is, is het interval h te groot gekozen of is er iets ergers aan de hand.)

Opm.

De uitwerking van het eerste en laatste element van $C * y^{(1)}$ stuit op moeilijkheden, omdat we voor de uitvoering van de operatoren $\delta\uparrow 4$ en $\mu * \delta\uparrow 3$ ook de buiten het interval gelegen waarden $y^{(1)}[-1]$ en $y^{(1)}[n + 1]$ zouden willen hebben. (Willen we ook de volgende differentiecorrecties hebben, dan moeten we ook een $y^{(1)}[-2]$ en $y^{(1)}[n + 2]$ extrapoleren.) Voor deze extrapolatie kan men gebruik maken van relatie (4), door deze voor $y = y^{(1)}$ en voor $i = 0$ resp. $= n$ op te lossen naar $y^{(1)}[-1]$ resp. $y^{(1)}[n + 1]$.

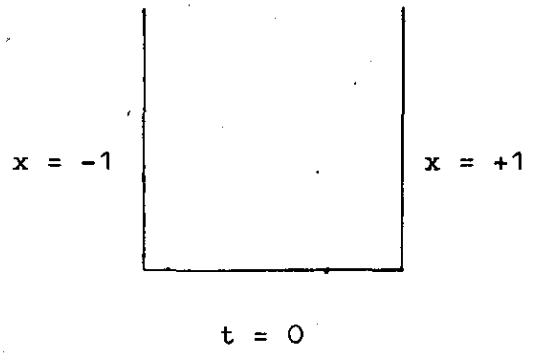
Het is speciaal dit gepriegel aan de uiteinden van het interval, dat deze methode voor automatisch rekenen iets minder aantrekkelijk maakt.

5.5 Een eenvoudige parabolische differentiaalvergelijking

Problemen van bv. warmtegeleiding en diffusie geven aanleiding tot partiële differentiaalvergelijkingen, waarvan zo ongeveer het simpelste type is

$$\frac{\partial^2 f}{dx^2} = \frac{\partial f}{\partial t} \qquad (1)$$

met randvoorwaarden op de lijnen $x = \pm 1$, $t > 0$ en $t = 0$, $-1 \leq x \leq 1$.



We hebben hier dus gemengde randvoorwaarden: in de x-richting hebben we de meer-punts randvoorwaarde, in de t-richting hebben we beginvoorwaarden.

We kiezen in de x-richting aequidistante steunpunten x[0] = -1, ..., x[n] = +1 en noemen de afstand tussen twee opeenvolgende steunpunten hx.

Deze discretisatie voert de partiële differentiaalvergelijking over in

$$(hx)\uparrow 2 * \frac{df[i]}{dt} = (f[i-1] - 2 * f[i] + f[i+1]) + Cx * f[i] \qquad (2)$$

waarin Cx de differentiecorrectie in de x-richting

$$Cx = -\frac{1}{12} * \delta\uparrow 4 + \frac{1}{90}\delta\uparrow 6 - \ldots \quad \text{is} \qquad (3)$$

(centrale differenties in de x-richting). We nemen echter aan, dat de differentie-correcties in de x-richting verwaarloosd kunnen worden.

Vergelijkingen (2) voor i = 1,... n - 1 representeren een stelsel van n - 1 gewone differentiaalvergelijkingen voor de n - 1 functies f[i](t), en we kunnen hier de inmiddels behan-

delde methoden op loslaten.

Om de stabiliteit te onderzoeken gaan wij over op de variabele $T = t / h \uparrow 2$ en beschouwen wij de homogene vergelijkingen - dwz. we stellen $f[0](t) = 0$, $f(n)(t) = 0$.

In vectornotatie luiden onze vergelijkingen dan

$$\frac{d}{dT} \underline{f} = -A * f$$

waarbij de matrix A van de orde n - 1 is met de gedaante

$$\begin{array}{ccccccc} +2 & -1 & & & & & \\ -1 & +2 & -1 & & & & \\ & -1 & +2 & -1 & & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -1 & +2 & -1 \\ & & & & & -1 & +2 \end{array}$$

en elders nullen. (Het minteken voor A is ingevoerd om de aansluiting met 5.3 duidelijker te maken.)

De algemene oplossing van dit stelsel is

$$\underline{f}(T) = e^{-A*T} * \underline{f}(0) .$$

De symmetrische matrix A heeft slechts reële eigenwaarden, die - zoals volgt uit de methode van Givens - alle verschillend zijn en - zoals volgt uit de stelling van Gershgorin - alle liggen op het interval $0 < \lambda < 4$. (Nauwkeurigere analyse toont aan, dat voor grote n de uiterste eigenwaarden inderdaad dicht bij deze grenzen komen te liggen.)

Als we dit stelsel numeriek integreren met de methode van Euler, dwz.

$$\underline{f}(T + \Delta T) = \underline{f}(T) + \Delta T * f'(T)$$

$$= (I - A * \Delta T) * \underline{f}(T)$$

dan krijgen we alleen maar een stabiel systeem, als de eigenwaarden van

$$I - A * \Delta T$$

in absolute waarde kleiner zijn dan 1. Dit betekent, dat de grootste eigenwaarde van $A * \Delta T < 2$ moet zijn, en omdat de grootste eigenwaarde als redelijke maiorant heeft, volgt hieruit dat

$$\Delta T < .5 .$$

M.a.w. hieruit volgt, dat $\Delta t < \frac{hx\uparrow 2}{2}$.

Als wij dus h verkleinen, dan krimpt de maximaal toelaatbare stap in de t-richting kwadratisch: de hoeveelheid werk, om de partiële differentiaalvergelijking tot een einde tijd t te integreren neemt dus toe met $h\uparrow(-3)$. Voor kleine waarden van h wordt dit alras prohibitief.

Als we het stelsel integreren met Runge-Kutta, dan benaderen we effectief

$$e^{-A*\Delta T} \approx E1 = I - A * \Delta T + (A * \Delta T)\uparrow 2/2 - (A * \Delta T)\uparrow 3/6 + (A * \Delta T)\uparrow 4/24$$

waarvan de grootste eigenwaarde slechts absoluut kleiner is dan 1, als de grootste eigenwaarde van $A * \Delta T < 2.8$ ... is.

(Deze bovengrens komt omdat voor $0 < x < 2.8$ geldt,

$$abs(1 - x + x\uparrow 2/2 - x\uparrow 3/6 + x\uparrow 4/24) < 1 \quad).$$

Dit geeft voor ΔT een iets mildere voorwaarde, nl.

$$\Delta T < .7$$

zij het, dat ook hier de maximale dt kwadratisch van hx afhangt.

Tenslotte hebben we de stabiele methode, die uitgaat van de trapeziumregel

$$\underline{f}(T + \Delta T) = \underline{f}(T) - A * \Delta T * (\underline{f}(T) + \underline{f}(T + \Delta T))/2$$

waarbij

$$\underline{f}(T + \Delta T) = (I + A * \Delta T/2)^{-1} * (I - A * \Delta T/2) * \underline{f}(T) .$$

Hierbij wordt door stabiliteitsoverwegingen geen bovengrens aan ΔT opgelegd en hangt de maximale stap in de t-richting dus niet meer kwadratisch van die in de x-richting af, maar meer van het gedrag van de functies in de t-richting. Wij zullen deze methode, de methode van Crank-Nicolson, omdat hij door experts wordt aanbevolen, iets uitvoeriger bespreken.

We beginnen met de afleiding van een andere integratieformule, nl. een, die ons een uitdrukking geeft voor de afbreekfout van de trapeziumregel.

$$\tfrac{1}{2} * h * (y'[k] + y'[k+1]) = \theta * \mu * y[k+\tfrac{1}{2}] =$$

$$(2 * \operatorname{arcsinh}(\delta/2)) * (1 + \delta^2/4)\uparrow(1/2) * y[k+\tfrac{1}{2}] =$$

$$(\delta + \tfrac{1}{12}\,\delta\uparrow 3 - \tfrac{1}{120}\,\delta\uparrow 5 + \ldots) * y[k+\tfrac{1}{2}] .$$

Bedenkend, dat $\delta * y[k+\tfrac{1}{2}] = y[k+1] - y[k]$, zien wij bij verwaarlozing van de hogere machten van δ het integratieschema via de trapeziumregel ontstaan:

$$y[k+1] - \tfrac{1}{2} * h * y'[k+1] = y[k] + \tfrac{1}{2} * h * y'[k] + C * y[k+\tfrac{1}{2}] \qquad (4)$$

waar

$$C = -\tfrac{1}{12}\,\delta\uparrow 3 + \tfrac{1}{120}\,\delta\uparrow 5 - \ldots \qquad (5)$$

Wij passen deze formule toe op $y = f[i]$ in (2) k, resp. k + 1 overeen latende komen met de tijdstippen t en t + ht. Met de afkorting

$$s = (hx)\uparrow 2 \,/\, ht$$

vinden we voor het linkerlid van (4) na vermenigvuldiging met 2s en met inachtneming van (2)

$$2 * s * f[i](t+ht) - s * ht * f'[i](t+ht) =$$

$$-\{f[i-1] - 2*(1+s)*f[i] + f[i+1] + Cx * f[i]\}_{t+ht}$$

met $\{\quad\}_{t+ht}$ aanduidend, dat de functiewaarden tussen accolades ten tijde t + ht genomen moeten worden. Het rechter lid van (4) wordt analoog hefleid en we vinden als we met Ct de differentiecorrectie als in (5) maar nu in de t-richting aanduiden het volgende stelsel vergelijkingen (voor i = 1,2,...n-1)

$$-\{f[i-1] - 2*(1+s)*f[i] + f[i+1] + Cx * f[i]\}_{t+ht} =$$

$$+\{f[i-1] - 2*(1-s)*f[i] + f[i] + Cx * f[i]\}_{t} + 2*s\{Ct * f[i]\}_{t+ht/2}$$

(6)

Dit is een stelsel vergelijkingen voor de grootheden ten tijde t + ht (op de differentiecorrectie $2 * s\{Ct * f[i]\}_{t+ht/2}$ na) links van het gelijkteken, dat we naar deze onbekenden kunnen oplossen. Bij de oplossing van dit stelsel wil men natuurlijk weer exploiteren, dat het hoofdbestanddeel van de coëfficiëntenmatrix een tridiagonaalmatrix is. Men kan iteratief te werk gaan als in 5.4 beschreven; doorgaans kan men zich iteratie besparen door bij de oplossing van 6 er van uit te gaan, dat $\{Cx * f[i]\}_t$ een goede benadering zal zijn voor $\{Cx * f[i]\}_{t+ht}$.

Om de differentiecorrectie Ct * f in rekening te brengen is veel en veel lastiger, omdat men daarvoor toekomstige waarden van f voor t + n * ht moet hebben,waarover men nog lang niet de beschikking heeft. Het is dan ook gebruikelijk om de correctie Ct * f maar te vergeten en het integratieproces met verschillende stappen ht uit te voeren.

Tot slot wil ik eigenvectoren en eigenwaarden van twee soorten matrices, die we in dit verband vaak tegenkomen, analytisch bepalen.

De eerste manier is de bandmatrix van orde n met elementen 2 op de hoofddiagonaal en elementen -1 op de nevendiagonalen, overige elementen = 0.

Noemen wij de elementen van een eigenvector x[1] ..... x[n] dan vinden we de vergelijkingen

$$(2 - \lambda) * x[1] - x[2] = 0$$

$$-x[i-1] + (2 - \lambda) * x[i] - x[i+1] = 0 \qquad 2 \leq i \leq n-1$$

$$-x[n-1] + (2 - \lambda) * x[n] = 0 .$$

De beide uiterste vergelijkingen kunnen we ook als de middelste schrijven, mits we de nevenconditie

$$x[0] = x[n+1] = 0$$

opleggen.

De recursievergelijking

$$x[i-1] - (2 - \lambda) * x[i] + x[i+1] = 0$$

heeft als algemene oplossing

$$x[i] = a * \sin(i * \varphi) + b * \cos(i * \varphi)$$

waarbij

$$(2 - \lambda) = 2 * \cos(\varphi) .$$

Uit de voorwaarde x[0] = 0 volgt, dat b = 0 is, uit de voorwaarde x[n + 1] = 0 volgt dat sin((n + 1) * φ) = 0 moet zijn. Hieraan is voldaan, mits

$$(n + 1) * \varphi = k * \pi \qquad 1 \leq k \leq n .$$

De elementen x[i,k] van de k-de eigenvector zijn dus gegeven door

$$x[i,k] = \sin(k * \pi / (n + 1)) ;$$

de bijbehorende eigenwaarde is gegeven door

$$\lambda = 2 * (1 - \cos(k * \pi/(n + 1))$$

$$= 4 * \sin(k * \pi/(2 * n + 2))\uparrow 2 .$$

We zien dus dat aan $0 < \lambda < 4$ voldaan is - dat wisten we al dank zij Gershgorin - we zien ook, dat voor grotere n deze grenzen willekeurig dicht genaderd worden.

De tweede matrix, waarvan we de eigenvector en eigenwaarden analytisch zullen bepalen is de matrix, waarvan de rijen verkregen worden door de elementen van de eerste rij cyclisch te permuteren, zodat op de hoofddiagonaal steeds hetzelfde element staat. Een dergelijke matrix krijgt men bv. bij een randwaardeprobleem, waarvan geeist wordt dat de oplossing periodiek is.

Als de matrix van orde n is en de eerste rij bestaat uit de elementen

$$a[0]\ \ a[1]\ \ \ldots\ldots\ldots\ \ a[n-1]$$

dan is gemakkelijk in te zien dat de vector

$$x[j] = W[k]\uparrow j \qquad (o \leqslant j, k \leqslant n-1)$$

de k-de eigenvector is, mits W[k] één van de n-de machtswortels van 1 is:

$$W[k] = \cos(k * 2 * \pi/n) + \underline{i} * \sin(k * 2 * \pi/n) .$$

De bijbehorende eigenwaarden volgen bv. uit de eerste rij (NB: x[0] = 1!):

$$\lambda[k] = \text{SIGMA}\ (j,0,n-1,\ a[j] * W[k]\uparrow j) .$$