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

The Mathematical Centre, founded the 11th of February 1946, is a non - profit institution aiming at the promotion of pure mathematics and its applications, and is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.) and the Central Organization for Applied Scientific Research in the Netherlands (T.N.O.), by the Municipality of Amsterdam and by several industries.

MC SYLLABUS

2.2

MC SYLLABUS

2.2

COLLOQUIUM STABILITEIT VAN DIFFERENTIESCHEMA'S

DEEL 2

DOOR

L. DEKKER
T.J. DEKKER
P.J. VAN DER HOUWEN
M.N. SPIJKER

MATHEMATISCH CENTRUM AMSTERDAM
1968

Inhoud

1. Numerieke bepaling van eigenwaarden en eigenvectoren

1.1 Theorie

Het algebraïsche eigenwaarde-probleem

Gegeven een vierkante matrix A van de orde n, worden gevraagd getallen λ, waarvoor de vector-vergelijking

(1.1.0) $$Ax = \lambda x$$

een oplossings-vector $x \neq \vec{0}$ heeft. Zo'n vector x bestaat dan en slechts dan, als

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

Deze vergelijking heet de karakteristieke vergelijking van A.
De determinant is een polynoom van de graad n, met coëfficiënt van λ^n gelijk aan 1, dus er zijn hoogstens n verschillende waarden λ, die voldoen. Deze waarden λ heten eigenwaarden van A.
Bij elke eigenwaarde λ hoort minstens één oplossing x van (1.1.0).
Zo'n oplossings-vector heet een bij λ behorende eigenvector van A.
Als x een eigenvector is horende bij een eigenwaarde λ, dan ook kx voor een willekeurig getal $k \neq 0$. Hierom worden de eigenvectoren meestal genormeerd, zodat een of andere norm (zie onder) gelijk aan 1 is.

Het getransponeerde eigenwaarde-probleem

Naast het eigenprobleem horend bij A is van belang het getransponeerde eigenprobleem, dat is het eigenprobleem horende bij A^T. De eigenwaarden van A^T zijn dezelfde als die van A. Een bij λ horende eigenvector y van A^T voldoet dus aan

(1.1.2) $$A^T y = \lambda y,$$

wat ook kan worden geschreven als

(1.1.3) $$y^T A = \lambda y^T.$$

De rij-vectoren y^T heten eigenrijen van A en ter onderscheiding hiervan heten de eigenvectoren x ook wel eigenkolommen.

(1.1.4) Stelling. Is x een eigenkolom horende bij λ en y^T een eigenrij horende bij $\mu \neq \lambda$, dan geldt $y^T x = 0$.

Enkelvoudige eigenwaarden

Eigenvectoren horende bij verschillende eigenwaarden zijn lineair onafhankelijk. Zijn alle wortels van de karakteristieke vergelijking enkelvoudig, dan zijn er dus n verschillende eigenwaarden $\lambda_1, \ldots, \lambda_n$. De n bijbehorende eigenvectoren $x_1, \ldots, x_n$ zijn lineair onafhankelijk en spannen dus de hele ruimte op.
Deze eigenvectoren vormen een matrix $X = (x_{ij})$, waarbij we afspreken dat x_{ij} = i-de element van de eigenvector x_j. De volledige oplossing van het eigenprobleem kan dan worden geschreven in de vorm

$$AX = X\Lambda, \tag{1.1.5}$$

waarbij $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Omdat X niet-singulier is, mogen we ook schrijven

$$X^{-1}AX = \Lambda. \tag{1.1.6}$$

Uit stelling (1.1.4) volgt $Y^T X$ = diagonaal-matrix.
Deze diagonaal-matrix is niet-singulier en de eigenrijen kunnen dus zo genormeerd worden, dat $Y^T X = I$. Dan kan (1.1.6) worden geschreven als

$$Y^T AX = \Lambda. \tag{1.1.7}$$

Gelijkvormigheids-transformaties

Een transformatie die, bij gegeven niet-singuliere X, matrices A overvoert in $X^{-1}AX$ heet gelijkvormigheids-transformatie. De matrix X heet transformerende matrix. Twee matrices A en B heten gelijkvormig, als er een gelijkvormigheids-transformatie is, die A in B overvoert. Is A gelijkvormig met een diagonaal-matrix, dan heet A diagonaliseerbaar. Blijkens het bovenstaande hebben we

(1.1.8) Stelling. Zijn alle eigenwaarden van A verschillend, dan is A diagonaliseerbaar.

Het begrip "gelijkvormig" is een equivalentie-relatie (reflexief, symmetrisch en transitief). Gelijkvormigheids-transformaties laten de eigenwaarden invariant, d.w.z. gelijkvormige matrices hebben dezelfde eigenwaarden.

Meervoudige eigenwaarden

Zijn x_1 en x_2 eigenvectoren bij één eigenwaarde λ, dan ook elke lineaire combinatie $k_1x_1 + k_2x_2$. De eigenvectoren bij een eigenwaarde spannen dus een lineaire deel-ruimte op, de zgn. eigenruimte van λ.
De eigenruimte van een enkelvoudige eigenwaarde is altijd één-dimensionaal. Bij een meervoudige eigenwaarde kan de eigenruimte een hogere dimensie hebben. Voor een eigenwaarde met multipliciteit m geldt namelijk: $1 \leq$ dimensie eigenruimte $\leq m$.
Is voor alle eigenwaarden van A de dimensie van de eigenruimte gelijk aan de multipliciteit, dan zijn er blijkbaar n lineair onafhankelijke eigenvectoren en de matrix A is dan dus diagonaliseerbaar. Zijn er een of meer eigenwaarden, waarvoor de dimensie van de eigenruimte kleiner dan de multipliciteit is, dan zijn er niet n lineair onafhankelijke eigenvectoren en A is dan niet diagonaliseerbaar. In dit geval heet A ook defect. Is voor een of meer eigenwaarden de dimensie van de eigenruimte > 1, dan heet de matrix derogatoir. Een matrix met uitsluitend enkelvoudige eigenwaarden is dus niet derogatoir.

De canonieke vorm van Jordan

Een canonieke vorm van A is een speciale met A gelijkvormige matrix.
Een willekeurige n×n-matrix A hoeft niet gelijkvormig te zijn met een diagonaal-matrix, maar is wel gelijkvormig met een matrix van de volgende gedaante, de canonieke vorm van Jordan:

$$(1.1.9) \qquad J = \begin{pmatrix} J_{i_1}(\lambda_1) & & & \\ & J_{i_2}(\lambda_2) & & \\ & & \ddots & \\ & & & J_{i_r}(\lambda_r) \end{pmatrix}$$

Hierbij is de "Jordan-kast" $J_{i_k}(\lambda_k)$ een matrix van de orde i_k met λ_k op de hoofddiagonaal en enen op de daarboven liggende nevendiagonaal:

$$(1.1.10) \qquad J_{i_k}(\lambda_k) = \begin{pmatrix} \lambda_k & 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \dots & \dots & 0 & \lambda_k \end{pmatrix}, \quad k = 1(1)r,$$

en $\lambda_1, \dots, \lambda_r$ zijn de eigenwaarden van A. Onder deze λ_k mogen gelijken voorkomen, d.w.z. bij één eigenwaarde mag meer dan een Jordan-kast optreden.

De som van de ordes der Jordan-kasten bij één eigenwaarde is gelijk aan de multipliciteit, zodat J netjes van de orde n is.

J heet ook wel de <u>directe som</u> der langs de hoofddiagonaal geordende Jordan-kasten, genoteerd als

$$(1.1.11) \qquad J = J_{i_1}(\lambda_1) \dotplus J_{i_2}(\lambda_2) \dotplus \dots \dotplus J_{i_r}(\lambda_r).$$

Afgezien van permutaties van deze directe "sommanden" is de Jordan-canonieke vorm van een matrix eenduidig bepaald.

De karakteristieke polynomen der Jordan-kasten heten de <u>elementair-delers</u> van A. Als A diagonaliseerbaar is, zijn deze elementair-delers dus lineair.

De elementen 1 op de nevendiagonaal zijn niet essentiëel. Door voor- en na-vermenigvuldiging met diagonaal-matrices D resp. D^{-1} zijn deze te vervangen door willekeurige elementen $\neq 0$.

Companian matrix

Laat het karakteristieke polynoom $\det(\lambda I - A)$ van A in expliciete vorm luiden

$$\lambda^n - p_{n-1}\lambda^{n-1} - \ldots - p_1\lambda - p_0. \tag{1.1.12}$$

Men kan gemakkelijk bewijzen, dat dit tevens het karakteristieke polynoom is van de matrix

$$C = \begin{pmatrix} p_{n-1} & \cdots & p_1 & p_0 \\ 1 & & & 0 \\ & \ddots & & \vdots \\ 0 & & 1 & 0 \end{pmatrix} \tag{1.1.13}$$

Deze matrix heet de companian matrix van A. C heeft blijkbaar dezelfde eigenwaarden als A, maar A is dan en slechts dan gelijkvormig met C, als A niet-derogatoir is.

De rationale canonieke vorm van Frobenius

Een willekeurige n×n-matrix A is gelijkvormig met de canonieke vorm van Frobenius, zijnde een directe som van companian matrices:

$$F = F_1 \dot{+} F_2 \dot{+} \ldots \dot{+} F_s, \tag{1.1.14}$$

waarbij het karakteristieke polynoom van F_k een deler is van het karakteristieke polynoom van F_{k-1} (k = 2(1)s). Als A niet-derogatoir is, geldt s = 1 en $F = F_1$ is de companian matrix van A. De relatie met de Jordan-vorm is als volgt. F_1 is gelijkvormig met de directe som van enige Jordan-kasten; van elke verschillende eigenwaarde treedt hierin precies één Jordan-kast van hoogste orde op. Een met F_2 gelijkvormige matrix wordt op dezelfde wijze gevormd uit de resterende directe som van Jordan-kasten, enz. De Frobenius-canonieke vorm heet ook rationale canonieke vorm, omdat voor matrices met rationale elementen deze met rationale gelijkvormigheids-transformaties te bereiken is.

Andere speciale vormen

Iedere matrix is gelijkvormig met een driehoeks-matrix. Op de hoofddiagonaal hiervan staan de eigenwaarden. Berekening van deze driehoeksvorm kan dus alleen maar door een iteratief proces geschieden. Een andere speciale vorm, die, evenals de driehoeks-vorm, de Jordan-canonieke vorm als speciaal geval bevat, is de tri-diagonale vorm, d.w.z. alleen de hoofddiagonaal en de aanliggende nevendiagonalen zijn niet nul. Deze vorm is een belangrijk tussen-stadium van sommige numerieke eigenwaarde-bepalingen. Een nog minder speciale vorm is de bijna-driehoeks-vorm of Hessenberg-vorm, waarin de niet-nul elementen staan in een driehoek en de aanliggende nevendiagonaal. Deze vorm is van belang als tussen-stadium bij het berekenen van eigenwaarden van asymmetrische matrices.

Principale vectoren

Bij een defecte matrix spannen de eigenvectoren niet de hele ruimte op. Daarom is het gewenst het begrip "eigenvector" uit te breiden tot een iets ruimer begrip. Als λ_i een eigenwaarde is, waarvan de eigenruimte een dimensie heeft kleiner dan de multipliciteit van λ_i, dan definiëert men bij λ_i principale vectoren van de graad j als vectoren $x_k \neq 0$, die voldoen aan

$$(A - \lambda_i I)^j x_k = 0, \tag{1.1.15}$$

waarbij j minimaal is. Vult men de eigenvectoren met deze principale vectoren aan, dan kan de hele ruimte opgespannen worden. In het bijzonder voor de Jordan-canonieke vorm geldt, dat de eenheids-vectoren $e_1, \ldots, e_n$ een volledig stelsel principale vectoren vormen.

Speciale matrices

Van groot belang in de praktijk zijn de Hermitische matrices, dat zijn matrices A, die gelijk zijn aan hun toegevoegd complex getransponeerde A^*. Zij hebben de volgende prettige eigenschappen (die samen nodig en voldoende zijn voor Hermiticiteit):

1) alle eigenwaarden zijn reëel
2) Hermitische matrices zijn diagonaliseerbaar
3) de eigenvectoren kunnen zo gekozen worden, dat zij een unitaire matrix vormen (een matrix U is unitair als $U^{-1} = U^{*}$).

Deze eigenschappen maken, dat het eigenwaarde-probleem numeriek eenvoudiger ligt. Het eigenwaarde-probleem van een Hermitische matrix is altijd goed geconditioneerd, d.w.z. kleine storingen in de matrix geven kleine variaties in de eigenwaarden. Deze storingen kunnen echter wel grote afwijkingen in de eigenvectoren veroorzaken, met name als een of meer eigenwaarden dicht bij elkaar liggen. (Dan kan een storing immers bewerken, dat twee eigenwaarden samenvallen, waardoor de eigenvectoren zelfs onbepaald worden.)
Bij de behandeling van numerieke methoden beperken we ons tot reële matrices. Reëel Hermitisch is blijkbaar hetzelfde als reëel symmetrisch. De matrix der eigenvectoren kan dan reëel unitair, dat is dus reëel orthogonaal, gekozen worden.

Een andere belangrijke klasse van matrices, de zgn. normale matrices, ontstaat, als we van bovengenoemde drie eigenschappen de eerste laten vervallen, m.a.w. we laten complexe eigenwaarden toe. Een normale matrix A is, wegens (1.1.5) na-vermenigvuldigd met $X^{-1} = X^{*}$, te schrijven als

$$A = X\Lambda X^{*}, \tag{1.1.16}$$

waarbij X dus unitair is. Vermenigvuldigen we A links of rechts met A^{*} dan volgt hieruit

$$AA^{*} = X\Lambda\Lambda^{*}X^{*} = A^{*}A, \tag{1.1.17}$$

want $\Lambda\Lambda^{*} = \Lambda^{*}\Lambda$.
Eigenschap (1.1.17) is nodig en voldoende voor normaliteit. Hebben alle eigenwaarden een absolute waarde = 1, dan geldt $\Lambda\Lambda^{*} = I$ en dus ook $AA^{*} = I$, m.a.w. A is dan een unitaire matrix.
Ook voor normale matrices is het eigen-probleem goed geconditioneerd. Bij niet-normale matrices kunnen daarentegen de eigenwaarden zeer ge-

voelig zijn voor kleine storingen in de matrix-elementen. Dit is met name het geval, als eigenvectoren (bij verschillende eigenwaarden) een zeer kleine hoek maken, m.a.w. als de matrix der eigenvectoren verre van unitair is. Maken we een limiet-overgang, waarbij we eigenvectoren laten samenvallen, dan ontstaat een defecte matrix. Defecte matrices zijn slecht geconditioneerd en de oplossing van het eigenwaarde-probleem is vaak niet eenvoudig, zeker als ook de eigenvectoren en de principale vectoren gevraagd worden.

1.2 Vector- en matrix-normen

Vector-normen

Een norm in een lineaire vector-ruimte R over het lichaam der reële of complexe getallen is een functie, die aan elke vector $x \in R$ toevoegt een reëel getal $||x||$ met de eigenschappen

1) $||x|| \geq 0$
2) $||x|| = 0 \iff x = \vec{0}$
3) $||x + y|| \leq ||x|| + ||y||$ (driehoeks-ongelijkheid)
4) $||\alpha x|| = |\alpha| \times ||x||$ (homogeniteit)

Voor een n-dimensionale vector-ruimte gebruiken we de volgende normen

$$(1.2.0) \qquad ||x||_p = \sqrt[p]{|x_1|^p + |x_2|^p + \ldots + |x_n|^p} \; .$$

Om aan de driehoeks-ongelijkheid te voldoen moet p minstens 1 zijn. In de praktijk gebruiken we $p = 1$, 2, of ∞. Voor $p = 2$ hebben we de inwendig-product-norm $||x||_2 = \sqrt{x^* x}$; voor $p = \infty$ krijgen we na limiet-overgang

$$||x||_\infty = \max_{k=1,\ldots,n} |x_k| \; .$$

Matrix-normen

Een n×n-matrix kunnen we beschouwen als een vector in een n^2-dimensionale ruimte. Een matrix-norm moet dan ook aan de bovengenoemde eisen voldoen.

Hieraan willen we voor matrix-normen graag de volgende eigenschappen toevoegen

5) $||AB|| \leq ||A|| \times ||B||$ multiplicativiteit

6) $||Ax|| \leq ||A|| \times ||x||$ consistentie m.b.t. $||x||$,

waarbij $||x||$ een of andere vector-norm is.

Formule (1.2.0) wordt voor matrices alleen toegepast met p = 2, wat de zgn. Euclidische norm of Schur-norm levert:

$$(1.2.1) \qquad ||A||_E = \sqrt{\sum_{i,j=1}^{n} |A_{ij}|^2} \ .$$

Andere matrix-normen krijgt men, door bij een gegeven vector-norm $||x||_p$ een minimale matrix-norm te zoeken, die nog aan eigenschap (6) voldoet:

$$(1.2.2) \qquad ||A||_p = \sup_{x \neq \vec{0}} \frac{||Ax||_p}{||x||_p} \ .$$

Deze norm heet de met $||x||_p$ geassocieerde matrix-norm. Voor p = 1,2,∞ zijn deze expliciet te schrijven als volgt

$$(1.2.3) \qquad ||A||_1 = \max_j \sum_i |A_{ij}|$$

$$(1.2.4) \qquad ||A||_\infty = \max_i \sum_j |A_{ij}|$$

$$(1.2.5) \qquad ||A||_2 = \sqrt{\lambda_{max}\,(A^*A)},$$

als λ_{max} de grootste eigenwaarde aanduidt.

Hierom heet $||A||_2$ ook wel de spectrale norm van A. Matrix-normen zijn van belang, omdat ze (wegens eigenschap 6) een bovengrens leveren voor de modulus van de eigenwaarden. De normen $||A||_E$, $||A||_1$ en $||A||_\infty$ zijn gemakkelijk te berekenen en blijven invariant, als alle elementen van A door hun absolute waarde vervangen worden. (Deze matrix wordt genoteerd als $|A|$.)

Dit geldt niet voor $||A||_2$, die theoretisch belangrijk is en ook vaak de scherpste bovengrens voor de eigenwaarden geeft.

Hier volgen nog enige belangrijke relaties.

(1.2.6) $$||A||_2 \leq ||A||_E \leq \sqrt{n}\, ||A||_2$$

en dus ook

(1.2.7) $$||\,|A|\,||_2 \leq ||\,|A|\,||_E = ||A||_E \leq n\, ||A||_2$$

(1.2.8) $$||A||_2^2 \leq ||A^*A||_1 \leq ||A^*||_1\, ||A||_1 = ||A||_\infty\, ||A||_1.$$

Als A normaal is geldt

(1.2.9) $$||A||_2 = |\lambda(A)|_{max}.$$

1.3 Methode van Jacobi voor reële symmetrische matrices

Een reële symmetrische matrix A van de orde n heeft reële eigenwaarden en n lineair onafhankelijke orthogonale eigenvectoren. Zij kan dus worden geschreven als

(1.3.1) $$A = X\Lambda X^T,$$

waarbij Λ een diagonaal-matrix is met de eigenwaarden van A op de hoofddiagonaal en X de orthogonale matrix der eigenvectoren. In de methode van Jacobi worden Λ en X iteratief bepaald als volgt. Uitgaande van de matrix $A = A^{(0)}$, wordt in elke stap een orthogonale gelijkvormigheids-transformatie uitgevoerd, in formule:

(1.3.2) $$A^{(k+1)} = R^{(k)T}\, A^{(k)}\, R^{(k)}.$$

Hierbij is $R = R^{(k)}$ een speciale orthogonale matrix, nl. een vlakke rotatie. Een rotatie in het (p,q)-vlak over een hoek θ is gedefiniëerd door

(1.3.3) $$\begin{cases} R_{pp} = R_{qq} = \cos\theta, \; R_{qp} = -R_{pq} = \sin\theta, \\ R_{ii} = 1 \; (i \neq p,q), \; R_{ij} = 0 \text{ voor alle andere } i,j. \end{cases}$$

De transformatie (1.3.2) laat, wegens de orthogonaliteit van R, de symmetrie intact en wijzigt alleen de p-de en q-de rij en kolom van $A^{(k)}$ en wel

$$(1.3.4)\quad \begin{cases} a_{ip}^{(k+1)} = a_{ip}^{(k)} \cos\theta + a_{iq}^{(k)} \sin\theta = a_{pi}^{(k+1)} \\ a_{iq}^{(k+1)} = -a_{ip}^{(k)} \sin\theta + a_{iq}^{(k)} \cos\theta = a_{qi}^{(k+1)} \end{cases} \qquad i \neq p,q$$

$$(1.3.5)\quad \begin{cases} a_{pp}^{(k+1)} = a_{pp}^{(k)} \cos^2\theta + a_{pq}^{(k)} \sin 2\theta + a_{qq}^{(k)} \sin^2\theta \\ a_{qq}^{(k+1)} = a_{pp}^{(k)} \sin^2\theta - a_{pq}^{(k)} \sin 2\theta + a_{qq}^{(k)} \cos^2\theta \end{cases}$$

$$(1.3.6)\quad a_{pq}^{(k+1)} = (a_{qq}^{(k)} - a_{pp}^{(k)})\cos\theta \sin\theta + a_{pq}^{(k)}(\cos^2\theta - \sin^2\theta) =$$

$$= a_{qp}^{(k+1)}.$$

De draaiingshoek θ wordt zó gekozen, dat $a_{pq}^{(k+1)} = 0$, dus

$$(1.3.7)\qquad \tan 2\theta = 2\, a_{pq}^{(k)} / (a_{pp}^{(k)} - a_{qq}^{(k)}).$$

Het doel van de iteratie is te verkrijgen een met A gelijkvormige matrix $A^{(m)}$, die in voldoende precisie gelijk is aan een diagonaalmatrix. Dan staan de eigenwaarden op de hoofddiagonaal van $A^{(m)}$ en de matrix der eigenvectoren is

$$(1.3.8)\qquad X \approx R^{(0)}\, R^{(1)} \ldots R^{(m-1)}.$$

Convergentie-onderzoek

Als maat voor de bereikte precisie dient de Euclidische norm van de matrix $E^{(k)}$ der niet-diagonale elementen, dus

$$(1.3.9)\quad \begin{cases} E^{(k)} = A^{(k)} - \mathrm{diag}\,(a_{ii}^{(k)}) \\ \| E^{(k)} \|_E = \sqrt{\sum_{i \neq j} (a_{ij}^{(k)})^2} \end{cases}$$

Uit (1.3.3) volgt voor $i \neq p,q$

$$(a_{ip}^{(k+1)})^2 + (a_{iq}^{(k+1)})^2 = (a_{ip}^{(k)})^2 + (a_{iq}^{(k)})^2.$$

Dus de enige verandering in de Euclidische norm der niet-diagonale elementen wordt teweeg gebracht door de wijziging van het (p,q)-de element.
Het grootste effect in een stap wordt bereikt, door dit element nul te maken, vandaar de keuze van θ volgens (1.3.7). Dan geldt

$$||E^{(k+1)}||_E^2 = ||E^{(k)}||_E^2 - 2(a_{pq}^{(k)})^2. \tag{1.3.10}$$

Wat betreft de keuze van het draaiingsvlak (p,q) bestaan de volgende drie strategieën

a) Klassieke Jacobi. In elke stap worden p en q gekozen zó, dat $|a_{pq}^{(k)}|$ maximaal is. Deze strategie heeft het nadeel, dat het zoeken van zo'n maximaal element te veel tijd kost (aantal operaties is evenredig met $\frac{1}{2}n(n-1)$, terwijl het aantal operaties, nodig voor de transformatie, evenredig is met n).

b) Serieele Jacobi. De elementen $a_{pq}^{(k)}$ worden afgewerkt in een vaste volgorde. Deze taktiek heeft het nadeel, dat de convergentie te langzaam is. Als $|a_{pq}^{(k)}|$ veel kleiner is, dan andere niet-diagonale elementen, heeft het weinig zin hieraan de moeite van een transformatie te besteden.

c) Drempel-strategie. De elementen $a_{pq}^{(k)}$ worden in een vaste volgorde afgewerkt, maar de transformatie wordt alleen uitgevoerd, als $|a_{pq}^{(k)}|$ boven een zekere drempel d ligt. Wegens (1.3.10) liggen na een eindig aantal stappen alle niet-diagonale elementen onder de drempel. Zodra dit het geval is, wordt de drempel verlaagd en het proces herhaald, totdat een drempel bereikt wordt, waarmee voldoende precisie overeenkomt.

Jacobi met drempel-strategie convergeert dus altijd. Als alle eigenwaarden verschillend zijn, dan is de convergentie op den duur kwadratisch, precieser: voor voldoende grote k geldt:

$$||E^{(k+N)}||_E \leq \text{constante} \times ||E^{(k)}||_E^2,$$

waarbij $N = \frac{1}{2}n(n-1)$, d.w.z. alle elementen (boven de drempel) zijn minstens eenmaal aan bod gekomen.

Voor- en nadelen van Jacobi's methode

De methode van Jacobi is betrekkelijk gemakkelijk te programmeren en is redelijk snel. De methode kan de eigenwaarden in een behoorlijke precisie leveren en de berekende eigenvectoren zijn ook netjes (in voldoende precisie) orthogonaal. Wanneer echter enige eigenwaarden dicht bij elkaar liggen, worden de bijbehorende eigenvectoren onbetrouwbaar. Dit is weliswaar een essentiële moeilijkheid van het probleem, maar andere methodes kunnen in dit opzicht beter zijn. Bovendien zijn er methodes, die sneller zijn dan Jacobi.

1.4 Reductie tot gelijkvormige bijna-driehoekige of tridiagonale matrix

De methode van Jacobi is essentiëel iteratief. Andere methodes beginnen met een iets minder volledige reductie van de gegeven matrix, bereikbaar door een direct, dat is niet-iteratief, proces, bestaande uit een, bij gegeven orde, vast aantal speciale gelijkvormigheids-transformaties. De resulterende gereduceerde vorm is in het algemene geval de bijna-driehoeksvorm. Voor reële symmetrische matrices gebruikt men, teneinde de symmetrie te handhaven, altijd orthogonale matrices, zodat de gereduceerde vorm dan dus symmetrisch tridiagonaal is. In Given's transformatie zijn de transformerende matrices vlakke rotaties. Elke rotatie introduceert één nul, zodat in totaal $\frac{1}{2}(n-1)(n-2)$ vlakke rotaties nodig zijn. Een ander, dubbel zo snel, procédé, uitgevonden door Householder en ontwikkeld door Wilkinson, zullen we hier in meer detail bespreken.

Householder's transformatie

Uitgaande van een reële matrix $A = A^{(1)}$ wordt een gelijkvormige bijna-bovendriehoeksvorm-matrix H (d.w.z. $H_{ij} = 0$ voor $i > j + 1$) verkregen door n-2 transformaties van de vorm

(1.4.0) $$A^{(r+1)} = P^{(r)}A^{(r)}P^{(r)}, \quad r = 1(1)n-2,$$

waarbij $P^{(r)}$ een orthogonale symmetrische matrix is van de gedaante

(1.4.1) $$P^{(r)} = I + k_r u_r u_r^T.$$

Hierin is k_r een scalar en u_r een kolom-vector, waarvan de eerste r elementen nul zijn, m.a.w.

(1.4.2) $$u_r^T = (0,\ldots, 0, u_{r+1,r},\ldots, u_{nr}).$$

De bedoeling is dat de r-de transformatie de gewenste nullen introduceert in de r-de kolom van de matrix $A^{(r)}$, zodat na n-2 stappen de gewenste bijna-driehoeksvorm ontstaat. Door al deze voorwaarden ligt de transformatie vrijwel eenduidig vast.
Ten eerste moet $P^{(r)}$ orthogonaal zijn (symmetrie is vanzelfsprekend), dus $(P^{(r)})^2 = I$, waaruit onmiddellijk volgt

(1.4.3) $$k_r u_r^T u_r = -2 \text{ of } 0,$$

waarbij de waarde 0 leidt tot het triviale geval $P^{(r)} = I$. Vanwege de nullen in de vector u en vanwege de reeds geïntroduceerde nullen in de eerste r-1 kolommen van de matrix blijven deze r-1 kolommen invariant. De r-de kolom wordt alleen door de linker factor $P^{(r)}$ gewijzigd, zodat we hiervoor hebben

(1.4.4) $$A_{ir}^{(r+1)} = (P^{(r)}A^{(r)})_{ir} = A_{ir}^{(r)} + k_r u_{ir}(u_r^T A^{(r)})_r$$

$$= A_{ir}^{(r)} + u_{ir}q_r,$$

waarbij

(1.4.5) $$q_r = k_r \sum_{i=r+1}^{n} u_{ir}A_{ir}^{(r)}.$$

Om de gewenste nullen te krijgen moet dus gelden

(1.4.6) $$0 = A_{ir}^{(r)} + u_{ir}q_r, \quad i = r + 2(1)n.$$

Noemen we $A^{(r+1)}_{r+1,r} = b_r$ dan geldt hiervoor

$$b_r = A^{(r)}_{r+1,r} + u_{r+1,r}q_r. \tag{1.4.7}$$

Kwadrateren en sommeren van (1.4.6 & 7) levert, als $S = \sum_{i=r+1}^{n} (A^{(r)}_{ir})^2$:

$$b_r^2 = S + 2q_r \sum_{i=r+1}^{n} u_{ir}A^{(r)}_{ir} + q_r^2 \sum u_{ir}^2 = \tag{1.4.8}$$

$$= S + q_r(2 + k_r u_r^T u_r) \sum_{i=r+1}^{n} u_{ir}A^{(r)}_{ir} = S$$

wegens (1.4.5 & 3).

Als $S = 0$, dan moet men de transformatie onderdrukken, m.a.w. $P^{(r)} = I$ nemen. In het andere geval kan de vector u altijd zó genormeerd worden, dat $q_r = -1$. Het teken van b_r worde zo gekozen, dat bij het berekenen van $u_{r+1,r}$ (uit (1.4.7)) geen cijfers wegvallen. Tenslotte volgt k_r uit (1.4.6 & 7) door te vermenigvuldigen met $k_r u_{ir}$ en te sommeren:

$$k_r u_{r+1,r} b_r = k_r \sum_{i=r+1}^{n} u_{ir}A^{(r)}_{ir} + q_r k_r u_r^T u_r = -q_r = 1$$

wegens (1.4.5 & 3) en de gekozen normering van u. Samenvattend hebben we dan (voor $S \neq 0$):

$$\begin{cases} b_r = -\operatorname{sgn}(A^{(r)}_{r+1,r})\sqrt{S} \\ u_{ir} = A^{(r)}_{ir}, \quad i = r + 2(1)n \\ u_{r+1,r} = A^{(r)}_{r+1,r} - b_r \\ k_r = 1/(u_{r+1,r}b_r). \end{cases} \tag{1.4.9}$$

Hiermee is de transformatie volkomen bepaald.

Het bovengenoemde uitzonderingsgeval $S = 0$ is onvermijdelijk. In de praktijk onderdrukt men de transformatie, als $\sqrt{S}$ onder een zekere drempel ligt, bv. $||A||_\infty \times$ machine-precisie.

Andere transformaties

Voor reële symmetrische matrices is Householder's transformatie, een factor 2 sneller zijnde dan Givens, wel optimaal. Omdat de transformatie orthogonaal is, blijft de symmetrie intact en er ontstaat dus een symmetrische tridiagonaal-vorm. Deze vorm is zeer handzaam voor de berekening van de eigenwaarden zowel als de eigenvectoren.
Bij asymmetrische matrices zijn we niet gebonden aan orthogonale transformaties. Er bestaat voor het asymmetrische geval een transformatie, die eveneens een bijna-driehoeksvorm levert en wederom een factor 2 sneller is dan Householder.
De nullen worden geïntroduceerd door Gauss-achtige eliminaties, waarbij, terwille van de numerieke stabiliteit, rij-verwisseling en overeenkomstige kolom-wisseling nodig zijn om een niet te klein element als pivot te kunnen gebruiken. Afgezien van de verwisselingen komt dit proces hierop neer, dat de transformerende matrix een driehoeksmatrix is.

Verdere reductie

Men kan een bijna-driehoekige matrix nog verder reduceren, eveneens met een direct proces, tot een Frobenius-vorm of tot een tridiagonale vorm. Als alle elementen van de nevendiagonaal niet-nul zijn, is de matrix niet-derogatoir en de Frobenius-vorm is dan dus de companian-matrix. Reductie tot Frobenius-vorm komt dus neer op berekenen van de coëfficiënten van het karakteristieke polynoom. Dit proces is numeriek vaak zeer instabiel; kleine storingen in de coëfficiënten van het karakteristieke polynoom kunnen grote veranderingen in de eigenwaarden teweeg brengen. De companian-matrix kan dan ook veel slechter geconditioneerd zijn, dan de oorspronkelijke matrix.
Reductie tot tridiagonaal-vorm is minder stabiel (en kan bij uitzondering vast lopen) dan de reductie tot Hessenberg-vorm, maar is meestal wel bruikbaar en veel beter dan reductie tot Frobenius-vorm.
Men kan de tridiagonaal-vorm verkrijgen door de Gauss-achtige eliminaties voort te zetten, waarbij nu echter geen verwisselingen

meer mogelijk zijn. Omdat men kleine pivots kan tegenkomen, is het aan te bevelen deze verdere reductie in dubbele lengte uit te voeren. Een hiermee samenhangend proces is dat van Lanczos, dat eveneens een tridiagonaal-vorm levert.

Conditie van een matrix m.b.t. eigenwaarde-probleem

Bij een matrix A kan men definiëren een conditie-getal (condition number) m.b.t. matrix-inversie:

(1.4.10) $$k(A) = ||A^{-1}|| \cdot ||A||,$$

waarbij $||A||$ een van de in (1.2) genoemde matrix-normen is. Is X de matrix der eigenvectoren van een matrix A, dan heet k(X) conditie-getal m.b.t. het eigenwaarde-probleem van A. Dit in verband met de volgende

(1.4.11) Stelling (van Bauer en Fike, [1] p. 87 en [3]):
Zij A een diagonaliseerbare matrix met eigenvectoren-matrix X, m.a.w.

$$X^{-1}AX = \mathrm{diag}(\lambda_i)$$

en zij λ een eigenwaarde van $A + \varepsilon B$, waarbij εB een (kleine) storingsmatrix is. Dan geldt:

$$\min |\lambda_i - \lambda| \leq \varepsilon ||B|| \cdot k(X).$$

In het bijzonder, als A een normale matrix is, dan is X unitair en het spectrale conditie-getal $k_2(X)$ (horende bij de spectrale norm van X) is dan dus gelijk aan 1. Bovengenoemde stelling levert dan

$$\min (\lambda_i - \lambda) \leq \varepsilon ||B||_2,$$

m.a.w. een kleine storing van de matrix brengt slechts een kleine wijziging in de eigenwaarden teweeg.

1.5 Berekening der eigenwaarden

Symmetrische matrices

In het reële symmetrische geval nemen we aan, dat we m.b.v. Householder's transformatie een symmetrische tridiagonale matrix verkregen hebben, die we noteren als

(1.5.0) $$S_n = \begin{pmatrix} a_1 & b_1 & 0 & \cdots & 0 \\ b_1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_{n-1} \\ 0 & \cdots & 0 & b_{n-1} & a_n \end{pmatrix}$$

We definiëren voor i = 0(1)n:

(1.5.1) $$f_i(\lambda) = \det(\lambda I_i - S_i),$$

waarbij I_i de eenheidsmatrix van de orde i is. Deze voldoen aan de recurrente betrekking:

(1.5.2) $$f_i(\lambda) = (\lambda - a_i)f_{i-1}(\lambda) - b_{i-1}^2 f_{i-2}(\lambda), \quad i = 2(1)n$$

met als start $f_0 = 1$ en $f_1(\lambda) = \lambda - a_1$.
De eigenwaarden van S_n laten zich goed bepalen, dank zij de volgende

(1.5.3) Stelling (van Givens):
Als de nevendiagonaal van S_n geen nullen bevat, dan vormt de rij functies f_i, i = 0(1)n, een Sturm-rij. Het aantal eigenwaarden van S_n groter dan a is gelijk aan het aantal tekenwisselingen in de rij $\{f_i(a)\}$, mits natuurlijk $f_n(a) \neq 0$. De nulpunten van elke f_i zijn verschillend en worden gescheiden door die van f_{i-1}.
Op grond van deze stelling kan men met behulp van bisectie elke eigenwaarde localiseren. Heeft men eenmaal een interval gevonden, waarin blijkens de tekenwisselingen precies één eigenwaarde ligt, dan kan men de iteratie aanzienlijk versnellen, door over te gaan op lineaire interpolatie (regula falsi).

Asymmetrische matrices

We nemen aan dat de matrix is getransformeerd in een gelijkvormige bijna-driehoeksmatrix

(1.5.4) $$H = \begin{pmatrix} h_{11} & \cdots & h_{1n} \\ b_1 & \ddots & \vdots \\ 0 & \ddots & \ddots \\ 0 & 0 & b_{n-1} \quad h_{nn} \end{pmatrix}$$

Voor een willekeurig getal λ kan de waarde van de karakteristieke determinant verkregen worden met Hyman's methode, mits de nevendiagonaalelementen b_i niet-nul zijn. In deze methode wordt getracht het stelsel $(\lambda I - H)x = 0$ op te lossen door uitgaande van $x_n = 1$ achtereenvolgens x_{i-1} te berekenen uit de i-de vergelijking voor $i = n(-1)2$, dus

(1.5.5) $$x_{i-1} = (\lambda x_i - \sum_{j=i}^{n} h_{ij} x_j)/b_{i-1}.$$

De determinant van $\lambda I - H$ blijft invariant, als de laatste kolom wordt vervangen door de vector $(\lambda I - H)x$. Aangezien van deze vector alleen het eerste element niet-nul is vinden we voor de karakteristieke determinant

(1.5.6) $$\det(\lambda I - H) = (\lambda x_1 - \sum_{j=1}^{n} h_{ij} x_j) \times b_1 \times \ldots \times b_{n-1}.$$

Merkwaardigerwijs is deze formule ook bruikbaar (althans voor het localiseren der nulpunten), als op de nevendiagonaal zeer dicht bij 0 liggende elementen b_i voorkomen. Deze formule stelt ons in staat de eigenwaarden iteratief te bepalen door middel van lineaire interpolatie (vooral voor reële eigenwaarden) of inverse kwadratische interpolatie (formule van Muller of van Traub), welke laatste het voordeel heeft, dat men, bij reële start, vanzelf in het complexe

vlak duikt en daardoor ook niet-reële eigenwaarden kan vinden. Voor de afgeleiden van de karakteristieke veelterm gelden soortgelijke formules, waarmee iteratie-processen zoals Newton's en Laguerre's formule toepasbaar worden. Ook de methode van Bairstow, bekend als methode voor expliciete polynomen, laat zich redelijk voor de karakteristieke veelterm van H formuleren.
Heeft men enige eigenwaarden λ_i, $i < k$ gevonden, dan deelt men, voor het vinden van verdere eigenwaarden, de karakteristieke veelterm door het product $\prod_{i<k} (\lambda - \lambda_i)$. Natuurlijk moet men dan wel oppassen, dat de iterates λ niet te dicht bij reeds gevonden eigenwaarden λ_i komen. Dit is een essentiële moeilijkheid bij deze methodes.
Ook afgezien hiervan kan convergentie helaas niet gegarandeerd worden, ofschoon het in de praktijk vaak wel lukt, mits men zorgt, dat de iterates binnen een cirkel (bv. met straal $||A||$) blijven.

De QR-algorithme

Deflatie wil zeggen, dat na het vinden van een eigenwaarde het probleem wordt herleid tot een lagere orde. Het voordeel is, dat reeds gevonden eigenwaarden niet meer plagen; integendeel, wegens de ordeverlaging verloopt het proces steeds sneller. Een belangrijke deflerende methode is de QR-transformatie. In een QR-stap wordt een gegeven matrix $H^{(k)}$ geschreven als het product van een unitaire matrix Q en een boven-driehoeksmatrix R. Deze ontbinding is meestal eenduidig.
Het omgekeerde product RQ is dan de volgende iterate $H^{(k+1)}$. Wordt als uitgangspunt genomen een Hessenberg-matrix $H = H^{(0)}$, dan hebben de matrices Q ook de Hessenberg-vorm.
De convergentie is lineair met convergentie-factor $\left|\frac{\lambda_n}{\lambda_{n-1}}\right|$, waarbij $|\lambda_n|$ de kleinste en $|\lambda_{n-1}|$ de op een na kleinste absolute waarde der eigenwaarden is. De convergentie verloopt dus aanzienlijk sneller, als men een shift kan vinden die λ_n nagenoeg 0 maakt.
Een QR-stap met shift ζ_k luidt als volgt:

$$H^{(k)} - \zeta_k I = QR$$

(1.5.7)

$$RQ + \zeta_k I = H^{(k+1)}.$$

Als shift kan men nemen het element $h_{nn}^{(k)}$ of de hier dichtstbij gelegen eigenwaarde van de 2×2-matrix in de rechter-onder-hoek van $H^{(k)}$. De convergentie is dan kwadratisch, mits de eigenwaarden enkelvoudig zijn. Als het nevendiagonaal-element $b_{n-1}^{(k)}$ klein genoeg is, kan $h_{nn}^{(k)}$ als benadering van een eigenwaarde worden afgeleverd en de berekening worden voortgezet na het schrappen van laatste rij en kolom. Een belangrijke eigenschap van de QR transformatie is, dat, wegens het unitair zijn van Q, de conditie $k_2(X^{(k)})$ m.b.t. het eigenwaarde probleem van de matrices $H^{(k)}$ invariant blijft (waarbij $X^{(k)}$ de matrix der eigenvectoren van $H^{(k)}$ is).

1.6 Berekening der eigenvectoren

Heeft men eenmaal een schatting van een eigenwaarde λ_i, dan kan een bijbehorende eigenvector zeer goed verkregen worden door inverse iteratie, d.w.z. uitgaande van een schatting $v^{(0)}$ berekent men nieuwe schattingen $v^{(k)}$ uit

(1.6.0) $$(A - \lambda_i I)v^{(k+1)} = v^{(k)}.$$

Men kan dit stelsel oplossen met Gauss' eliminatie met verwisselingen. Deze verwisselingen zijn ook nodig, als A tridiagonaal of bijna driehoekig is.
Merkwaardigerwijs geldt: hoe beter de benadering λ_i is, dus hoe meer singulier $A - \lambda_i I$ is, des te sneller convergeert dit en vaak zijn 1 of 2 stappen voldoende. Een maat voor de precisie van $v^{(k+1)}$ is: $||v^{(k)}||/||v^{(k+1)}||$. Na de inverse iteratie moet men uiteraard, om een eigenvector van de oorspronkelijke matrix te vinden, nog terug transformeren. De formule hiervoor volgt gemakkelijk uit de transformatieformule (vgl. (1.4.0)).

1.7 Iteratieve methodes

Hier zij vermeld de matrix-maal-vector iteratie (power method), eventueel voorafgegaan door matrix-kwadratering. Deze methode convergeert slechts lineair en is daardoor alleen te verkiezen, wanneer enkele van de grootste of de kleinste eigenwaarden verlangd worden. De methode kan aantrekkelijk zijn voor grote matrices met overwegend nullen.
Een andere iteratieve methode, die volgens Wilkinson veelbelovend is, is die van Eberlein [4]. Deze methode lijkt op de Jacobi-iteratie, maar nu gebruikt men in plaats van vlakke rotaties niet-unitaire vlakke transformaties (d.w.z. de transformerende matrix verschilt in een diagonale 2×2-submatrix van de eenheidsmatrix). Deze vlakke transformaties worden zo gekozen, dat de geïtereerde matrices steeds "normaler" worden, d.w.z. dat de eigenvectoren-matrix nadert tot een unitaire matrix. Is de matrix voldoende normaal, dan zijn de verdere transformaties unitaire (in het algemeen niet-reële) vlakke rotaties en verloopt de convergentie naar diagonaal-vorm als bij Jacobi.

1.8 Literatuur

1. J.H. Wilkinson, The algebraic eigenvalue problem (Oxford 1965).

2. D.K. Faddeev & V.N. Faddeeva, Computational methods of linear algebra (1960, transl. from Russian 1963).

3. F.L. Bauer & C.T. Fike, Norms and exclusion theorems, Num. Mat. 2 (1960) 137-141.

4. P.J. Eberlein, J. SIAM 10 (1962) 74-88.

2. Keuze van Δt en Δx bij een diffusieprobleem

2.1 Inleiding

Een benadering van de oplossing van een probleem in differentiaalvorm wordt vaak verkregen door met behulp van een rekenmachine de oplossing te bepalen van een toegevoegd probleem in differentievorm. De wijze van toevoeging van het probleem in differentievorm aan dat in differentiaalvorm ligt niet bij voorbaat eenduidig vast. Als doel staat voor ogen om m.b.v. de rekenmachine snel en met geringe kosten een voldoend nauwkeurige benadering te vinden.
Om te kunnen aangeven òf een benadering voldoend nauwkeurig zal zijn, is voorkennis van het gegeven probleem zowel in differentiaalvorm als in differentievorm noodzakelijk. Het verschil in gedrag van de oplossingen van beide problemen moet gering zijn. Typerend voor het gedrag zullen vaak ook een aantal afgeleiden zijn. Voldoend nauwkeurig houdt dus in, dat het differentieprobleem de oplossing alsmede de significante afgeleiden van het differentiaalprobleem goed vertolkt.
De rekentijd, die nodig is om het differentieprobleem op te lossen, zal afhangen van de aard van het toegevoegde differentieprobleem alsmede van de methode van oplossen. Indien het bijvoorbeeld een iteratieve methode is, is de snelheid van convergentie van belang. Bij de toevoeging van het differentieprobleem gaat het niet alleen om de keuze van het type differentievergelijking, maar spelen ook een belangrijke rol andere vrijheidsgraden zoals: de toe te passen waarden van de differenties bij de onafhankelijke veranderlijken; de manier van aanpassen van de gegeven nevenvoorwaarden; de methode van oplossen.
We zullen deze aspecten nader bekijken voor het volgende (lineaire) ééndimensionale diffusieprobleem:

$$\text{(1a)} \qquad \begin{cases} \dfrac{\partial z}{\partial t} - \dfrac{\partial^2 z}{\partial x^2} - pz = 0 \\ z(0,t) = \phi(t),\ z(1,t) = \eta(t); \end{cases}$$

$$\text{(1b)} \qquad z(x,0) = \gamma(x).$$

Voor de keuze van de differentievergelijking bestaan vele mogelijkheden. We zullen ons voornamelijk beperken tot de volgende 1-parameterklasse van differentievergelijkingen:

$$(2)\quad \frac{\delta_{t-} z(x_m,t_n)}{\Delta t} - a\,\frac{\delta_x^2 z(x_m,t_n)}{(\Delta x)^2} - (1-a)\,\frac{\delta_x^2 z(x_m,t_{n-1})}{(\Delta x)^2} - apz(x_m,t_n) - (1-a)pz(x_m,t_{n-1}) = 0,$$

waarbij δ en δ_- worden gebruikt om resp. het centrale en het achterwaartse differentiequotiënt aan te duiden:

$$(3)\quad \frac{\delta_x^2 z(x_m,t_n)}{(\Delta x)^2} = \frac{z(x_{m-1},t_n) - 2z(x_m,t_n) + z(x_{m+1},t_n)}{(\Delta x)^2}$$

$$\frac{\delta_{t-} z(x_m,t_n)}{\Delta t} = \frac{z(x_m,t_n) - z(x_m,t_{n-1})}{\Delta t}\,.$$

Bijzondere gevallen van (2) zijn de differentievergelijking van Milne ($a = 0$), Crank-Nicolson ($a = \frac{1}{2}$), Laasonen ($a = 1$).

De differenties Δt en Δx moeten niet kleiner worden genomen dan strikt nodig is, omdat kleinere waarden van Δt en Δx resulteren in grotere rekenkosten. In verband hiermee kan het soms zinvol zijn de rand- en/of beginvoorwaarden van een differentiaalprobleem aan te passen alvorens deze te gebruiken bij het toegevoegde differentieprobleem.
Bij het bestuderen van de argumenten op grond waarvan de stapgrootten Δt en Δx kunnen worden gekozen, blijkt het nuttig te zijn een tweetal begrippen in te voeren: "uitwendige oplossing" en "inwendige oplossing". De uitwendige oplossing is een generalisatie van de bekende begrippen: stationaire of periodiek-stationaire oplossing. De inwendige oplossing is het inschakelverschijnsel, dat gesuperponeerd is op de uitwendige oplossing. Tezamen vormen ze de totale oplossing.
Niet zelden zullen de randvoorwaarden niet snel veranderen en dus weinig significante afgeleiden bezitten. Dit houdt dan ook in, dat de uitwendige oplossing weinig significante afgeleiden heeft, zowel naar t als naar x. De inwendige oplossing zal wel snel veranderen in zowel

x- als t-richting, zodat vaak de inwendige oplossing bepalend zal zijn voor de keuze van Δt en Δx.

2.2 Uitwendige en inwendige oplossing

Probleem (1) is een "rand- en beginwaarden"-probleem. Het kan b.v. worden opgevat als de wiskundige beschrijving van de temperatuur in een staaf als functie van de plaats x en de tijd t. Voor $p < 0$ vindt in elk punt x warmteafvoer evenredig met z plaats door thermische geleiding naar een omgeving met temperatuur 0. Warmte kan alleen toegevoerd worden via de randen $x = 0$ en $x = 1$. Op fysische gronden is het dus duidelijk, dat voor $p < 0$ de oplossing van het diffusieprobleem stabiel is, d.w.z. een verandering in de beginvoorwaarde zal voor grote waarden van t naar 0 gaan.
De oplossing is volledig bepaald door de begin- en randvoorwaarden. Bij twee verschillende beginvoorwaarden zal de oplossing voor voldoend grote waarden van t op grond van de stabiliteit vrijwel niet verschillen. De oplossing, die dan aanwezig is, moet dus nagenoeg geheel gedefiniëerd zijn door de randvoorwaarden. De randvoorwaarden dringen blijkbaar aan het fysische systeem een oplossing op. Dit is niet alleen voor grote t waar, maar het geldt op elk tijdstip.
De door de randvoorwaarden op elk tijdstip aan het fysische systeem "opgedrongen" oplossing noemen we de uitwendige oplossing. Bij constante of periodieke randvoorwaarden is de uitwendige oplossing identiek met resp. de stationaire en de periodiek-stationaire oplossing.
De uitwendige oplossing moet op elk tijdstip t_0 zijn vastgelegd door "informatie" omtrent ϕ en η in het verleden. We zullen onze beschouwing beperken tot oneindig vaak differentiëerbare functies $\phi(t)$ en $\eta(t)$. Voor een oneindig vaak differentiëerbare functie geldt, dat verleden zowel als toekomst volledig bepaald is door de afgeleiden op het tijdstip t_0. Het ligt dus voor de hand te trachten de uitwendige oplossing $z_e(x,t)$ op elk tijdstip t op te bouwen uit de afgeleiden van $\phi(t)$ en $\eta(t)$ en wel als een lineaire samenstelling omdat probleem (1) lineair is:

$$(4) \qquad z_e(x,t) = \sum_{k=0}^{\infty} \{g_k(x)\phi^{(k)}(t) + h_k(x)\eta^{(k)}(t)\};\ p = 0.$$

Deze reeksontwikkeling moet uiteraard een identiteit zijn, ongeacht de keuze van de functies $\phi(t)$ en $\eta(t)$. Substitutie van (4) in (1a) leidt o.a. tot de conclusie, dat de coëfficiënten van (4) kunnen worden uitgedrukt in Bernoulli-polynomen:

$$(5) \qquad g_k(x) = h_k(1 - x) = \frac{2^{2k+1}}{(2k + 1)!} B_{2k+1}(1 - \tfrac{x}{2}).$$

Voor een nadere beschouwing van (4) (zoals: convergentie, samenvallen met het begrip periodiek-stationaire oplossing bij periodieke ϕ en η) en het geval $p \neq 0$ zie: [1].
Reeksontwikkeling (4) houdt ten nauwste verband met de reeks van Lidstone [2]:

$$(6) \qquad f(x) = \sum_{k=0}^{\infty} \{g_k(x)f^{(2k)}(0) + g_k(1 - x)f^{(2k)}(1)\},$$

mits

$$(7) \qquad |f^{(2k)}(0)|,\ |f^{(2k)}(1)| \leq c\pi^{2k},\ c = \text{constante} < 1.$$

Ook bij gemengde randvoorwaarden is de uitwendige oplossing te formuleren als een reeksontwikkeling (4), uiteraard met andere rijen functies $g_k(x)$ en $h_k(x)$.

Het begrip uitwendige oplossing is ook bruikbaar bij andere problemen. Zo heeft voor niet-homogene gewone differentievergelijkingen N.E. Nörlund het begrip uitwendige oplossing ingevoerd onder de naam "Hauptlösung" [3]. Bij een inhomogene integraalvergelijking van de 2e soort leidt de constructie van de uitwendige oplossing als een reeksontwikkeling tot de bekende reeks van Neumann.

De inwendige oplossing $z_i(x,t)$ is per definitie het verschil tussen de oplossing van (1) en de uitwendige oplossing, zodat:

(8a) $$\frac{\partial z_i}{\partial t} - \frac{\partial^2 z_i}{\partial x^2} - pz_i = 0,\ z_i(0,t) = z_i(1,t) = 0;$$

(8b) $$z_i(x,0) = z(x,0) - z_e(x,0) = \gamma(x) - z_e(x,0).$$

De inwendige oplossing voldoet aan de homogene differentiaalvergelijking alsmede homogene randvoorwaarden en brengt blijkbaar in rekening de afwijking van de gegeven beginvoorwaarde t.o.v. de beginwaarde van de uitwendige oplossing.

Onder elementaire oplossingen van (8a) zullen we verstaan oplossingen met gescheiden variabelen. Deze blijken te zijn:

(9) $$\sin j\pi x\ e^{\{p-(j\pi)^2\}t},\ j = 1, 2, \ldots .$$

De inwendige oplossing kan worden geschreven als een oneindige reeks van elementaire oplossingen:

(10) $$z_i(x,t) = \sum_{j=1}^{\infty} c_j \sin j\pi x\ e^{\{p-(j\pi)^2\}t},\ z_i(x,0) = \sum_{j=1}^{\infty} c_j \sin j\pi x.$$

Uit (10) volgt, dat de inwendige oplossing stabiel is, zolang $p < \pi^2$.

Ook bij gebruik van de benadering (2) is het mogelijk de uitwendige oplossing te formuleren als een reeksontwikkeling naar differentiequotiënten van $\phi(t_n)$ en $\eta(t_n)$.
Formeel kan (2) ook worden geschreven als:

(11) $$\frac{\delta_{t-}}{\{(a-1)\delta_{t-} + 1\}\Delta t} z(x_m,t_n) - \{\frac{\delta_x^2}{(\Delta x)^2} + p\}z(x_m,t_n) = 0,$$

waarbij we het delen door $(a - 1)\delta_{t-} + 1$ definiëren als:

$$\frac{f(t_n)}{(a - 1)\delta_{t-} + 1} = \sum_{k=0}^{\infty} \{(1 - a)\delta_{t-}\}^k f(t_n). \tag{12}$$

Bij gebruik van de afkorting $0_t(a) = \frac{\delta_{t-}}{\{(a - 1)\delta_{t-} + 1\}\Delta t}$ luidt de bij het differentieprobleem behorende uitwendige oplossing:

$$z_e^*(x_m, t_n) = \sum_{k=0}^{\infty} \{g_k^*(x_m) 0_t^k(a)\phi(t_n) + g_k^*(1 - x_m) 0_t^k(a)\eta(t_n)\} \; ;p=0. \tag{13}$$

Ook nu zijn de functies $g_k^*(x_m)$ polynomen. Voor convergentie van (13) en eigenschappen van de polynomen $g_k^*(x_m)$ alsmede voor $p \neq 0$ wordt weer verwezen naar [1].
We beperken onze beschouwing verder tot die randvoorwaarden $\phi(t)$ en $\eta(t)$, waarbij de reeksontwikkelingen (4) en (13) praktisch na enkele termen mogen worden afgebroken. In dat geval zijn de uitwendige oplossingen (4) en (13) vrijwel gelijk. Dit geldt omdat voor kleine waarden van k allereerst de polynomen $g_k(x_m)$ en $g_k^*(x_m)$ weinig verschillen (voor k = 1, 2 zijn ze gelijk) en er ten tweede dan voor een goede overeenkomst tussen $\phi^{(k)}(t_n)$, $\eta^{(k)}(t_n)$ en resp. $0_t^k\phi(t_n)$, $0_t^k\eta(t_n)$ grote waarden van Δt en Δx toelaatbaar zijn.

Ook voor het gediscretizeerde probleem (1), (2) is de inwendige oplossing per definitie het verschil tussen de totale oplossing en de uitwendige oplossing. We zullen nu laten zien, dat om een goede gelijkenis te verkrijgen tussen de inwendige oplossingen vaak kleine waarden van Δt en Δx noodzakelijk zijn.

De elementaire oplossingen (of wel eigenoplossingen) van het differentieprobleem zijn:

(14) $$\sin j\pi x_m \, g_j^{t_n} \quad , \; j = 1, \ldots, M-1; \; M\Delta x = 1,$$

waarbij voor de groeifactor g_j geldt:

(15) $$g_j = \left\{ \frac{1 - (1 - a)q_j \Delta t}{1 + a q_j \Delta t} \right\}^{\frac{1}{\Delta t}}, \quad q_j = \frac{2(1 - \cos j\pi\Delta x)}{(\Delta x)^2} - p.$$

Hieruit volgt voor de inwendige oplossing $z_i^*(x_m, t_n)$:

(16) $$z_i^*(x_m, t_n) = \sum_{j=1}^{M-1} d_j \sin j\pi x_m \, g_j^{t_n}; \quad z_i^*(x_m, 0) = \sum_{j=1}^{M-1} d_j \sin j\pi x_m.$$

Het differentieprobleem kent slechts een eindig aantal eigenoplossingen. De "hoogfrequente" elementaire oplossingen van het differentiaalprobleem met $j \geq M$ worden niet als zodanig vertolkt bij het differentieprobleem. De aanwezigheid van de hoogfrequente elementaire oplossingen uit zich bij het differentieprobleem door een verandering van de amplituden van de "laagfrequente" elementaire oplossingen, $j < M$. In het geval van gelijke beginvoorwaarden $z_i(x_m, 0)$ en $z_i^*(x_m, 0)$, wat bij de gemaakte veronderstelling omtrent ϕ en η vrijwel juist is, geldt: zie (10) en (16):

(17) $$d_j - c_j = \sum_{r=1}^{M} (c_{2rM+j} - c_{2rM-j}), \; j = 1, \ldots, M-1,$$

wat volgt uit de relatie: $\sin(2rM \pm j)\pi x_m = \pm \sin j\pi x_m$.
Het verschijnsel van het veranderen van de amplituden van de "laagfrequente" elementaire oplossingen staat bekend als "vouweffect".

2.3 Aanpassen van de beginvoorwaarde

Volgens (9) dempen de elementaire oplossingen sneller uit, naarmate j groter is. Dit betekent dat bij het differentiaalprobleem bij toenemende tijd steeds minder elementaire oplossingen van betekenis zijn. We zullen de stapgrootten Δt en Δx, waarmee op het tijdstip T de uitwendige oplossing alsmede de dan in het differentiaalprobleem merkbaar aanwezige elementaire oplossingen juist voldoende worden vertolkt, aanduiden met $\Delta t_{opt.}$ en $\Delta x_{opt.}$.

Veelal zal het vouweffect de keuze $\Delta t = \Delta t_{opt.}$, $\Delta x = \Delta x_{opt.}$ niet toelaten, omdat dan de fout in de amplitude d_j te groot zal zijn voor de laagfrequente elementaire oplossingen. Deze fout kan op twee manieren worden verminderd.

Ten eerste door Δx te verkleinen, daar de rij $|c_j|$ afnemend zal zijn met j. Een gevolg is dat het aantal elementaire oplossingen toeneemt (want M neemt toe). Uit (9) en (15) blijkt dat voor goed vertolken van een meer hoogfrequente elementaire oplossing een kleinere Δt is vereist. Een consequentie van verkleinen van Δx om het vouweffect te verminderen zal dus vaak zijn, dat ook Δt moet worden verminderd.

Een andere mogelijkheid om de fout in de amplitude d_j te verkleinen is het elimineren van het vouweffect. Dit kan gebeuren door in de beginvoorwaarde van de inwendige oplossing van het differentiaalprobleem alleen die fouriercomponenten mee te nemen bij het differentieprobleem, waarvan de bijbehorende elementaire oplossingen op het tijdstip T nog merkbaar aanwezig zijn.

Bij een voorbeeld in [1], waarbij $c_j \approx \frac{1}{(2j+1)^3}$, werd bij een toegelaten fout van 10^{-4} gevonden dat zonder aanpassen $\Delta t = \frac{1}{120}$, $\Delta x = \frac{1}{18}$ moest worden toegepast, terwijl na aanpassen van de beginvoorwaarde $\Delta t = \Delta t_{opt.} = \frac{1}{30}$, $\Delta x = \Delta x_{opt.} = \frac{1}{9}$.

Bij niet-lineaire problemen kan het resultaat van "aanpassen van de beginvoorwaarde" aan het differentieprobleem niet precies worden voorspeld, omdat alleen bij een lineair probleem de elementaire oplossingen elkaar niet beïnvloeden. Het is van de mate van niet-lineariteit afhankelijk òf aanpassen van de beginvoorwaarde zinvol zal zijn. Indien de oplossing van (1) voor een aantal tijdstippen moet worden bepaald, kunnen bij de berekening Δt en Δx geleidelijk worden vergroot.

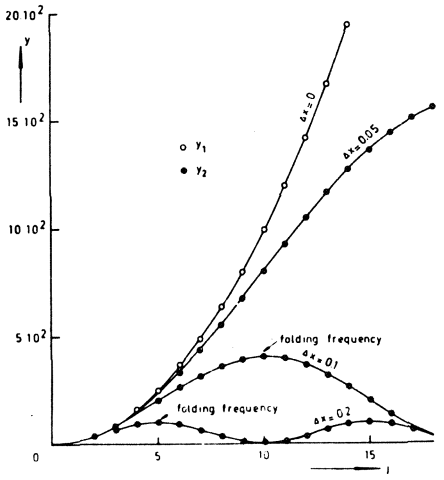
2.4 Gedrag van de elementaire oplossingen als functie van t

Uit (9) en (14), (15) blijkt, dat overeenkomstige elementaire oplossingen verschillen, wat betreft de tijdsafhankelijkheid.
Het discretizeren van x is oorzaak van het vervangen van $y_1 = -(j\pi)^2$ in de exponent van de groeifactor van de j^e elementaire oplossing van het differentiaalprobleem, $e^{p-(j\pi)^2}$, door $y_2 = 2(\cos j\pi\Delta x - 1)/(\Delta x)^2$.
In figuur 1 zijn y_1 en y_2 weergegeven als functie van j. Het vouweffect weerspiegelt zich in de symmetriepunten $j = k/\Delta x$, $k = 1, 2, \ldots$. Voor kleinere waarden van Δx stroken y_1 en y_2 over een groter j-interval.
Door het discretizeren van t wordt e^{p-y_2} als e-macht aangetast; zie (15). Beschouw de groeifactor per tijdstap $g_{j,\Delta t} = g_j^{\Delta t}$. Om het aantal parameters te verminderen schrijven we: $g_{j,\Delta t} = g_u$:

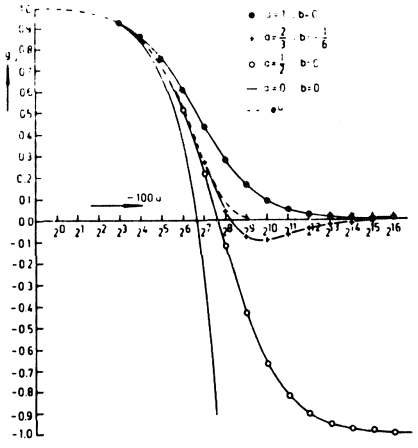
$$(18a) \qquad g_u = \frac{1 + (1 - a)u}{1 - au}\,, \quad u = \left\{p - \frac{2(1 - \cos j\pi\Delta x)}{(\Delta x)^2}\right\}\Delta t.$$

$$(18b) \qquad = 1 + u + au^2 + \ldots + a^{n-1}u^n + \ldots, \quad |au| < 1.$$

In figuur 2 is g_u uitgezet als functie van log u voor een aantal waarden van de parameter a. De kromme $a = \frac{2}{3}$, $b = -\frac{1}{6}$ heeft betrekking op een ander type differentievergelijking (waarvan (2) het bijzondere geval b = 0 voorstelt) die in de volgende paragraaf ter sprake komt.



Figuur 1 Invloed van het discretizeren van x.



Figuur 2 Groeifactor g_u als functie van u voor een aantal differentievergelijkingen.

We zullen nu eerst de groeifactoren van de j^e elementaire oplossingen met elkaar vergelijken voor kleine waarden van Δt en $j\Delta x$ (dus ook van u); gemakshalve beperken we ons tot problemen (1) met $p = 0$.

Door een Taylor-ontwikkeling toe te passen op g_u en $e^{-(j\pi)^2\Delta t}$ vinden we:

$$g_u - e^{-(j\pi)^2\Delta t} = \{u + (j\pi)^2\Delta t\} + \{au^2 - \frac{(j\pi)^4(\Delta t)^2}{2!}\} +$$

$$+ \{a^2u^3 + \frac{(j\pi)^6(\Delta t)^3}{3!}\} + \ldots = \tag{19a}$$

$$= \{\frac{(j\pi)^4(\Delta x)^2 t}{12} \ldots\} + \{(a - \frac{1}{2!})(j\pi)^4(\Delta t)^2 \ldots\} +$$

$$+ \{-(a^2 - \frac{1}{3!})(j\pi)^6(\Delta t)^3\} + \ldots . \tag{19b}$$

Hieruit volgt, dat voor kleine Δt en Δx de beste keuze voor a is $a = \frac{1}{2}$, overeenkomende met de differentievergelijking van Crank-Nicolson. Bij $a = \frac{1}{2}$ is $g_u - e^{-(j\pi)^2\Delta t}$ in eerste benadering evenredig met $(j\pi)^4\{(\Delta x)^2 - (j\pi)^2(\Delta t)^2\}\Delta t$. Een merkwaardige conclusie is dus, dat om de j_0^e elementaire oplossing zo goed mogelijk te vertolken Δt en Δx moeten voldoen aan:

$$\frac{\Delta x}{\Delta t} = j_0\pi . \tag{20}$$

In dit geval geldt:

$$g_{u_0} - e^{-(j_0\pi)^2\Delta t} = \frac{1}{180}(j\pi\Delta x)^5 \ldots . \tag{21}$$

Het verschil $g_u^{\frac{t}{\Delta t}} - e^{-(j\pi)^2 t}$ is een functie van t. Er treedt een maximum op ter waarde

(22a) $$\approx e^{-1}(j\pi)^2 \left\{\frac{(\Delta x)^2}{12} + (a - \frac{1}{2})\Delta t\right\} \text{ voor}$$

(22b) $$t \approx \frac{1}{(j\pi)^2} .$$

Voor $a = \frac{1}{2}$ wordt dit:

(23a) $$\approx \frac{e^{-1}(j\pi)^2}{12} \left\{(\Delta x)^2 - (j\pi\Delta t)^2\right\}$$

en indien $\frac{\Delta x}{\Delta t} = j_0\pi$ is het maximum voor j_0:

(23b) $$\approx \frac{e^{-1}}{180} (j_0\pi\Delta x)^4.$$

Uit het voorgaande is duidelijk, dat bij kleine waarden van $|u|$ de differentievergelijking van Crank-Nicolson de kleinste fout geeft. Echter voor grote waarden van $|u|$ is de differentievergelijking van Laasonen ($a = 1$) de beste, zoals volgt uit:

(24) $$\lim_{|u|\to\infty} g_u = -\frac{1 - a}{a}.$$

Voor $a = \frac{1}{2}$ wordt $g_u \approx -1$ voor grote waarden van $|u|$. Grote waarden van $|u|$ treden op voor grote waarden van j en dus voor de hoogfrequente elementaire oplossingen (tenzij $\Delta t << (\Delta x)^2$). Indien Δt niet klein genoeg is, zijn bij Crank-Nicolson de hoogfrequente elementaire oplossingen alternerend en nagenoeg ongedempt, terwijl deze bij het differentiaalprobleem niet alterneren en sterk gedempt zijn.
Zoals reeds eerder gezegd, moeten (indien aanpassen van de beginvoorwaarde achterwege blijft) Δx en Δt zo klein worden gekozen dat resp. de fout t.g.v. het vouweffect voldoende klein is en alle elementaire oplossingen voldoende nauwkeurig worden vertolkt. Het zal dus duidelijk zijn, dat het bij $a = \frac{1}{2}$ vaak zal voorkomen dat Δt gekozen moet worden op grond van een goede vertolking van de hoogfrequente elementaire oplossingen, d.w.z. $\Delta t < (\Delta x)^2$. In dat geval blijkt uit (19) dat dan de keuze $a = \frac{1}{2}$ ook voor kleine waarden van $|u|$ en Δt geen voordeel biedt.

2.5 Een differentievergelijking, die zowel de laagfrequente als de hoogfrequente elementaire oplossingen goed vertolkt

Bij de differentievergelijking van Crank-Nicolson nadert voor de hoogfrequente elementaire oplossingen de groeifactor naar -1. De reden hiervoor is, dat teller en noemer in (18) polynomen in u zijn van gelijke graad.
In het geval dat p = 0 geeft het toepassen van de operator $\frac{\delta_x^2}{(\Delta x)^2}$ op een elementaire oplossing $z_j = \sin j\pi x_m \, g_j^{t_n}$ als resultaat: $\frac{u}{\Delta t} z_j$. Een tweede macht van u in de noemer van g_u kan dus worden verkregen door aan de differentievergelijking een term $b\Delta t \frac{\delta_x^4 z(x_m, t_n)}{(\Delta x)^4}$, b = constant toe te voegen. De verkregen differentievergelijking luidt:

$$(25) \qquad \frac{\delta_t - z(x_m,t_n)}{\Delta t} - a\frac{\delta_x^2 z(x_m,t_n)}{(\Delta x)^2} - (1-a)\frac{\delta_x^2 z(x_m,t_{n-1})}{(\Delta x)^2} - b\Delta t \frac{\delta_x^4 z(x_m,t_n)}{(\Delta x)^4} = 0,$$

terwijl voor g_u geldt:

$$(26a) \qquad g_u = \frac{1 + (1-a)u}{1 - au - bu^2}$$

$$(26b) \qquad = 1 + u + (a+b)u^2 + (a^2 + ab + b)u^3 + \ldots, \quad |au + bu^2| < 1.$$

Vergelijking van (18b) met (26b) leert, dat voor kleine $|u|$ de beste keus voor a + b is: $a + b = \frac{1}{2}$. We kunnen dan nog één van de constanten vrij kiezen. Bijv. kunnen we zorgen, dat de 3e term van (19b) in 1e benadering verdwijnt; dit geeft $a = \frac{2}{3}$, $b = -\frac{1}{6}$. Echter dan wordt g_u voor $-u > 3$ negatief (dus alternerende oplossingen) en bereikt een minimum $\approx -0{,}1$ voor $u \approx -8$. Om alterneren te voorkomen, dus $g_u \geq 0$, moet $a \geq 1$. Een aantrekkelijke keuze is $a = 1$, $b = -\frac{1}{2}$, omdat dan de differentievergelijking (25) een term minder bevat:

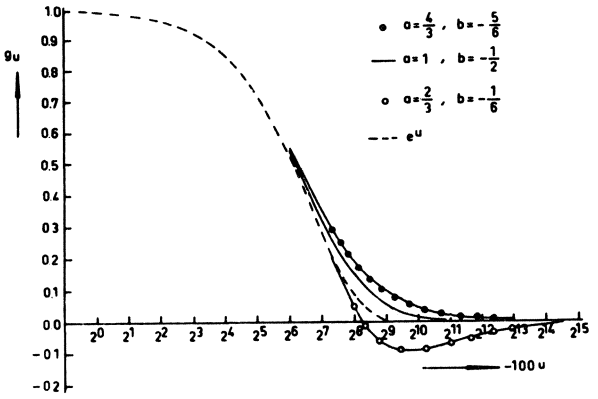
$$(27)\qquad \frac{\delta_{t-}z(x_m,t_n)}{\Delta t} - \frac{\delta_x^2 z(x_m,t_n)}{(\Delta x)^2} + \frac{\Delta t}{2}\,\frac{\delta_x^4 z(x_m,t_n)}{(\Delta x)^4} = 0.$$

Het berekenen van $z(x_m,t_n)$ met behulp van differentievergelijking (25) (en dus ook (27)) vereist als gevolg van de aanwezigheid van $\frac{\delta_x^4 z(x_m,t_n)}{(\Delta x)^4}$ de waarden van z op het tijdstip t_n in de punten x_{m-2}, x_{m-1}, x_{m+1}, x_{m+2}. Dit geeft een moeilijkheid bij het berekenen van z in de punten x_1 en x_{M-1}. De punten x_{-1} en x_{M+1} zijn punten buiten het x-interval $[0,1]$, waarvoor de oplossing niet is vastgelegd. Omdat geldt: $\sin j\pi x_{-1} = -\sin j\pi x$, $\sin j\pi x_{M+1} = -\sin j\pi x_{M-1}$, kunnen $z(x_{-1},t_n)$ en $z(x_{M+1},t_n)$ - wat betreft de inwendige oplossing - gevonden worden door lineaire extrapolatie. We hebben verondersteld, dat de uitwendige oplossing weinig verandert als functie van x en t. Binnen het gebied, waar Δt en Δx gekozen moeten worden op grond van een juiste vertolking van de inwendige oplossing, zijn de waarden van Δt en Δx veel kleiner dan nodig is voor een juiste vertolking van de uitwendige oplossing. Dus ook voor de uitwendige oplossing kan worden volstaan met lineaire extrapolatie om $z(x_{-1},t_n)$ en $z(x_{M+1},t_n)$ te vinden:

$$(28)\quad z(x_{-1},t_n) = 2z(0,t_n) - z(x_1,t_n);\quad z(x_{M+1},t_n) = 2z(1,t_n) - z(x_{M-1},t_n).$$

Voor $p \neq 0$ komt in de noemer van (26a) i.p.v. $-bu^2$ te staan $-b(u - p\Delta t)^2$. Het verloop van g_u als functie van u is dan afhankelijk van de waarde van p. Er blijft echter gelden $\lim_{|u|\to\infty} g_u = 0$. Differentievergelijking (25) zal dus ook bij $p \neq 0$ zowel de laagfrequente als hoogfrequente elementaire oplossingen goed vertolken.

In figuur 3 is voor enkele waarden van a en b het verloop van g_u als functie van u gegeven.



Figuur 3 Groeifactor g_u als functie van u voor differentievergelijking (25)

2.6 Invloed Δt en Δx op rekensnelheid bij iteratief oplossen van het differentieprobleem

Beschouw als niet-lineair diffusieprobleem:

$$(29)\quad \frac{\partial z}{\partial t} - \frac{\partial^2 z}{\partial x^2} - f(x,t,z) = 0;\ z(0,t) = \phi(t),\ z(1,t) = \eta(t);\ z(x,0) = \gamma(x).$$

Op elk tijdstip t_n is de oplossing van het toegevoegde differentieprobleem vastgelegd door $z(0,t_n)$, $z(1,t_n)$ en $z(x_m,t_{n-1})$, $m = 1, \ldots, M-1$ als een stelsel niet-lineaire vergelijkingen.
Omdat de oplossingen van twee opeenvolgende tijdstippen weinig verschillen, is op elk tijdstip een goede beginschatting van de oplossing beschikbaar. Iteratief oplossen van de stelsels niet-lineaire vergelijkingen zal dus mogelijk zijn. Als voorspelling van de oplossing op tijdstip t_n kan bijv. worden genomen de oplossing op het vorige tijdstip. Voordat aan de berekening op t_n, $n \geq 2$ wordt begonnen, is de oplossing van twee vorige tijdstippen nog in het geheugen van de rekenmachine aanwezig. Men kan dus ook de voorspelling van de oplossing op t_n vinden door middel van lineaire extrapolatie.

In het volgende zullen we veronderstellen, dat de voorspelling van de oplossing op tijdstip t_n zo goed is, dat gedurende het iteratieproces slechts kleine veranderingen in z optreden. Dan is het geoorloofd het "iteratieprobleem" (dit is het probleem, wat bij het itereren wordt opgelost) te linearizeren, wat betreft het gedrag van de iteratiefout.

We nemen bij de keuze van het aan het differentieprobleem toegevoegde iteratieprobleem aan, dat het oplossen van de iteratieformule zal gebeuren in de natuurlijke volgorde, dus: $x_1, \ldots, x_{M-1}$. Verder zullen in de iteratieformule bij voorkeur de meest recent verkregen waarden van z worden benut. De iteratiefout $z(x_m,t_n,s) - z(x_m,t_n)$ wordt genoteerd als $\varepsilon(m,s)$. Kortheidshalve schrijven we i.p.v. $z(x_m,t_n)$, $z(x_m,t_n,s)$, $f\{x_m,t_n,z(x_m,t_n)\}$ en $f\{x_m,t_n,z(x_m,t_n,s)\}$ respectievelijk $z(m,n)$, $z_s(m,n)$, $f(m,n,z)$ en $f(m,n,z_s)$. Om een expliciete iteratieformule te verkrijgen, zal bij de s^e iteratie in de (niet-lineaire) functie f als benadering van z moeten worden genomen $z_{s-1}(m,n)$.
Bij gebruik van differentieformule (2) luidt dan de iteratieformule:

$$(30a)\quad z_s(m,n) = C_1\{z_s(m-1,n)+z_{s-1}(m+1,n)\} + C_2 f(m,n,z_{s-1})+F(m,n)$$

$$(30b)\quad F(m,n) = C_3 z(m,n-1) + C_4\{z(m-1,n-1)+z(m+1,n-1)\} + C_5 f(m,n-1,z),$$

$$(30c)\quad C_1 = \frac{a\gamma}{1+2a\gamma}\ ,\quad C_2 = \frac{a\Delta t}{1+2a\gamma}\ ,\quad C_3 = \frac{1-2(1-a)\gamma}{1+2a\gamma}\ ,$$

$$C_4 = \frac{(1-a)\gamma}{1+2a\gamma}\ ,\quad C_5 = \frac{(1-a)\Delta t}{1+2a\gamma}\ ,\quad \gamma = \frac{\Delta t}{(\Delta x)^2}$$

Gelinearizeerd geldt voor de iteratiefout, indien $\frac{\partial f}{\partial z} = p$ = constant wordt genomen:

$$(31a)\quad \varepsilon(m,s) = C_1\{\varepsilon(m-1,s) + \varepsilon(m+1,s-1)\} + C_2 p\varepsilon(m,s-1);\ \varepsilon(0,s)=\varepsilon(1,s)=0;$$

$$(31b)\quad \varepsilon(m,0) = \mu(m,n) - z(m,n),$$

waarbij $\mu(m,n)$ voorstelt de voorspelling van $z(x_m,t_n)$ op het tijdstip t_n.

Als elementaire oplossing van (31a) voldoet:

(32) $$\lambda_j^{m+2s} \sin j\pi x_m,$$

mits λ_j een wortel is van de vierkantsvergelijking:

(33) $$\lambda^2 - (2C_1 \cos j\pi\Delta x) - C_2 p = 0.$$

De wortels λ_{M-j} zijn tegengesteld gelijk aan de wortels λ_j. Omdat bovendien $(-\lambda)^{m+2s} \sin(M-j)\pi x_m = -\lambda^{m+2s} \sin j\pi x_m$, zijn dus als onafhankelijke elementaire oplossingen te onderscheiden:

(34) $$\lambda_{j1,2}^{m+2s} \sin j\pi x_m,\ j=1,\ldots,\ \frac{M}{2} \text{ bij } M=\text{even};\ j=1,\ldots,\ \frac{M-1}{2} \text{ bij } M=\text{oneven}.$$

Er geldt $C_2 = (\Delta x)^2 C_1$. Voor praktische waarden van p en Δx mogen we dus wel stellen $|C_2 p| << 2C_1$. Voor kleine waarden van j zal $\lambda_1 \approx 1$ en $\lambda_2 \approx 0$ zijn. (Opmerking: bij $p = 0$ geldt exact $\lambda_2 = 0$; de bij λ_{j2} behorende componenten van $\varepsilon(m,0)$ zijn dan na een eindig aantal iteraties $\equiv 0$; zie [1].) Voor toenemende j neemt λ_{j1} af; voor $j \approx \frac{M}{2}$ is behalve λ_{j2} ook $\lambda_{j1} \approx 0$.
Naarmate γ groter wordt, komt λ_{j1} voor kleine waarden van j dichterbij 1 te liggen (want $2C_1 \approx 1$), d.w.z. de convergentiesnelheid van het iteratieproces neemt af. De rekentijd per tijdstap zal dus toenemen. Vergroten van Δt behoeft dus bij iteratief oplossen van het differentieprobleem niet te leiden tot vermindering van de totale rekentijd, hoewel het aantal tijdstippen, waarvoor de oplossing wordt bepaald, afneemt!
Beschouw nu differentievergelijking (25) en wel gemakshalve voor $p = 0$. We zullen voor $z(x_{m+2},t_n)$ in de iteratieformule niet de meest recente benadering $z_{s-1}(m+2,n)$ nemen, maar $z_{s-2}(m+2,n)$. De reden hiervoor is, dat dan de elementaire oplossingen van het iteratieprobleem weer de gedaante hebben, zoals aangegeven in (32).

De iteratieformule voor $\varepsilon(m,s)$ luidt:

(35a) $$\varepsilon(m,s) = K_1\{\varepsilon(m-1,s) + \varepsilon(m+1,s-1)\} + K_2\{\varepsilon(m-2,s) + \varepsilon(m+2,s-2)\},$$

(35b) $$K_1 = \frac{a\gamma-4b\gamma^2}{1+2a\gamma-6b\gamma^2}, \quad K_2 = \frac{b\gamma^2}{1+2a\gamma-b\gamma^2}.$$

Hieruit volgt, dat $\lambda_j \sin j\pi x_m$ een elementaire oplossing is, indien:

(36) $$\lambda^2 - 2K_1 \cos j\pi\Delta x - 2K_2 \cos 2j\pi\Delta x = 0.$$

Ook hier zijn de wortels λ_{M-j} tegengesteld gelijk aan de wortels λ_j. Dus (34) is weer geldig.
De in absolute waarde grootste λ treedt op voor $j = 1$. Voor kleine waarden van j en grote waarden van γ geldt:

(37) $$\lambda^2 - \frac{4}{3}\lambda + \frac{1}{3} \approx 0 \qquad \lambda_1 \approx 1, \ \lambda_2 \approx \frac{1}{3}.$$

Voor $\gamma \to \infty$ naderen zowel C_1 als K_1, K_2 tot constanten; zie (30c) en (35b). Dit gebeurt voor de constanten K_1, K_2 echter veel sneller dan voor C_1 door de aanwezigheid van termen met γ^2 in teller en noemer van K_1, K_2.
Dit betekent, dat het toepassen van een grotere waarde van Δt bij differentievergelijking (25) - hoewel toelaatbaar uit het oogpunt van nauwkeurigheid - bezwaarlijk is is verband met de slechtere convergentie van het gekozen iteratieproces. Hierbij moet wel bedacht worden, dat vervangen van $z_{s-2}(m+2,n)$ in de iteratieformule door $z_{s-1}(m+2,n)$ de convergentie van het iteratieproces ten goede zal komen.

We merken nog op, dat het optreden van een elementaire oplossing $\lambda^{m+2s} \sin j\pi x_m$ met $\lambda \to 1$ voor $\gamma \to \infty$ bij een bepaalde iteratieformule inhoudt, dat deze elementaire oplossing ook zal optreden voor $\gamma \to \infty$ bij elke andere keuze van de iteratieformule.

Uit het voorgaande zal het duidelijk zijn, dat bij toepassen van een grote waarde van Δt versnelling van de convergentie voor de "laagfrequente" elementaire oplossingen van het iteratieproces gewenst is. Het iteratieproces kan bijv. zo lang worden voortgezet, dat diens elementaire oplossingen met kleine $|\lambda|$ voldoende klein zijn om daarna door middel van extrapolatie te trachten de overige elementaire oplossingen drastisch te verkleinen.

[1] Dekker, L., Numerical Aspects of the One-Dimensional Diffusion Equation. Proefschrift, Delft juni 1964.

[2] Whittaker, J.M., On Lidstone's Series and Two-Point Expansions of Analytic Functions. Proc. London Math. Soc. (2) vol 36 (1934), pp. 451-469.

[3] Nörlund, N.E., Vorlesungen über Differenzenrechnung, Springer Berlin, 1924.

3. Behoud van stabiliteit bij wijziging van een differentie-schema

Zoals reeds in de hoofdstukken 3 en 4 van deel 1 is uiteengezet, is het van belang te weten of een differentie-schema stabiel is, en wel om de volgende redenen:

a) Als een methode stabiel is, dan werken locale storingen (zoals afrondingsfouten) niet al te fel in het eindantwoord door.

b) Is een consistente methode stabiel, dan is deze methode ook convergent (verg. deel 1, stelling 3.3, p. 43).

c) Voor een stabiele methode kan niet alleen de convergentie bewezen worden, maar kan ook de orde van nauwkeurigheid van het eindantwoord worden bepaald (verg. deel 1, de stellingen 3.1 en 3.4, p. 43).

In dit hoofdstuk zal de stabiliteit besproken worden van (niet lineaire) differentie-schema's. Het zal blijken dat de stabiliteit invariant is bij bepaalde wijzigingen in de differentie-methode. Hiervan kunnen we gebruik maken bij het stabiliteitsonderzoek van sommige (niet lineaire) differentie-schema's.

3.1 Inleiding

We bekijken het volgende begin-randwaarde probleem:

$$(3.1)\qquad \begin{cases} U_t(x,t) = U_{xx}(x,t) + F(x,t,U(x,t)),\ 0 < x < 1,\ 0 < t \leq T \\ U(x,0) = U_0(x),\ 0 \leq x \leq 1 \\ U(0,t) = \phi(t),\ U(1,t) = \eta(t),\ 0 < t \leq T \end{cases}$$

Alle voorkomende functies nemen reële waarden aan; een functie $U(x,t)$ wordt gezocht zo, dat aan (3.1) is voldaan. We veronderstellen dat de functie F aan een Lipschitz voorwaarde voldoet:

$$(3.2)\qquad |F(x,t,U) - F(x,t,\tilde{U})| \leq L \cdot |U - \tilde{U}|,\ 0 < x < 1,\ 0 \leq t \leq T,$$
$$-\infty < U,\ \tilde{U} < +\infty.$$

We nemen aan dat $\Delta t > 0$, $\Delta x > 0$, $r = \Delta t/(\Delta x)^2$, en

$$x_m = m \cdot \Delta x \ (m = 0,1,2,\ldots,M+1) \text{ met } (M+1)\Delta x = 1,$$

$$t_n = n \cdot \Delta t \ (n = 0,1,2,\ldots \leq T/\Delta t).$$

Met $u_{m,n}$ duiden we een benadering van $U(x_m,t_n)$ aan, en

$$f_{m,n} = F(x_m,t_n,u_{m,n}).$$

Een differentie-schema voor de oplossing van (3.1), analoog aan het schema van Crank-Nicolson (verg. hoofdstuk 2, p. 24), is:

$$(3.3)\qquad \begin{cases} \dfrac{u_{m,n+1} - u_{m,n}}{\Delta t} = \dfrac{1}{2}\left\{ \dfrac{u_{m+1,n} - 2u_{m,n} + u_{m-1,n}}{(\Delta x)^2} + \right. \\ \left. + \dfrac{u_{m+1,n+1} - 2u_{m,n+1} + u_{m-1,n+1}}{(\Delta x)^2} \right\} + (1-b)f_{m,n} + bf_{m,n+1} \\ (m = 1,2,\ldots,M;\ n = 0,1,2,\ldots) \end{cases}$$

In (3.3) moeten we natuurlijk $u_{0,n} = \phi(t_n)$ en $u_{M+1,n} = \eta(t_n)$ kiezen. b is een parameter, $0 \leq b \leq 1$. We voeren (evenals in deel 1, p. 62-63) een vector notatie in:

$$u_n = \begin{bmatrix} u_{1,n} \\ u_{2,n} \\ \cdot \\ \cdot \\ \cdot \\ u_{M,n} \end{bmatrix}, f_n = f_n(u_n) = \begin{bmatrix} f_{1,n} \\ f_{2,n} \\ \cdot \\ \cdot \\ \cdot \\ f_{M,n} \end{bmatrix}, \quad c_n = \begin{bmatrix} \frac{r}{2}\left[\phi(t_n) + \phi(t_{n+1})\right] \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ \frac{r}{2}\left[\eta(t_n) + \eta(t_{n+1})\right] \end{bmatrix};$$

$B = (\beta_{i,j})$ is de $M \times M$ matrix met $\beta_{i,i} = 1 + r$ $(i = 1,2,\ldots,M)$, $\beta_{i,i+1} = -r/2$ $(i = 1,2,\ldots,M-1)$, $\beta_{i+1,i} = -r/2$ $(i = 1,2,\ldots,M-1)$ en met de overige $\beta_{i,j} = 0$ $(|i-j| > 1)$;
$A = 2I - B$, waarbij I de $M \times M$ eenheidsmatrix is.
Met deze notaties is (3.3) gelijkwaardig met

$$(3.4)\qquad Bu_{n+1} = Au_n + c_n + \Delta t \cdot \left[(1-b)f_n + bf_{n+1}\right].$$

We definiëren:

$$\Phi_n(u_n,u_{n+1};\Delta t) = (I-B)u_{n+1} + Au_n + c_n,$$

$$\Psi_n(u_n,u_{n+1};\Delta t) = (1-b)f_n + bf_{n+1}, \text{ en}$$

(3.5) $\Omega_n(u_n,u_{n+1};\Delta t) = \Phi_n(u_n,u_{n+1};\Delta t) + \Delta t \cdot \Psi_n(u_n,u_{n+1};\Delta t)$.

Methode (3.4) is dus gelijkwaardig met:

(3.6) $$u_{n+1} = \Omega_n(u_n,u_{n+1};\Delta t).$$

Als $F(x,t,U) \equiv 0$, dan is $\Psi_n \equiv 0$ en is (3.6) de methode van Crank-Nicolson (zoals behandeld in deel 1, p. 63-64). In dit geval is (3.6) een stabiel schema (voor elke waarde van r), d.w.z. er is een getal $\alpha > 0$ zo dat

als $y_{n+1} = \Omega_n(y_n,y_{n+1};\Delta t) + v_{n+1}$, en

$\tilde{y}_{n+1} = \Omega_n(\tilde{y}_n,\tilde{y}_{n+1};\Delta t) + \tilde{v}_{n+1}$ $(n = 0,1,2,\ldots)$,

dan

$$||y_N - \tilde{y}_N|| \leq \alpha \cdot ||y_0 - \tilde{y}_0|| + \alpha \cdot \sum_{i=1}^{N} ||v_i - \tilde{v}_i||$$

voor $N\Delta t = t_N \leq T$.

(y_0, $\tilde{y}_0$, v_i en $\tilde{v}_i$ zijn willekeurige vectoren; als norm van de vector w met coördinaten w_1, w_2, ..., w_M gebruiken we hier

$$||w|| = \sqrt{\sum_{m=1}^{M} w_m^2 \cdot \Delta x}\ ).$$

We stellen nu de vraag voor welke waarden van b methode (3.6) (die gelijkwaardig is met (3.3)) ook in bovenstaande zin stabiel is voor het geval dat $F \not\equiv 0$. Het antwoord op deze en soortgelijke vragen wordt gegeven door de stelling die volgt in 3.2.

3.2 Een stabiliteitscriterium

We bekijken een algemene differentie-methode die vectoren u_n uit een genormeerde vectorruimte V bepaalt (n = k, k+1, k+2, ...), wanneer startwaarden u_0, u_1, ..., u_{k-1} gegeven zijn (k is hierbij een vast natuurlijk getal ≥ 1). We nemen aan ,dat deze u_n successievelijk berekend worden uit de formule

(3.7) $$u_{n+k} = \Omega_n(u_0,u_1,\ldots,u_{n+k};\Delta t),\ n = 0,1,2,\ldots\ .$$

Formule (3.6) is een voorbeeld van (3.7) met k = 1 en Ω_n onafhankelijk van $u_0,u_1,\ldots,u_{n-1}$.

Definitie 3.1

Methode (3.7) heet stabiel als er een $\alpha > 0$ en een $\beta > 0$ is zo dat

$$||y_N - \tilde{y}_N|| \leq \alpha \cdot \sum_{i=0}^{k-1} ||y_i - \tilde{y}_i|| + \alpha \cdot \sum_{i=k}^{N} ||v_i - \tilde{v}_i|| \text{ als}$$

$$y_{n+k} = \Omega_n(y_0,y_1,\ldots,y_{n+k};\Delta t) + v_{n+k}, \text{ en}$$

$$\tilde{y}_{n+k} = \Omega_n(\tilde{y}_0,\tilde{y}_1,\ldots,\tilde{y}_{n+k};\Delta t) + \tilde{v}_{n+k} \quad (n = 0,1,2,\ldots), \text{ en}$$

$$N\Delta t = t_N \leq T,\ 0 < \Delta t \leq \beta.$$

Analoog aan (3.5) nemen we aan dat $\Omega_n(u_0,\ldots,u_{n+k};\Delta t) \equiv$ $\equiv \Phi_n(u_0,\ldots,u_{n+k};\Delta t) + \Delta t \cdot \Psi_n(u_0,\ldots,u_{n+k};\Delta t)$, waarbij de functies Ψ_n aan een soort Lipschitz voorwaarde voldoen:

(3.8) $||\Psi_n(y_0,y_1,\ldots,y_{n+k};\Delta t) - \Psi_n(\tilde{y}_0,\tilde{y}_1,\ldots,\tilde{y}_{n+k};\Delta t)|| \leq$

$$L_0 \cdot \max_{0\leq i\leq n+k} ||y_i - \tilde{y}_i|| \text{ (met } L_0 \text{ onafhankelijk van n en } \Delta t).$$

Nu geldt:

Stelling 3.1

De methode Ω_n is dan en slechts dan stabiel, als de methode, bepaald door Φ_n, stabiel is.

Voor het bewijs van deze stelling, zie [4], p. 165-166.

3.3 Voorbeelden

Voorbeeld 3.1

We zullen bovenstaand stabiliteitscriterium toepassen op methode (3.6). Dit is mogelijk, want met $L_0 = L$ is aan (3.8) voldaan (verg. (3.2)).

Immers:

$$||\Psi_n(y_n,y_{n+1};\Delta t) - \Psi_n(\tilde{y}_n,\tilde{y}_{n+1};\Delta t)|| \leq (1-b)||f_n(y_n)-f_n(\tilde{y}_n)|| + b||f_{n+1}(y_{n+1})$$

$$-f_{n+1}(\tilde{y}_{n+1})|| \leq (1-b)L||y_n-\tilde{y}_n|| + bL||y_{n+1}-\tilde{y}_{n+1}|| \leq L \cdot \max_{i\leq n+1} ||y_i-\tilde{y}_i||.$$

Hiermee is bewezen, dat schema (3.3) stabiel is voor elke F die aan (3.2) voldoet en voor elke waarde van b.

Voorbeeld 3.2

We bekijken het begin-randwaardeprobleem:

$$(3.9)\quad \begin{cases} U_t(x,t) = G(x,t,U(x,t),U_x(x,t),U_{xx}(x,t)) + \\ \quad + \int_0^t F(x,t,U(x,t),s,U(x,s))ds, \quad 0 < x < 1, \ 0 < t \leq T \\ U(x,0) = U_0(x), \quad 0 \leq x \leq 1 \\ U(0,t) = \phi(t), \ U(1,t) = \eta(t), \quad 0 < t \leq T \end{cases}$$

(verg. [2], p. 49-50, voor soortgelijke vergelijkingen).

We nemen aan, dat F voldoet aan een Lipschitz voorwaarde:

$$|F(x,t,U_1,s,U_2) - F(x,t,\tilde{U}_1,s,\tilde{U}_2)| \leq L \cdot (|U_1 - \tilde{U}_1| + |U_2 - \tilde{U}_2|).$$

Wanneer we nu in een differentie-schema voor de oplossing van (3.9) de optredende integraal door een geschikte som vervangen, dan is (op grond van stelling 3.1) het zo ontstane schema dan en slechts dan stabiel als het schema stabiel zou zijn voor het geval dat F = 0.

Voorbeeld 3.3

We bekijken het beginwaarde probleem

$$(3.10)\quad \begin{cases} U_t(x,t) = U_{xx}(x,t) + F(x,t,U(x,t)), \quad -\infty < x < +\infty, \ 0 < t \leq T \\ U(x,0) = U_0(x), \quad -\infty < x < +\infty. \end{cases}$$

Op grond van stelling 3.1 is een differentie-schema voor de oplossing van (3.10) stabiel, indien het stabiel is voor het geval dat $F(x,t,U) \equiv 0$ (we nemen aan dat de functie F in (3.10) aan een Lipschitz voorwaarde m.b.t. U voldoet).
Men kan zich afvragen of stelling 3.1 zo uitgebreid kan worden, dat een dergelijke uitspraak ook mogelijk is voor het geval dat F in (3.10) de meer algemene gedaante $F(x,t,U(x,t),U_x(x,t))$ heeft. Het volgende voorbeeld toont aan, dat zo'n generalisatie onmogelijk is. We beschouwen het beginwaarde probleem:

$$(3.11)\quad \begin{cases} U_t(x,t) = U_{xx}(x,t) + U_x(x,t), \quad -\infty < x < \infty, \quad 0 < t \leq T. \\ U(x,0) = U_0(x), \quad -\infty < x < +\infty. \end{cases}$$

Voor de oplossing van (3.11) stellen we:

$$(3.12)\quad \begin{cases} \dfrac{u_{m,n+1} - u_{m,n}}{\Delta t} = \dfrac{u_{m+1,n} - 2u_{m,n} + u_{m-1,n}}{(\Delta x)^2} + \dfrac{u_{m+1,n} - u_{m,n}}{\Delta x} \\ (m = 0, \pm 1, \pm 2, \ldots ; n = 0,1,2,\ldots). \end{cases}$$

Het is eenvoudig na te gaan, dat het schema

$$(3.13)\qquad \frac{u_{m,n+1} - u_{m,n}}{\Delta t} = \frac{u_{m+1,n} - 2u_{m,n} + u_{m-1,n}}{(\Delta x)^2}$$

stabiel is voor $r = \Delta t/(\Delta x)^2 = 1/2$ wanneer we $||w|| = \sup |w_i|$ als norm kiezen voor de vector $w = (\ldots w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}^i, \ldots)$ uit V. Schema (3.12) is evenwel niet stabiel voor $r = 1/2$ (verg. het bewijs van de instabiliteit van schema (3.13) voor $r > 1/2$, deel 1, hoofdstuk 1, p. 14).

3.4 Een toepassing van het stabiliteitscriterium uit 3.2

We bekijken de vergelijking

$$(3.14)\quad \begin{cases} \dfrac{d}{dt} U(t) = F(t,U(t),U(G(t))), \quad 0 \leq t \leq T \\ \text{met de voorwaarde } U(t) = \phi(t), \quad -\infty < t \leq 0. \end{cases}$$

Alle functies in (3.14) nemen reële waarden aan; de functie G(t) voldoet aan $G(t) \leq t$ (voor $0 \leq t \leq T$) en F voldoet aan een Lipschitz voorwaarde:

$$|F(t,U_1,U_2) - F(t,\tilde{U}_1,\tilde{U}_2)| \leq L \cdot (|U_1-\tilde{U}_1| + |U_2-\tilde{U}_2|). \tag{3.15}$$

Om (3.14) numeriek op te lossen kunnen we (analoog aan de methode van Euler, zie [3], p. 9-63) stellen:

$$\begin{cases} u_{n+1} = u_n + \Delta t \cdot F(t_n,u_n,v_n),\ n = 0,1,2,\ldots, \text{ waarbij} \\ v_n = \phi(G(t_n)) \text{ voor } G(t_n) \leq 0, \text{ en} \\ v_n = \dfrac{G(t_n) - t_m}{\Delta t} \cdot u_{m+1} + \dfrac{t_{m+1} - G(t_n)}{\Delta t} \cdot u_m \text{ voor} \\ 0 \leq t_m < G(t_n) \leq t_{m+1}. \end{cases} \tag{3.16}$$

Met $\Psi_n(u_0,u_1,\ldots,u_{n+1};\Delta t) = F(t_n,u_n,v_n)$ en $L_0 = 2L$ (zie (3.15)) is aan voorwaarde (3.8) voldaan (met $||w|| = |w|$ als w reëel is). De methode $u_{n+1} = u_n$ $(n = 0,1,2,\ldots)$ is stabiel. Op grond van stelling 3.1 is (3.16) dus ook stabiel. Hieruit volgt dat $u_n - U(t_n) = O(\Delta t)$ als U(t) op $0 \leq t \leq T$ een begrensde tweede afgeleide heeft, $u_0 = \phi(0)$ en u_n voor $n \geq 1$ uit (3.16) wordt bepaald (verg. deel 1, hoofdstuk 3, p. 43, of [4], p. 167).

De volgende methoden kunnen als generalisatie van (3.16) beschouwd worden:

$$\sum_{i=0}^{k} \alpha_i u_{n+i} = \Delta t \cdot \sum_{i=0}^{k} \beta_i F(t_{n+i},u_{n+i},v_{n+i}), \tag{3.17}$$

$$n = 0,1,2,\ldots .$$

$k \geq 1$ is een vast natuurlijk getal, α_i en β_i reële constanten, $\alpha_k = 1$ en v_{n+i} is weer een op interpolatie gebaseerde benadering van $U(G(t_{n+i}))$. Door deze interpolatie geschikt uit te voeren en α_i, β_i goed te kiezen kunnen methoden geconstrueerd worden die (locaal bezien) veel nauwkeuriger zijn dan (3.16) (verg. [3], deel II voor een bespreking van "lineaire multistep methoden").

Volgens stelling 3.1 is (3.17) stabiel dan en slechts dan, als de methode

$u_{n+k} = -\sum_{i=0}^{k-1} \alpha_i u_{n+i}$ (n = 0,1,2,...) stabiel is. Voor de stabiliteit van deze methode is nodig en voldoende, dat de coëfficiënten α_i voldoen aan:

(3.18) alle complexe wortels z van de vergelijking $\sum_{i=0}^{k} \alpha_i z^i = 0$

hebben modulus ≤ 1 en wortels met modulus = 1 zijn enkelvoudig.

(verg. [4], p. 168). Hieruit volgt dus:

Stelling 3.2

Methode (3.17) is dan en slechts dan stabiel, als de coëfficiënten α_i voldoen aan (3.18).

Een principiëel andere oplossingsmethode voor vergelijkingen van type (3.14) is te vinden in [1], hoofdstuk 4.

Literatuur

[1] Bellman, R.E., and R.E. Kalaba: Quasilinearization and nonlinear boundary-value problems.
Santa Monica: The Rand Corporation 1965.

[2] Douglas Jr, J.: A survey of numerical methods for parabolic differential equations. Advances in Computers.
New York: Academic Press 1961.

[3] Henrici, P.: Discrete variable methods in ordinary differential equations.
New York: J. Wiley & Sons 1962.

[4] Spijker, M.N.: Convergence and stability of step-by-step methods for the numerical solution of initial-value problems.
Numerische Mathematik 8, 161-177 (1966).

4. Het Noordzee-probleem

In dit hoofdstuk wordt de stabiliteit onderzocht van enkele numerieke oplossingsmethoden voor het Noordzee-probleem. Er zullen uitsluitend expliciete differentie-schema's ter sprake komen.

4.1 De partiële differentiaal-vergelijkingen

Het Noordzee-probleem is een begin-randwaarde probleem voor het volgende stelsel vergelijkingen

$$
\begin{aligned}
U_t &= -\lambda U + \Omega V - gh\, Z_x + F_1 \\
V_t &= -\Omega U - \lambda V - gh\, Z_y + F_2 \qquad (4.1) \\
Z_t &= -U_x - V_y .
\end{aligned}
$$

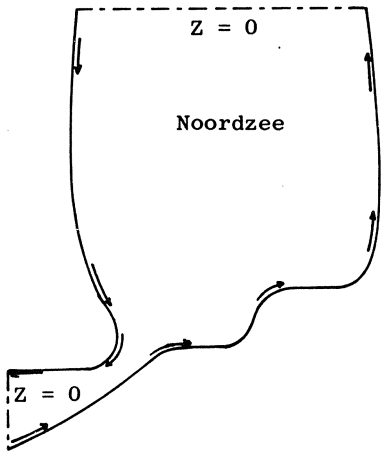
Hierin zijn:

x, y	rechthoekige coördinaten in het zee-vlak,
t	de tijd,
U, V	de over de diepte van de zee geïntegreerde snelheids-componenten van het zeewater (totale stroom genoemd) langs de x en de y-as,
Z	de verhoging van het wateroppervlak boven het ongestoorde niveau Z = 0,
λ	wrijvings-coëfficiënt van het water met de zeebodem,
Ω	Coriolis-coëfficiënt,
g	gravitatie-constante,
h	diepte van de zee,
F_1, F_2	door de windsnelheid bepaalde functies.

De twee eerste vergelijkingen zijn bewegingsvergelijkingen: de op het zeewater werkende krachten zijn de negatieve bodemwrijving $-\lambda(U,V)$, de loodrecht op de beweging werkende Coriolis-kracht $\Omega(V,-U)$, de "verval"-kracht $-gh\,\nabla Z$ en het windveld (F_1,F_2). De derde vergelijking is de continuïteits-vergelijking.
We zullen (4.1) als een matrix-vergelijking schrijven:

$$(4.1')\quad \vec{S}_t = N\vec{S} + \vec{F},\ N = \begin{pmatrix} -\lambda & \Omega & -gh\frac{\partial}{\partial x} \\ -\Omega & -\lambda & -gh\frac{\partial}{\partial y} \\ -\frac{\partial}{\partial x} & -\frac{\partial}{\partial y} & 0 \end{pmatrix},\ \vec{S} = \begin{pmatrix} U \\ V \\ Z \end{pmatrix},\ \vec{F} = \begin{pmatrix} F_1 \\ F_2 \\ 0 \end{pmatrix}.$$

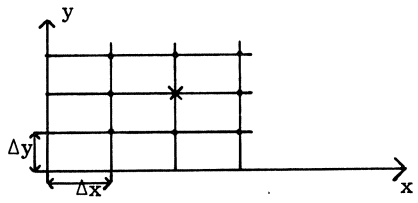
De beginwaarden worden gegeven door de begintoestand $\vec{S}$ voor $t = 0$, de randvoorwaarden worden bepaald door de eis, dat de stroming langs de rand gericht is, wanneer de rand met de kustlijn correspondeert en dat de verhoging $Z = 0$ is, wanneer de rand met een oceaan-begrenzing correspondeert (zie figuur). Voor de wiskundige formulering van de kustvoorwaarden, indien de kust niet eenvoudig langs de x-as ($V = 0$) of langs de y-as ($U = 0$) verloopt, wordt naar [2] verwezen.



Het is duidelijk dat een dergelijk algemeen beginrandwaarde probleem niet analytisch oplosbaar is.
Met behulp van differentie-schema's kan men het probleem voor willekeurige kust en oceaan-situaties en een willekeurig bodemprofiel oplossen.

4.2 Differentie-schema I

In het (x,y)-vlak kiezen we een rechthoekig rooster met netpuntafstanden Δx en Δy; voor functies gegeven op dit rooster definiëren we de operatoren:



$$D_x = \frac{aY_+ + b + aY_-}{2a + b} \cdot \frac{X_+ - X_-}{2\Delta x}$$

$$D_y = \frac{aX_+ + b + aX_-}{2a + b} \cdot \frac{Y_+ - Y_-}{2\Delta y}\ .$$

Hierin zijn a en b gewichts-factoren en $X_\pm$ en $Y_\pm$ translaties over $\pm\,\Delta x$ respectievelijk $\pm\,\Delta y$ in de x- en de y-richting.

De operatoren D_x en D_y zijn benaderingen voor de operatoren $\frac{\partial}{\partial x}$ en $\frac{\partial}{\partial y}$:

$$D_x = \frac{\partial}{\partial x} + O(\Delta x^2), \quad D_y = \frac{\partial}{\partial y} + O(\Delta y)^2.$$

Vervangen we de operator $\frac{\partial}{\partial t}$ door voorwaartse differenties, dan is het schema (Schema I)

$$\vec{s}_{k+1} = (I + \tau D + \tau C)\, \vec{s}_k + \tau \vec{f}_k, \tag{4.2}$$

waarin

$$D = \begin{pmatrix} 0 & 0 & -gh\, D_x \\ 0 & 0 & -gh\, D_y \\ -D_x & -D_y & 0 \end{pmatrix} \text{ en } C = \begin{pmatrix} -\lambda & \Omega & 0 \\ -\Omega & -\lambda & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

consistent met de vergelijking van (4.1'); de benaderingsfout is $O(\Delta t + \Delta x^2 + \Delta y^2)$, waarin Δt de tijdstap is.
Schema I kan gebruikt worden om de toestand $\vec{s}_k$ in de inwendige punten van het rooster te berekenen. In de randpunten gelden andere formules [2], maar daar gaan we niet op in.
Schema I is expliciet, zodat de karakteristieken van (4.1') een eerste verband geven tussen Δt, Δx en Δy.

4.3 Het karakteristieken-criterium

Volgens Courant-Friedrichs (1948) [1], p. 384, worden de karakteristieke vergelijkingen van (4.1') gegeven door:

$$\det \left[p\, I + q \begin{pmatrix} 0 & 0 & gh \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} + r \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & gh \\ 0 & 1 & 0 \end{pmatrix} \right] = 0. \tag{4.3}$$

p, q en r zijn de richtings-cosinussen van de lijn-elementen (dt, dx, dy) loodrecht op de karakteristieke richtingen.
Uitwerking van vergelijking (4.3) geeft:

$$p(p^2 - gh\, q^2 - gh\, r^2) = 0. \tag{4.3'}$$

Aan deze vergelijking wordt voldaan door lijn-elementen evenwijdig met

het (x,y)-vlak en lijn-elementen evenwijdig aan de beschrijvenden van de kegel

$$t^2 - gh\,x^2 - gh\,y^2 = 0.$$

De karakteristieke richtingen zijn dus richtingen evenwijdig aan de t-as en evenwijdig aan de beschrijvenden van de kegel:

$$t^2 - \frac{1}{gh}\,x^2 - \frac{1}{gh}\,y^2 = 0.$$

Een karakteristiek lijn-element (dt, dx, dy) moet voldoen aan

$$dt = \frac{1}{\sqrt{gh}}\sqrt{dx^2 + dy^2}. \tag{4.4}$$

Uit (4.4) volgt dat verstoringen zich met een snelheid $\sqrt{gh}$ voortplanten. Voor de tijdstap Δt vinden we het criterium

$$\Delta t \leq \frac{1}{\sqrt{gh}}\sqrt{\Delta x^2 + \Delta y^2}. \tag{4.5}$$

Voor een concreet geval, waarin $\Delta x = \Delta y = 2\cdot 10^4$ m, $g = 10$ m sec^{-2} en h tussen 20 en 200 m variëert, betekent dit dat de tijdstap Δt moet voldoen aan:

$$\Delta t \leq \frac{4}{5}\cdot 10^4 \cdot \frac{1}{\sqrt{h}}\,\text{sec} \approx \begin{cases} 2000 \text{ sec als } h = 20 \text{ m} \\ 1000 \text{ sec als } h = 80 \text{ m} \\ 600 \text{ sec als } h = 200 \text{ m}. \end{cases} \tag{4.5'}$$

Bij gebruik van uniforme roosters vormt $\Delta t = 10$ min een bovengrens voor de tijdstappen, welke met schema I genomen kunnen worden. Het karakteristieken-criterium (3.5) is wel noodzakelijk, echter niet voldoende om er zeker van te kunnen zijn, dat de resultaten goed zijn; daartoe moet het schema stabiel zijn.

4.4 Stabiliteit van schema I

In deel 1, hoofdstuk 5 is een methode beschreven om de stabiliteit van differentie-schema's te onderzoeken door de eigenwaarden van de dif-

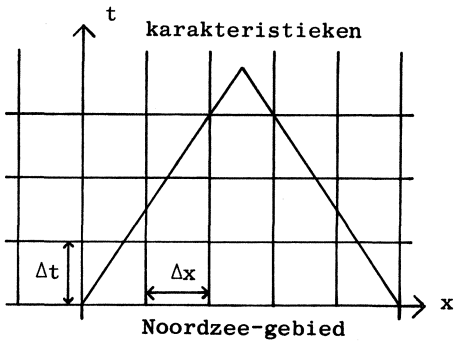
ferentie-operator te beschouwen. Indien de differentie-operator constante coëfficiënten heeft, de netpunten in een rechthoek liggen en de randvoorwaarden door periodiciteitsvoorwaarden te vervangen zijn, zijn de eigenwaarden de eigenwaarden van de amplification-matrix $\hat{A}$ van de differentie-operator A; de amplification-matrix wordt gedefiniëerd door de relatie

$$\hat{A} \exp i \vec{\omega} \cdot \vec{x} = A \exp i \vec{\omega} \cdot \vec{x}, \tag{4.6}$$

waarin $\vec{x}$ de netpunten doorloopt en $\vec{\omega} = (\omega_1, \omega_2)$ een zeker aantal punten in het (x,y)-vlak doorloopt. $\hat{A}$ is een gewone matrix afhankelijk van $\vec{\omega}$; de eigenfuncties van A zijn van de vorm:

$$\vec{a} \exp i \vec{\omega} \cdot \vec{x},$$

waarin $\vec{a}$ een eigenvector van $\hat{A}$ is.
Volgens een vermoeden van O'Brien, Hyman en Kaplan (1950) [4] kan de stabiliteit van differentie-schema's met veranderlijke coëfficiënten bestudeerd worden door aan elk inwendig netpunt een schema toe te voegen met constante coëfficiënten, en wel met die coëfficiënten, welke het gegeven schema in dat netpunt aanneemt. Voor stabiliteit zou het voldoende zijn dat al deze <u>(locaal geldende) schema's</u> stabiel zijn. Volgens Rjabenki en Filippow (1960) [5] is deze hypothese in de praktijk inderdaad houdbaar gebleken. Rest nog de moeilijkheid met de randvoorwaarden; alleen wanneer we een rechthoekig Noordzee-model beschouwen en $\Omega = 0$ stellen, kunnen de randvoorwaarden door periodiciteitsvoorwaarden vervangen worden. In het algemene geval kunnen we slechts <u>noodzakelijke</u> stabiliteits-criteria afleiden. We breiden daartoe de begintoestand $\vec{s}_0$ over het gehele rooster in het (x,y)-vlak op een min of meer willekeurige wijze uit. Passen we een locale differentie-operator op de nieuwe begintoestand $\vec{s}_0$ toe dan lossen we in feite een Cauchy-probleem op, waarvan de oplossing binnen de karakteristieke kegel, gedefiniëerd op het "Noordzee-gebied" met de gezochte oplossing samenvalt. (In nevenstaande figuur is de situatie in een ééndimensionaal model getekend.)



Wil de oplossing binnen deze kegel stabiel zijn dan moet het volledige Cauchy-probleem stabiel zijn, m.a.w. het is noodzakelijk dat de locale schema's zonder randvoorwaarden stabiel zijn. In NSP VIII [3] wordt dit interne stabiliteit genoemd. De interne stabiliteit van de locaal geldende schema's kan geanalyseerd worden met de door (4.6) gedefiniëerde amplification-matrices. In de Noordzee berekeningen zijn deze noodzakelijke criteria ook voldoende gebleken. We construeren nu de amplification-matrix $A(\vec{\omega})$ behorend bij het locale differentie-schema in een zeker netpunt $\vec{x}_0$. Daartoe gaan we het effect na van de operatoren D_x en D_y op de functie $\exp i\,\vec{\omega}\cdot\vec{x}$:

$$D_x \exp i\,\vec{\omega}\cdot\vec{x} = i\,\beta_1 \exp i\,\vec{\omega}\cdot\vec{x},\quad \beta_1^2 = \frac{\xi(b + 2a\sqrt{1-\eta})^2}{(b+2a)^2\Delta x^2}$$

(4.7)

$$D_y \exp i\,\vec{\omega}\cdot\vec{x} = i\,\beta_2 \exp i\,\vec{\omega}\cdot\vec{x},\quad \beta_2^2 = \frac{\eta(b + 2a\sqrt{1-\xi})^2}{(b+2a)^2\Delta y^2}\ .$$

ξ en η zijn reële parameters variërend tussen 0 en 1 ($\xi = \sin^2 \omega_1 \Delta x$ en $\eta = \sin^2 \omega_2 \Delta y$).
Voor $\hat{A}(\vec{\omega})$ vinden we

(4.8) $$\hat{A}(\vec{\omega}) = I + \Delta t\,\hat{D} + \Delta t\,\hat{C},$$

waarin

$$\hat{D} = \begin{pmatrix} 0 & 0 & -gh\,i\beta_1 \\ 0 & 0 & -gh\,i\beta_2 \\ -i\beta_1 & -i\beta_2 & 0 \end{pmatrix}\ .$$

Eisen we dat de oplossing voor berekeningen over een eindig interval $[0,T]$ uniform stabiel is op $(0,\Delta x_0]$, $(0,\Delta y_0]$ en $(0,\Delta t_0]$ (stabiliteit in de zin van Rjabenki en Filippow), dan moet de norm van $\hat{A}$ voldoen aan:

(4.9) $$||\hat{A}|| \leq 1 + O(\Delta t).$$

Nu is $||\hat{A}|| \leq ||I + \Delta t\ \hat{D}|| + O(\Delta t)$, zodat $||I + \Delta t\ \hat{D}|| \leq 1 + O(\Delta t)$ moet zijn. We kiezen in het netpunt $\vec{x}_0$ een nieuwe tijdseenheid S, zodanig dat gh = 1. Men verifiëert eenvoudig dat $(I + \Delta t\ \hat{D})$ een normaal-matrix is (hoofdstuk 1), zodat operator-norm en spectraal-norm gelijk zijn:

$$\begin{aligned} ||I + \Delta t\ \hat{D}|| &= \text{Max}|\mu(I + \Delta t\ \hat{D})| \\ &= \text{Max}\ (1, |1 \pm i\ \Delta t\sqrt{\beta_1^2 + \beta_2^2}|) \\ &= \sqrt{1 + \Delta t^2(\beta_1^2 + \beta_2^2)}. \end{aligned}$$

Definiëren we:

$$\beta_m^2 = \underset{\xi,\eta}{\text{Max}}\ (\beta_1^2 + \beta_2^2),$$

dan is

$$||A(\vec{\omega})|| \leq \sqrt{1 + \Delta t^2\ \beta_m^2} + O(\Delta t) \text{ voor alle } \vec{\omega}.$$

Blijkbaar moet gelden $\Delta t\ \beta_m^2$ = constant voor $\Delta x, \Delta y, \Delta t \to 0$. In de oorspronkelijke eenheden dus

(4.10) $$\Delta t \leq \frac{\text{constante}}{\beta_m^2\ gh}\ ,\ \Delta x, \Delta y, \Delta t \to 0.$$

Hierin is β_m^2 een functie van Δx, Δy, a en b. Uit (4.7) volgt dat

(4.11) $$\beta_m^2\ (\Delta x, \Delta y, a, b) \leq \frac{\Delta^2 x + \Delta^2 y}{\Delta^2 x\ \Delta^2 y}\ .$$

Deze maximale waarde wordt b.v. aangenomen als a = 0:

(4.12) $$\beta_m^2\ (\Delta x, \Delta y, 0, b) = \frac{\Delta x^2 + \Delta y^2}{\Delta x^2\ \Delta y^2}\ .$$

We willen echter β_m^2 zo klein mogelijk laten zijn; voor b = 0 vinden we een kleinere waarde voor β_m; er geldt nl. in dat geval

$$\beta_m^2 = \underset{\xi,\eta}{\text{Max}} \left(\frac{\xi(1-\eta)}{\Delta x^2} + \frac{\eta(1-\xi)}{\Delta y^2}\right).$$

Dit is een harmonische functie in ξ en η, zodat het maximum op de rand van het gebied $0 \leq \xi \leq 1$, $0 \leq \eta \leq 1$ aangenomen wordt, dus

$$\beta_m^2 (\Delta x, \Delta y, a, 0) = \frac{1}{\min(\Delta x^2, \Delta y^2)} . \tag{4.13}$$

Het is duidelijk dat b = 0 de voorkeur verdient boven a = 0.
Stel nu dat a en b gekozen zijn; criterium (4.10) legt geen beperking aan de tijdstap op, maar bij verfijning van het rooster moet gelden $\Delta t\ \beta_m^2$ = constante, dus $\Delta t = O(\Delta x^2 + \Delta y^2)$. Hebben we dus voor een zeker rooster op grond van het karakteristieken-criterium een tijdstap Δt gekozen en willen we de nauwkeurigheid opvoeren door bv. Δx en Δy te halveren, dan betekent dit een vier maal kleinere tijdstap, dus 16 maal meer rekentijd!
In NSP VIII [3] is de stabiliteit van schema I beschouwd voor berekeningen over zeer grote tijden (stabiliteit in de zin van O'Brien, Hyman en Kaplan [4]). Inderdaad zijn bij de Noordzee-berekeningen tijden van 60 uur niet ongebruikelijk, hetgeen bij tijdstappen Δt van de orde van 5 tot 10 minuten als zeer lang beschouwd mag worden. We eisen nu dat alle eigenwaarden van de differentie-operator A binnen de eenheidscirkel liggen, want dan convergeert $A^k \vec{s}_0$ naar nul voor $k \to \infty$ [1].
Dit impliceert wel niet dat de norm van A kleiner of gelijk 1 is (hoofdstuk 3, stelling 3.4), maar wel dat de operatoren $||A^k||$ voor alle k uniform begrensd zijn, zodat het schema stabiel is in de zin van O'Brien, Hyman en Kaplan (deel 1, hoofdstuk 3, definitie 3.10).
Volgens NSP VIII [3] kunnen stabiliteits-criteria verkregen worden door het Hurwitz-criterium toe te passen op de eigenwaarden-vergelijking van de amplification-matrix; deze luidt

$$\mu^3 + a_1 \mu^2 + a_2 \mu + a_3 = 0,$$

waarin a_1, a_2 en a_3 functies zijn van Δt, β_1, β_2, Ω en λ (zie [3]); het Hurwitz-criterium luidt

$$
\begin{aligned}
&1 + a_1 + a_2 + a_3 > 0 \\
&1 - a_1 + a_2 - a_3 > 0 \\
&3 + a_1 - a_2 - 3a_3 > 0 \\
&1 - a_2 + a_1 a_3 - a_3^2 > 0.
\end{aligned}
\tag{4.14}
$$

Onder deze voorwaarden geldt $|\mu| < 1$. Uitwerking [3] geeft

$$\Delta t < \text{Min} \left(\frac{2}{3\lambda} , \frac{2\lambda}{\lambda^2 + \Omega^2}\right) \tag{4.15a}$$

$$\Delta t < \frac{\lambda}{\beta_m^2 gh} . \tag{4.15b}$$

De eerste voorwaarde geeft een directe bovengrens voor Δt; nu zijn λ en Ω ongeveer

$$\lambda \approx 25\ 10^{-6}\ \text{sec}^{-1} , \quad \Omega \approx 125\ 10^{-6}\ \text{sec}^{-1}. \tag{4.16}$$

Substitutie geeft een bovengrens van ongeveer 50 min zodat (4.15a) geen beperking vormt (vergelijk (4.5')).
Het tweede criterium is van dezelfde vorm als dat in het geval van stabiliteit voor eindige intervallen, alleen moet nu de constante gelijk aan de wrijvings-coëfficiënt gekozen worden. Deze is echter zo klein dat een onaanvaardbaar kleine tijdstap gevonden wordt; kiezen we weer $\Delta x = \Delta y = 20\ 10^3$ m, $g = 10$ m/sec^2, $\lambda = 25\ 10^{-6}$ sec^{-1} en $b = 0$, dan geldt:

$$\Delta t < \frac{\lambda}{gh} \Delta x^2 = \frac{1000}{h} \text{ sec},$$

zodat Δt van de orde van enkele seconden gekozen moet worden.
Het is duidelijk, dat schema I onbruikbaar is voor gevallen waarin de bodemwrijving gering is; dit suggereert een schema te construeren met meer demping, m.a.w. de toevoeging van een <u>viscositeits-term</u>.

4.5 Differentie-schema II

Stellen we de stroming (U,V) voor door de vector $\vec{W}$, dan zijn de Noordzee-vergelijkingen onder weglating van de uitwendige krachten (windveld, bodemwrijving, Coriolis-kracht), te schrijven als

$$\vec{W}_t = -gh\,\nabla Z \quad , \qquad Z_t = -\nabla\vec{W}. \tag{4.17}$$

Differentiatie naar t en eliminatie van $\vec{W}$ geeft

$$Z_{tt} = gh\,\Delta Z.$$

Zonder de uitwendige krachten wordt de verhoging door een ongedempte golfbeweging beschreven. De bodemwrijving $-\lambda(U,V)$ zal (indirect) deze golfbeweging enigszins dempen; we kunnen deze demping verhogen door een directe dempings-term toe te voegen b.v. van de vorm $\varepsilon\Delta Z_t$ (voor de physische betekenis van een dergelijke term zij verwezen naar [6], pg. 163): Z wordt dan in ongestoorde toestand beschreven door de vergelijking

$$Z_{tt} = gh\,\Delta Z + \varepsilon\Delta Z_t,$$

en de Noordzee-vergelijkingen (4.17) worden

$$\vec{W}_t = -gh\,\nabla Z \quad , \qquad Z_t = -\nabla\vec{W} + \varepsilon\Delta Z. \tag{4.17'}$$

Om een goede benadering te krijgen voor het oorspronkelijke model, moet ε voldoende klein gekozen worden.
De volledige Noordzee-vergelijkingen worden nu

$$\vec{S}_t = N'\vec{S} + \vec{F} \ , \ N' = \begin{pmatrix} -\lambda & \Omega & -gh\,\frac{\partial}{\partial x} \\ -\Omega & -\lambda & -gh\,\frac{\partial}{\partial y} \\ -\frac{\partial}{\partial x} & -\frac{\partial}{\partial y} & \varepsilon\Delta \end{pmatrix} . \tag{4.18}$$

Het differentie-schema wordt (schema II)

$$\vec{s}_{k+1} = (I + \Delta t\,D' + \Delta t\,C)\vec{s}_k + \Delta t\,\vec{f}_k, \tag{4.19}$$

waarin

$$D' = \begin{pmatrix} 0 & 0 & -gh\, D_x \\ 0 & 0 & -gh\, D_y \\ -D_x & -D_y & \varepsilon(D_x^2 + D_y^2) \end{pmatrix} .$$

Indien we $\varepsilon = O(\Delta t)$ kiezen is schema II consistent met de oorspronkelijke vergelijkingen (4.1), zodat de benaderingsfout weer $O(\Delta t + \Delta^2 x + \Delta^2 y)$ is. We merken op dat voor Noordzee-modellen met $\lambda = 0$, dus zonder demping van de stroming W, het volgende consistente schema gebruikt kan worden voor numerieke berekeningen:

$$\vec{s}_{k+1} = (I + \Delta t\, D' + \Delta t\, C')\vec{s}_k + \Delta t\, \vec{f}_k, \tag{4.20}$$

waarin

$$C' = \begin{pmatrix} -\theta & \Omega & 0 \\ -\Omega & -\theta & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \theta = O(\Delta t).$$

Er wordt dan een model berekend met fictieve bodemwrijving $-\theta\vec{W}$, welke echter naar nul gaat wanneer $\Delta t \to 0$. In de stabiliteitsanalyse zal blijken dat de stabiliteit van een differentie-schema eerst toeneemt met θ maar dan weer afneemt en dat de optimale waarde voor θ zeer klein is [8].

4.6 Enkele Stabiliteits-stellingen

Beschouw het differentie-schema

$$\vec{s}_{k+1} = A\,\vec{s}_k = A^{k+1}\,\vec{s}_0, \tag{4.21}$$

waarin A niet van k afhangt. In deel 1, hoofdstuk 3 werd A wel afhankelijk van k verondersteld en de stabiliteit werd geanalyseerd aan de hand van de grootheid

$$\alpha = \max_{0 \le k \le N-1} ||A_k||, \quad N = T/\Delta t.$$

Het schema was dan stabiel wanneer $\alpha \le 1 + O(\Delta t)$ voor $\Delta t \to 0$ (Lax en Richtmyer) of wanneer $\alpha \le 1$ voor $T \to \infty$ (O'Brien-Hyman-Kaplan). Indien A niet van k afhangt kunnen scherpere stellingen voor deze twee vormen van stabiliteit gegeven worden.

Stelling 4.1

Wanneer schema (4.21) R-F stabiel is dan blijft het schema R-F stabiel, wanneer A door $A + \Delta t\, C$ vervangen wordt en C uniform begrensd is voor $\Delta x, \Delta y, \Delta t \to 0$.

Bewijs

Het nu volgende bewijs werd gegeven door T.L. Johnson (Universiteit van Oslo) tijdens zijn bezoek aan het Mathematisch Centrum (zomer 1967).

Bewezen moet worden dat $(A + \Delta t\, C)^N$ voor $N = T/\Delta t$ en $\Delta x, \Delta y, \Delta t \to 0$ uniform begrensd is wanneer gegeven is dat A^N en C voor $\Delta x, \Delta y, \Delta t \to 0$ uniform begrensd zijn. Binomiaal ontwikkeling van $(A + \Delta t\, C)^N$ leidt tot een som van termen van de vorm

$$(*) \qquad A^{n_0}(\Delta t\, C)^{m_1} A^{n_1}(\Delta t\, C)^{m_2} \ldots A^{n_{r-1}}(\Delta t\, C)^{m_r} A^{n_r},$$

waarin n_i en m_i niet negatieve gehele getallen zijn zodanig dat $\sum_{i=0}^{r} n_i + \sum_{i=0}^{r} m_i = N$, $n_i \geq 1$ voor $i = 1, 2, \ldots, r-1$ en $m_i \geq 1$ voor $i = 1, 2, \ldots, r$. In deze termen worden de factoren A^{n_i} uniform begrensd door een constante M welke we groter dan 1 mogen veronderstellen. Hieruit volgt dat

$$||A^{n_0}(\Delta t\, C)^{m_1} \ldots (\Delta t\, C)^{m_r} A^{n_r}|| \leq M^k ||\Delta t\, C||^{m_1+m_2+\ldots+m_r},$$

waarin k een geheel getal is groter of gelijk aan het aantal getallen $n_i \neq 0$ in (*).

Men verifieert eenvoudig met behulp van volledige inductie dat

$$||(A + \Delta t\, C)^N|| \leq M \sum_{K=0}^{N-1} \binom{N-1}{K} M^K ||\Delta t\, C||^K + ||\Delta t\, C||\, M \sum_{K=0}^{N-1} \binom{N-1}{K} M^K ||\Delta t\, C||^K,$$

waarin elke bijdrage tot de eerste term een som van termen (*) met $n_r \neq 0$ majoriseert en waarin elke bijdrage tot de tweede som een som van termen (*) met $n_r = 0$ majoriseert. Hieruit volgt

$$||(A + \Delta t\, C)^N|| \leq M(1 + M||\Delta t\, C||)^{N-1}(1 + ||\Delta t\, C||) \leq M(1 + \Delta t||C||) \exp(TM||C||).$$

Aangezien C uniform begrensd is voor $\Delta x, \Delta y, \Delta t \to 0$, volgt hieruit de stelling.

Stelling 4.2

Wanneer voor elke $\vec{\omega}$ de eigenwaarden μ van $\hat{A}$ binnen of op de eenheidscirkel liggen en een multipliciteit 1 hebben wanneer $|\mu| = 1$, dan is schema (4.21) stabiel in de zin van O'Brien, Hyman en Kaplan.

Bewijs

Er moet bewezen worden dat $||A^k||$ voor zekere Δx, Δy en Δt uniform in k begrensd is; we beperken ons tot de amplification-matrices $\hat{A}(\vec{\omega})$. Nu geldt volgens Varga [7], p. 65 dat

$$||\hat{A}^k|| \sim c \cdot \binom{k}{p-1} \left[\sigma(\hat{A})\right]^{k-p+1},$$

waarin c een constante is en p de orde van de grootste Jordan-kast J_{i_1} (hoofdstuk 1) met spectraal-norm $\sigma(J_{i_1}) = \sigma(\hat{A})$. Hieruit en uit de voorwaarden van de stelling volgt dat $||\hat{A}^k||$ voor elke $\vec{\omega}$ uniform in k begrensd is en aangezien we voor vaste Δx, Δy, Δt volstaan kunnen met een eindig aantal $\vec{\omega}$'s is $||\hat{A}^k||$ ook uniform begrensd in $\vec{\omega}$, waarmee de stelling bewezen is.

4.7 Stabiliteit van schema II

In de eerste plaats merken we op dat het afhankelijkheids-gebied van $\vec{s}_{k+1}$ in een netpunt $\vec{x}_0$ van een rechthoek met zijden $2\Delta x$ en $2\Delta y$ vergroot is tot een rechthoek met zijden $4\Delta x$ en $4\Delta y$, zodat de door het karakteristieken-criterium toegestane tijdstap Δt verdubbeld is:

$$\Delta t \leq \frac{2}{\sqrt{gh}} \cdot \sqrt{\Delta x^2 + \Delta y^2}. \tag{4.22}$$

Vervolgens onderzoeken we de bruikbaarheid van schema II voor berekeningen over een eindig tijdsinterval $[0,T]$ waarin $\Delta x, \Delta y, \Delta t \to 0$. Volgens stelling 4.1 kunnen we ons beperken tot het onderzoek van de matrix $I + \Delta t\, D'$. De bijbehorende amplification-matrix wordt gegeven door

$$I + \Delta t\, \hat{D}' = \begin{pmatrix} 1 & 0 & -i\,\Delta t\, gh\beta_1 \\ 0 & 1 & -i\,\Delta t\, gh\beta_2 \\ -i\,\Delta t\beta_1 & -i\,\Delta t\beta_2 & 1 - \varepsilon\Delta t(\beta_1^2 + \beta_2^2) \end{pmatrix} .$$

De eigenwaarden ν van $I + \Delta t\, \hat{D}'$ voldoen aan de vergelijking

$$(\nu - 1)(\nu^2 - S\nu + P) = 0,$$

waarin

$$S = 2 - \varepsilon\Delta t(\beta_1^2 + \beta_2^2),$$

$$P = 1 + (gh\,\Delta t - \varepsilon)\Delta t(\beta_1^2 + \beta_2^2).$$

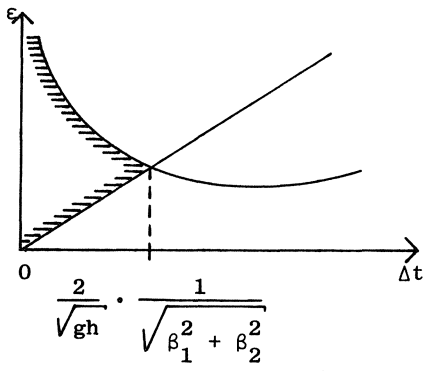
De eigenwaarden ν liggen binnen of op de eenheidscirkel wanneer S en P voldoen aan de ongelijkheden (o.a. af te leiden uit het Hurwitz-criterium)

(4.23) $\qquad P \leq 1,\ 1 - S + P \geq 0,\ 1 + S + P \geq 0.$

Substitutie van S en P geeft:

$$gh\,\Delta t - \varepsilon \leq 0,$$

$$(\beta_1^2 + \beta_2^2)gh\,\Delta t^2 - 2\varepsilon(\beta_1^2 + \beta_2^2)\Delta t + 4 \geq 0.$$



Het door deze betrekkingen in het $(\Delta t, \varepsilon)$ vlak bepaalde gebied is in nevenstaande figuur gearceerd.
De meest gunstige waarde voor ε is

$$\varepsilon = gh\,\Delta t.$$

Met deze waarde voor ε liggen de eigenwaarden van $I + \Delta t\, \hat{D}'(\vec{\omega})$ op de eenheidscirkel $(P = 1)$ wanneer

$$\Delta t \leq \frac{2}{\sqrt{gh}} \cdot \frac{1}{\sqrt{\beta_1^2 + \beta_2^2}} \cdot$$

Volgens Richtmyer [1], p. 65 is er stabiliteit wanneer aan dit criterium voor alle $\vec{\omega}$ voldaan is en wanneer de eigenvectoren van de amplification matrix $I + \Delta t\, \hat{D}'(\vec{\omega})$ voor alle $\vec{\omega}$ linear onafhankelijk zijn. Er geldt

$$\nu_1 = 1,\ \nu_{2,3} = 1 - \tfrac{1}{2} gh(\beta_1^2 + \beta_2^2)\Delta t^2 \pm \sqrt{(\tfrac{1}{4} gh(\beta_1^2 + \beta_2^2)\Delta t^2 - 1)(\beta_1^2 + \beta_2^2)\Delta t^2}.$$

Dus voor

$$\Delta t \leq \frac{2}{\sqrt{gh}} \cdot \frac{1}{\sqrt{\beta_1^2 + \beta_2^2}} \leq \frac{2}{\sqrt{gh}\ \beta_m} \tag{4.24}$$

en

$$\beta_1^2 + \beta_2^2 \neq 0$$

zijn de drie eigenwaarden verschillend en hebben dus onafhankelijke eigenvectoren; indien $\beta_1 = \beta_2 = 0$ is er een drievoudige wortel $\nu = 1$, zodat $I + \Delta t\, \hat{D}' = I$ met eveneens onafhankelijke eigenvectoren. Hiermee is bewezen dat onder de voorwaarde

$$\Delta t \leq \frac{2}{\sqrt{gh}} \cdot \frac{1}{\beta_m} \tag{4.25}$$

schema II R-F stabiel is.
Criterium (4.25) heeft het voordeel boven het overeenkomstige criterium voor schema I, dat Δt lineair met Δx en Δy samenhangt en dat in plaats van $1/h$ de minder snel variërende functie $1/\sqrt{h}$ in het rechterlid optreedt. Vergelijken we (4.25) met het karakteristieken-criterium (4.22) voor het geval $\Delta x = \Delta y$ en $b = 0$, dan vinden we:

Karakteristieken-criterium: $\Delta t \leq \sqrt{2} \cdot \frac{2\Delta x}{\sqrt{gh}}$

Stabiliteits-criterium: $\Delta t \leq \frac{2\Delta x}{\sqrt{gh}}$.

Deze criteria leveren dus vergelijkbare tijdstappen.
Een bezwaar van schema II in vergelijking met schema I, is de toename van het rekenwerk per tijdstap ten gevolge van de operatoren D_x^2 en D_y^2; we zullen nu schema II zo wijzigen dat (4.25) blijft gelden, terwijl de hoeveelheid rekenwerk gelijk wordt aan dat in schema I.

4.8 Differentie-schema III

Schema II wordt gedefiniëerd door de vergelijking:

$$\vec{s}_{k+1} = (I + \Delta t\, D' + \Delta t\, C)\vec{s}_k + \Delta t^2(D\, C\, \vec{s}_k - D\, \vec{f}_k) + \Delta t\, \vec{f}_{k+1}. \tag{4.26}$$

Het is duidelijk, dat indien $\Delta x = O(\Delta t)$ en $\Delta y = O(\Delta t)$, schema III consistent is met de Noordzee-vergelijkingen (4.1) en dat de criteria (4.22) en (4.25) ook voor dit schema gelden.
(4.26) kan geschreven worden als

$$\begin{cases} \vec{w}_{k+1} = \vec{w}_k - \Delta t\, gh \begin{pmatrix} D_x u_k \\ D_y v_k \end{pmatrix} + \Delta t \begin{pmatrix} -\lambda & \Omega \\ -\Omega & \lambda \end{pmatrix} \vec{w}_k + \Delta t \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}_k \\ z_{k+1} = z_k - \Delta t(D_x u_{k+1} + D_y v_{k+1}). \end{cases} \tag{4.26'}$$

Voor numerieke berekeningen zijn de schema's I en III per tijdstap nagenoeg gelijk, want ze verschillen alleen hierin, dat in schema I voorwaartse differenties voor de verhoging gebruikt worden en in schema III achterwaartse differenties.
We merken op dat het schema waarin de stroming met achterwaartse differenties en de verhoging met voorwaartse differenties berekend wordt, neerkomt op demping van de stroming in plaats van demping van de verhoging; de amplification-matrix $\hat{E}$ is dan van de vorm

$$\hat{E} = \begin{pmatrix} 1 - gh\,\Delta t^2\beta_1^2 & -gh\,\Delta t^2\,\beta_1\beta_2 & -i\,gh\,\beta_1 \\ -gh\,\Delta t^2\,\beta_1\beta_2 & 1 - gh\,\Delta t^2\,\beta_1\beta_2 & -i\,gh\,\beta_2 \\ -i\beta_1 & -i\beta_2 & 1 \end{pmatrix}.$$

De stabiliteit is nu echter veel lastiger te analyseren, omdat de eigenwaarden niet expliciet bekend zijn. Bovendien kan ook demping van de stroming verkregen worden door de toevoeging van een term $-\Delta t\theta\vec{\omega}_k$, $\theta = O(\Delta t)$, aan de bewegings-vergelijkingen, zodat we op deze mogelijkheid niet in zullen gaan.

4.9 Stabiliteit van schema III

Schema II lijkt uitermate geschikt voor numerieke berekeningen. Echter moet nog nagegaan worden in hoeverre het bruikbaar is voor berekeningen over de zeer lange tijdsintervallen die gevraagd worden. Volgens stelling 4.2 kan de stabiliteit van de oplossingen als $T \to \infty$ bestudeerd worden door de eigenwaarden van de amplification-matrix

$$\hat{A} = I + \Delta t\,\hat{D}' + \Delta t\,C + \Delta t^2\,\hat{D}\,C$$

te onderzoeken. De eigenwaarden van $\hat{A}$ voldoen aan de vergelijking:

$$\mu^3 + a_1\mu^2 + a_2\mu + a_3 = 0,$$

waarin

$$a_1 = -3 + 2\lambda\Delta t + (\beta_1^2 + \beta_2^2)\Delta t^2,$$
$$a_2 = 3 - 4\lambda\Delta t + (\lambda^2 + \Omega^2)\Delta t^2 - (\beta_1^2 + \beta_2^2)(1 - \lambda\Delta t)\Delta t^2,$$
$$a_3 = -1 + 2\lambda\Delta t - (\lambda^2 + \Omega^2)\Delta t^2,$$

en waarin weer $gh = 1$ gesteld is.
Toepassing van het Hurwitz-criterium (4.14) levert de ongelijkheden

(4.27a) $$\Delta t < \text{Min}\left(\frac{2}{\lambda}, \frac{\lambda}{\lambda^2 + \Omega^2}\right) = \frac{\lambda}{\lambda^2 + \Omega^2},$$

(4.27b) $$\lambda(\beta_1^2 + \beta_2^2) > 0,$$

(4.27c) $$(\beta_1^2 + \beta_2^2) > -(\lambda^2 + \Omega^2) \cdot \frac{\Delta t(\lambda^2 + \Omega^2) - 2\lambda}{\Delta t(\lambda^2 + \Omega^2) - \lambda},$$

(4.27d) $$(\beta_1^2 + \beta_2^2) < 2\,\frac{4(1 - 2\lambda\Delta t) + \Delta t^2(\lambda^2 + \Omega^2)}{\Delta t^2(2 - \lambda\Delta t)}.$$

De eerste ongelijkheid geeft een directe bovengrens voor Δt en is te vergelijken met criterium (4.15a). Aan de tweede ongelijkheid is voldaan voor alle $\vec{\omega}$ met $\beta_1^2 + \beta_2^2 \neq 0$; wanneer $\beta_1 = \beta_2 = 0$ is de amplification-matrix

$$\hat{A} = \begin{pmatrix} 1 - \lambda\Delta t & \Omega\Delta t & 0 \\ -\Omega\Delta t & 1 - \lambda\Delta t & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

met de eigenwaarden vergelijking

$$(\mu - 1)(\mu^2 - 2(2\lambda\Delta t - 1)\mu + (1 - \lambda\Delta t)^2 + \Omega^2\Delta t^2) = 0.$$

Een der oplossingen is $\mu = 1$, de andere twee wortels liggen binnen de eenheidscirkel wanneer Δt voldoet aan

$$\Delta t < \frac{2\lambda}{\lambda^2 + \Omega^2}.$$

Hier is volgens (4.27a) altijd aan voldaan, zodat de voorwaarden van stelling 4.2 voor $\beta_1 = \beta_2 = 0$ vervuld zijn.
Ongelijkheid (4.27c) volgt weer uit (4.27a).
De laatste ongelijkheid schrijven we als

$$\beta_1^2 + \beta_2^2 < 2 \cdot \frac{(2 - \lambda\Delta t)^2}{\Delta t^2(2 - \lambda\Delta t)} + 2\,\frac{\Omega^2}{2 - \lambda\Delta t}.$$

Hieraan is zeker voor alle $\vec{\omega}$ voldaan, wanneer

$$\beta_m^2 < \frac{4 - 2\lambda\Delta t + \Omega^2\Delta t^2}{\Delta t^2} \quad ,$$

ofwel

$$(\beta_m^2 - \Omega^2)\Delta t^2 + 2\lambda\Delta t - 4 < 0.$$

In de oorspronkelijke eenheden wordt dit

$$(gh\beta_m^2 - \Omega^2)\Delta t^2 + 2\lambda\Delta t - 4 < 0.$$

Nu is in het algemeen $gh\beta_m^2 > \Omega^2$ (b.v. $\beta_m = 1/\Delta x = 5\ 10^{-5}$, dus $gh\beta_m^2 = 25h\ 10^{-9}$, terwijl $\Omega^2 \approx 125^2\ 10^{-12} \approx 15\ 10^{-9}$), dus Δt moet voldoen aan

$$\Delta t < \frac{-2\lambda + \sqrt{4\lambda^2 + 16(gh\beta_m^2 - \Omega^2)}}{2(gh\beta_m^2 - \Omega^2)} \ .$$

Herleiding geeft

$$\Delta t < \frac{2}{\beta_m\sqrt{gh}} \cdot \frac{1}{\sqrt{1 - \frac{\Omega^2}{\beta_m^2 gh}}} \cdot \left[\sqrt{1 + \frac{\lambda^2}{4(gh\beta_m^2 - \Omega^2)}} - \frac{\lambda}{2\sqrt{gh\beta_m^2 - \Omega^2}}\right].$$

Voor relatief kleine waarden van Ω en λ is dit nagenoeg gelijk aan het criterium (4.25), zodat in benadering gevonden wordt

(4.28) $$\Delta t < \text{Min}\left(\frac{\lambda}{\lambda^2 + \Omega^2}\ ,\ \frac{2}{\beta_m\sqrt{gh}}\right).$$

Er is dus geen stabiliteit wanneer er geen bodemwrijving is, maar de toevoeging van een zeer geringe wrijvingsterm is al voldoende om er voor te zorgen dat het eerste deel van (4.28) geen enkele beperking

vormt. Stel n.l. dat we in het geval $\lambda = 0$ een kunstmatige wrijvings-coëfficiënt $\theta = c\Delta t$ toelaten, dan moet gelden

$$\Delta t < \frac{c\Delta t}{c^2\Delta t^2 + \Omega^2} ,$$

ofwel

$$\Delta t < \frac{1}{c}\sqrt{c - \Omega^2} .$$

Nu neemt $\frac{1}{c}\sqrt{c - \Omega^2}$ een maximum aan voor $c = 2\Omega^2$, hetgeen resulteert in

$$\Delta t < \frac{1}{2\Omega} \simeq 4 \cdot 10^3 \text{ sec} \simeq 1 \text{ uur.}$$

Dus de introductie van de wrijvings-coëfficiënt $\theta = 2\Omega^2\Delta t$ voorkomt instabiliteit. Het is duidelijk dat niet alleen de consistentie behouden blijft, maar dat dit ook een verwaarloosbare verstoring van het model betekent (voor een uitvoerige behandeling van dergelijke kunstmatige wrijvingstermen zie v.d. Houwen [8]).

Literatuur

[1] Forsythe, G.E. and W.R. Wasow, Finite difference methods for partial differential equations. John Wiley & Sons, Inc., New York-London (1960).

[2] v.d. Houwen, P.J., On the stability of a difference-scheme for the North Sea Problem. Report 100, Math. Centrum, Amsterdam (1966).

[3] Lauwerier, H.A. and B.R. Damsté, A numerical treatment - The North Sea Problem VIII. Proc. Kon. Ned. Ak. v. Wetensch. A66 (1963).

[4] O'Brien, Hyman and Kaplan, Numerical solution of partial differential equations. J. Math. and Physics, Vol. 29 (1950).

[5] Rjabenki, V.S. and A.F. Filippow, Über die Stabilität von Differenzen-gleichungen. VEB. Deutscher Verlag der Wissenschaften, Berlin (1960).

[6] Morse, P.M. and H. Feshbach, Methods of theoretical physics I. McGraw-Hill Book Company, Inc, New York (1953).

[7] Varga, R., Matrix iterative analysis. Prentice-Hall, Inc, Englewood Cliffs, New Jersey (1957).

[8] v.d. Houwen, P.J., Numerical treatment of the North Sea Problem without friction. TN 47, Math. Centre, Amsterdam (1967).

5. Elliptische randwaardeproblemen

Tot dusver zijn uitsluitend begin-randwaardeproblemen ter sprake gekomen. Op zuivere randwaardeproblemen, zoals het Dirichlet randwaardeprobleem, zijn de besproken oplossingsmethoden niet zonder meer van toepassing. Vat men echter een elliptisch randwaardeprobleem op als de limiet voor $t \to \infty$ van een begin-randwaardeprobleem dan is de gegeven theorie zonder meer van toepassing. We zullen er echter gebruik van maken dat alleen de oplossing voor $t \to \infty$ van belang is.

5.1 Definitie van iteratieve processen

Beschouw het elliptische randwaardeprobleem:

$$\begin{aligned} \Delta\, U(x,y) &= F(x,y) \quad \text{in} \quad R, \\ U(x,y) &= \Phi(x,y) \quad \text{op} \quad B, \end{aligned} \tag{5.1}$$

waarin Δ de Laplace-operator is in de Cartesische coördinaten x en y, en waarin F en Φ bekende functies zijn gedefinieerd op het gebied R resp. op de rand B van R; U(x,y) is de gezochte functie en kan opgevat worden als de stationnaire oplossing van het volgende begin-randwaardeprobleem:

$$\begin{aligned} T\,U(x,y,t) &= \Delta\, U(x,y,t) - F(x,y) \quad \text{in } R, \\ U(x,y,t) &= \Phi\,(x,y) \quad \text{op } B, \\ U(x,y,t) &= U_0(x,y) \quad \text{voor } t=0. \end{aligned} \tag{5.2}$$

U hangt nu ook van de tijd t af en T is een operator welke een limietfunctie $\lim_{t\to\infty} U(x,y,t)$ toelaat. Het is duidelijk, dat als $\lim_{t\to\infty} TU(x,y,t) = 0$ de stationnaire oplossing van (5.2) aan (5.1) voldoet; we zullen veronderstellen dat T hieraan voldoet. We definieren nu een <u>iteratief proces</u> als een consistent differentieschema voor (5.2). Het differentieschema moet zo gekozen worden, dat de stationnaire oplossing zo snel mogelijk bereikt wordt. Het is daarom voldoende om het gedrag van het verschil

$$V(x,y,t) = U(x,y,t) - \lim_{t\to\infty} U(x,y,t) \tag{5.3}$$

te beschouwen.

De fout V voldoet aan het homogene begin-randwaardeprobleem:

$$\begin{aligned} TV &= \Delta V \quad \text{in} \quad R, \\ V &= 0 \quad \text{op} \quad B, \\ V &= V_0 \quad \text{voor} \quad t=0. \end{aligned} \tag{5.4}$$

Het differentieanalogon voor (5.4) is van de vorm

$$v_0 = \text{beginapproximatie}, \; v_{k+1} = A_k v_k, \; k=0,1,2,\ldots . \tag{5.5}$$

Ook algemenere randwaardeproblemen dan (5.1) kunnen op deze manier tot de vorm (5.5) teruggebracht worden.
Het probleem is nu operatoren A_k te vinden zodanig dat

$$\prod_{k=0}^{K} A_k \to 0 \quad \text{als} \quad K \to \infty .$$

In de praktijk zal men wensen dat voor een zekere K, $\left\| \prod_{k=0}^{K} A_k \right\|$ zo klein mogelijk is. Als een maat hiervoor heeft men de <u>gemiddelde convergentiesnelheid over K iteraties</u> gedefinieerd:

$$R(K) = \frac{- \ln \sigma \left(\prod_{k=0}^{K-1} A_k \right)}{K} . \tag{5.6}$$

Hierin is $\sigma \left(\prod_{k=0}^{K-1} A_k \right)$ de spectraalnorm van $\prod_{k=0}^{K-1} A_k$.
Indien A_k niet van k afhangt wordt het iteratieve proces (5.5) <u>stationnair</u> genoemd; de convergentiesnelheid R(K) is dan onafhankelijk van K:

$$R(K) = R = - \ln \sigma(A). \tag{5.6}$$

Indien A_k wel van k afhangt, maar periodiek voorkomt, wordt het proces semi-iteratief (Varga) genoemd.

5.2 De methode van Richardson

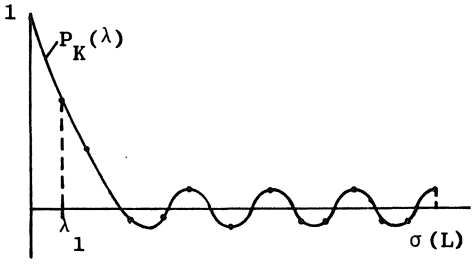
In vele gevallen is (5.5) te schrijven als

$$v_{k+1} = (I - \omega_k L)\, v_k, \tag{5.7}$$

waarin L onafhankelijk van k is en positieve eigenwaarden heeft; ω_k is een reële parameter de z.g. relaxatie-parameter.
We beschouwen semi-iteratieve processen van de graad K; deze worden gedefinieerd door de relaxatie-parameters $\{\omega_k\}_{k=0}^{K-1}$:

$$v_K = \prod_{k=0}^{K-1} (I - \omega_k L)\, v_0 = P_K(L) v_0. \tag{5.7'}$$

$P_K(L)$ is een polynoom-operator van de graad K in L met eigenwaarden $P_K(\lambda_i)$, waarin λ_i voor i = 1, 2, ..., M de eigenwaarden van L zijn.



Het spectrum van $P_K(L)$ bestaat dus uit punten op de kromme $P_K(\lambda)$. In nevenstaande figuur is $P_K(\lambda)$ als functie van λ getekend. De convergentiesnelheid van (5.7') wordt gegeven door

$$R(K) = -\frac{\ln \sigma(P_K(L))}{K} \geq -\frac{\ln ||P_K(\lambda)||}{K}, \tag{5.8}$$

waarin voor de norm $||\ ||$ de maximum norm over het interval $[\lambda_1, \sigma(L)]$ genomen is. Om een zo groot mogelijke convergentiesnelheid te krijgen minimaliseren we $||P_K(\lambda)||$. In dit verband is de volgende stelling van belang.

Stelling 5.1

Het polynoom $T_K^{-1}(\frac{b+a}{b-a}) \cdot T_K(\frac{b+a-2\lambda}{b-a})$, waarin $T_K(y)$ de Chebyshef-polynoom van de graad K in y voorstelt, heeft van alle K^{de} graads-polynomen in λ, welke in $\lambda = 0$ gelijk 1 zijn, de kleinste maximum norm over het interval $a \leq \lambda \leq b$ (Markoff).

In het vervolg zal met $P_K(L)$ steeds de door stelling 5.1 gedefinieerde Chebyshefoperator bedoeld worden.
De operator

$$P_K(L) = T_K^{-1}\left(\frac{b+a}{b-a}\right) \cdot T_K\left(\frac{b+a-2L}{b-a}\right),\ a = \lambda_1,\ b = \sigma(L) \tag{5.9}$$

definieert de methode van Richardson t.o.v. de operator L. Voor de relaxatie-parameters ω_k betekent dit het volgende: de nulpunten van $P_K(\lambda)$ worden gegeven door

$$\lambda = 1/\omega_k,\ k = 0, 1, \ldots, K-1 \tag{5.10}$$

en de nulpunten van $T_K\left(\frac{b+a-2\lambda}{b-a}\right)$ worden gegeven door

$$\lambda = \frac{1}{2}(a+b) + \frac{1}{2}(a-b)\cos\frac{2l+1}{2K}\pi,\ l = 0, 1, \ldots, K-1, \tag{5.11}$$

zodat uit (5.9) volgt dat de methode van Richardson gedefinieerd wordt door de relaxatie-parameters

$$\left(\frac{1}{2}(\lambda_1 + \sigma(L)) + \frac{1}{2}(\lambda_1 - \sigma(L))\cos\frac{2l+1}{2K}\pi\right)^{-1},\ l = 0,1,\ldots,K-1, \tag{5.12}$$

in een of andere volgorde.
De convergentiesnelheid van het Richardson proces volgt uit de relatie:

$$T_K\left(\frac{b+a}{b-a}\right) \sim \frac{1}{2}\exp\left[2K\sqrt{\frac{a}{b}}\right], \tag{5.13}$$

waarin $a << b$ en $K >> 1$.
Dus volgens (5.8) en het feit dat $|P_K(\sigma(L))| = T_K\left(\frac{b+a}{b-a}\right)$ geldt

$$R(K) \sim 2\sqrt{\frac{\lambda_1}{\sigma(L)}} - \frac{\ln 2}{K}\ . \tag{5.14}$$

Opmerking 5.1

Indien de eigenwaarden λ_i van L expliciet bekend zijn kan men voor ω_k nemen de waarden $1/\lambda_i$ (mits K = M) en indien er geen afrondingsfouten optreden wordt de oplossing na M iteraties exact gevonden. In de praktijk kent men de λ_i echter niet (Richardson zelf (1910) verdeelde het interval $[\lambda_1, \sigma(L)]$ in K equidistante intervallen I met $[1/\omega_{k_1} - 1/\omega_{k_2}] = I$).

Opmerking 5.2

Een bezwaar van de Richardson-methode is dat voor kleine waarden van λ_1 zeer grote relaxatie-parameters optreden welke tot numerieke instabiliteit aanleiding kunnen geven. Door echter de volgorde van de ω_k's geschikt te kiezen (zie paragraaf 5.3) blijft de methode voor relatief grote waarden van K en $\sigma(L)$ nog stabiel.

Opmerking 5.3

Alles wat men van de positieve operator L moet weten om methode (5.9) toe te passen is de kleinste eigenwaarde λ_1 en de spectraal-norm $\sigma(L)$. De laatste volgt uit een stelling van Gerschgorin; de eigenwaarde λ_1 is echter moeilijker te schatten, hetgeen een tweede bezwaar van de Richardson-methode is (zie 5.4).

Opmerking 5.4

Uit (5.14) volgt dat men zo mogelijk $\sigma(L)$ zo klein mogelijk moet kiezen. Inderdaad is dit soms eenvoudig te bereiken. Daarentegen wordt λ_1 meestal door de aard van het analytische probleem bepaald en is dezelfde voor alle differentieschema's (zie 5.5). Wanneer men nu de eerste eigenwaarden $\lambda_1, \lambda_2, \ldots$ negeert en $a > \lambda_1$ kiest, maar later de eigenfuncties behorend bij deze eigenwaarden elimineert, dan neemt de asymptotische convergentiesnelheid ($K \to \infty$) toe (zie 5.4).

5.3 Numerieke Stabiliteit van de Richardson-methode

We zullen stabiliteit in de zin van O'Brien, Hyman en Kaplan beschouwen (immers $k \to \infty$). Uit (5.13) volgt dat de methode zeker stabiel t.o.v. de beginvoorwaarden is (deel 1, definitie 3.10), m.a.w. de methode is zwakstabiel (deel 1, stelling 4.1). In tegenstelling tot de stabiliteitsbegrippen van Rjabenki en Filippow impliceert dit geen stabiliteit t.o.v. de inhomogene term (zie [1], pg. 40), ofwel sterke stabiliteit. In deel 1, hoofdstuk 4 is gesteld, dat wanneer de verstoringen van de inhomogene term (zo kunnen afrondingsfouten opgevat worden) at random zijn als functie van plaats en tijd, zwakke stabiliteit sterke

stabiliteit impliceert. Voor een willekeurige volgorde van de relaxatie-parameters in de methode van Richardson is het echter helemaal niet zeker dat de afrondingsfouten at random op zullen treden. Dit is als volgt in te zien:
Stel dat we de fout v_k kunnen ontwikkelen naar de eigenfuncties e(i) van L, dus

$$v_k = \sum_{i=1}^{M} \gamma_k(i)\, e(i).$$

Hierin zijn de $\gamma_0(i)$ coëfficiënten van de ontwikkeling van v_0, en de $\gamma_k(i)$, voor $k \neq 0$, worden gegeven door

$$\gamma_k(i) = (1 - \omega_{k-1}\, \lambda_i)\, \gamma_{k-1}(i).$$

Wanneer nu een aantal malen geitereerd wordt met een grote relaxatie-parameter, zullen de hoogfrequente eigenfuncties, welke meestal met grotere eigenwaarden corresponderen, in sterke mate aanwezig zijn; dit wordt nog versterkt door de hoogfrequente afrondingsfouten, welke nu eenmaal altijd zullen optreden. Het gevolg is dat er grote waarden voor bepaalde $\gamma_k(i)$ kunnen optreden; toch is $\gamma_K(i)$ klein, hetgeen veroorzaakt wordt doordat de grote $\gamma_k(i)$ met kleine factoren $(1 - \omega_k \lambda_i)$ vermenigvuldigd worden; hierbij zal echter precisie-verlies optreden, zelfs zo dat de numerieke fout de fout v_k gaat overheersen. De rij $\{\omega_k\}_{k=0}^{K-1}$ moet dus zo gekozen worden, dat de $\gamma_k(i)$ niet groot kunnen worden, dus het liefst zo snel mogelijk afnemen. Dit betekent dat de beste ordening der ω_k's die zal zijn waarbij de norm van de tussenresultaten v_k zo snel mogelijk naar nul gaat. Dit is een eenvoudig criterium om experimenteel de beste ordening te bepalen. Ook theoretisch kan men echter m.b.v. deze eis tot bepaalde ordeningen komen, waarvan verwacht mag worden dat ze voldoen; daartoe gebruiken we de volgende stelling

Stelling 5.2

Indien K deelbaar is door een oneven getal d, en indien ω_k, k = 0, 1, 2, ..., K-1, de relaxatie-parameters zijn van de Chebyshef-

operator $P_K(L)$, dan vormen de parameters ω_k, $k = md + \frac{1}{2}(d-1)$, $m = 0, 1, \ldots, K/d - 1$ de Chebyshef-operator $P_{K/d}(L)$.

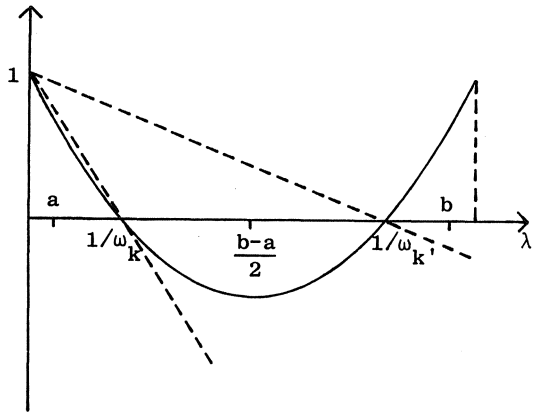
Indien K niet exact deelbaar is door d, mag verwacht worden dat m.b.v. de stelling een redelijk goede benadering voor een Chebyshef-operator van de graad $[K/d] \overset{def}{=}$ entier (K/d) verkregen wordt. Stel nu dat een ordening gevonden is welke stabiel is voor $K \leq K_1$ en we zoeken een ordening die stabiel is voor $K >> K_1$; we kiezen een oneven getal d met $K/d \leq K_1$; de volgens stelling 5.2 bepaalde $[K/d]$ relaxatie-parameters ω_k kunnen stabiel geordend worden, waarmee de fout v_k zo klein gemaakt wordt als in $[K/d]$ iteraties mogelijk is, terwijl tevens al een groot aantal grote relaxatie-parameters "verwerkt" wordt. Men vervolgt dit proces met $[K/d]$ naburige relaxatie-parameters, enz. Het is te verwachten dat men d zo klein als mogelijk moet kiezen, want dan stellen een groep van $[K/d]$ naburige relaxatie-parameters een operator samen welke nog enigszins een benadering voor beginoperator $P_{[K/d]}(L)$ is, dus een goede kans maakt ook numeriek stabiel te zijn.
Rest nog een stabiele begin-ordening te vinden; het schijnt dat Young [2] de enige is die zich met dit ordenings-probleem bezig heeft gehouden. Hij beveelt aan om de ω_k's in paren $(\omega_k, \omega_{k'})$ te ordenen, welke voldoen aan

$$\frac{1}{\omega_k} + \frac{1}{\omega_{k'}} = b - a \tag{5.15}$$

en

$$\sigma((1 - \omega_k L)(1 - \omega_{k'} L)) > \sigma((1 - \omega_{k-1} L)(1 - \omega_{(k-1)'} L)). \tag{5.16}$$

Gemiddeld is de relaxatie-parameter dus $\frac{1}{2}(b-a)$. Experimenten met deze ordening [2] bleven stabiel voor (volgens Forsythe en Wasow [3]) relatief grote K. De spectraal-normen van de operatorparen nemen echter voortdurend toe, zodat te verwachten is dat deze ordening (met toenemende

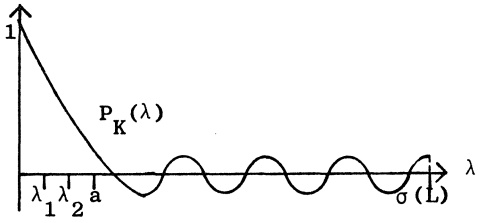


Spectra van de operatoren $1 - \omega_k L$, $1 - \omega_{k'} L$ en $(1 - \omega_k L)(1 - \omega_{k'} L)$

K) instabiel zal worden. Uit in 5.5 beschreven experimenten zal blijken dat wanneer de ordening van Young instabiel is, de hierboven beschreven ordening nog steeds stabiel is.
Voor een uitvoeriger discussie van de numerieke stabiliteit van Richardson's methode zij verwezen naar v.d. Houwen [5].

5.4 Versnelling van de Richardson-methode

Wanneer we in de operator $P_K(L)$, $a > \lambda_1$ kiezen, zodat er een aantal eigenwaarden $\lambda_1, \lambda_2, \ldots, \lambda_{M_1}$ buiten het interval $[a,\sigma(L)]$ liggen, wordt de fout v_k in de deelruimte, opgespannen door de eigenfuncties e(i) van L behorend bij $\lambda_1, \lambda_2, \ldots, \lambda_{M_1}$ geprojecteerd. Stel dat we deze eigenfuncties in K^* iteraties m.b.v. een operator $E_K^*(L)$ kunnen elimineren, dan is de gemiddelde convergentiesnelheid over $K + K^*$ iteraties volgens (5.6) en (5.14)

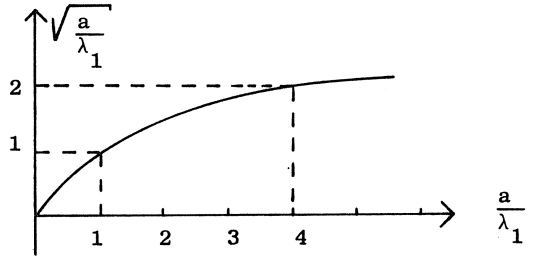


$$R(K + K^*) \sim 2\sqrt{\frac{a}{\sigma(L)}} - \frac{2K^*\sqrt{\frac{a}{\sigma(L)}} + \ln 2 + \ln \sigma(E_K^*)}{K + K^*} . \tag{5.17}$$

De asymptotische convergentie-snelheid R(∞) wordt dus gegeven door

$$R(\infty) = 2\sqrt{\frac{\lambda_1}{\sigma(L)}} \cdot \sqrt{\frac{a}{\lambda_1}} .$$

Wanneer we dit met de asymptotische convergentie-snelheid van het Richardson-proces vergelijken zien we dat een factor $\sqrt{\frac{a}{\lambda_1}}$ gewonnen wordt. In nevenstaande figuur is deze winst-factor uitgezet tegen $\frac{a}{\lambda_1}$;



we zien dat het effect van een grotere waarde voor a in het begin het grootst is, maar snel afneemt als a toeneemt; aangezien het aantal te elimineren eigenfuncties ook met a toeneemt moet men a niet te groot kiezen. We zullen nemen:

$$\lambda_1 < a \leq 4\lambda_1 .$$

Een voordeel is dat men de eerste eigenfunctie λ_1 niet exact hoeft te kennen, zoals bij de oorspronkelijke Richardson-methode wenselijk was.

Vervolgens moeten de eigenfuncties e(1), e(2), ... geëlimineerd worden. Indien K voldoende groot is zal e(1) sterk domineren, dus

$$v_K \simeq \text{const. } e(1).$$

Laat $||\ ||$ een of andere norm zijn, dan geldt:

$$(5.18)\quad \frac{1}{\omega_K} - \frac{||Lu_{K+1} - f||}{||u_{K+1} - u_K||} = \frac{1}{\omega_K} - \frac{||Lv_{K+1}||}{||v_{K+1} - v_K||} = \frac{1}{\omega_K} - \frac{|\lambda_1(1 - \omega_K\lambda_1)|}{|\omega_K\lambda_1|} = \lambda_1 .$$

Dus we vinden automatisch de eerste eigenwaarde (een controle kan verkregen worden door λ_1 met verschillende normen $||\ ||$ te berekenen). We passen nu een polynoom-operator $E_{K_1}^*(\lambda_1, L)$ toe met de eigenschappen

(5.19) $$E_{K_1}^{*}(\lambda_1,0) = 1 \ , \quad E_{K_1}^{*}(\lambda_1,\lambda_1) = 0.$$

Stelling 5.3

Het polynoom $P_{K_0}^{*}(\lambda)$ heeft van alle K_0^{*} de graadspolynomen in λ, welke in $\lambda = 0$ gelijk 1 en in $\lambda = \lambda_0$ gelijk nul zijn, de kleinste maximum norm over het interval $\lambda_0 \leq \lambda \leq b$, wanneer

$$a = a_0^{*} = \frac{2\lambda_0 + b(\cos(\pi/2K_0^{*}) - 1)}{\cos(\pi/2K_0^{*}) + 1} \ .$$

Een bewijs van deze stelling vindt men in v.d. Houwen [4].
Kiezen we $\lambda_0 = \lambda_1$ en $b = \sigma(L)$ dan voldoet de door stelling 5.3 gedefinieerde operator $P_{K_1}^{*}(L)$ aan (5.19), waarmede op de graad K_1^{*} na de beste eliminatie-operator $E_{K_1}^{*}$ voor e(1) gevonden is. Op deze manier kunnen ook e(2), e(3), ... geëlimineerd worden zodra λ_2, λ_3, ... bekend zijn. Rest nog de juiste graad K_i^{*} te bepalen. Een redelijke, maar niet optimale, keus voor de K_i^{*} is het kleinste getal waarvoor

(5.20) $$\sigma(E_{K_i}^{*}(\lambda_i,L)) \leq 1.$$

In dit geval kan K_i^{*} direct in λ_i en $\sigma(L)$ uitgedrukt worden.
Stellen we λ_i gelijk aan het eerste nulpunt van $E_{K_i}^{*}(\lambda_i,\lambda)$ dan is volgens (5.11)

$$\lambda_i = \frac{1}{2}\,(a_i^{*} + \sigma(L)) + \frac{1}{2}\,(a_i^{*} - \sigma(L))\cos\frac{\pi}{2K_i^{*}} \ ,$$

waaruit voor K_i^{*} volgt:

(5.21) $$K_i^{*} = \frac{1}{2}\,\pi\,\arccos^{-1}\left(\frac{a_i^{*} + \sigma(L) - 2\lambda_i}{\sigma(L) - a_i^{*}}\right).$$

Indien $\lambda_i << \sigma(L)$ geldt:

$$K_i^* \sim \frac{1}{4}\pi\sqrt{\frac{\sigma(L) - a_i^*}{\lambda_i - a_i^*}} \quad . \tag{5.21'}$$

Voorwaarde (5.20) betekent dat a_i^* positief moet zijn, maar zo klein mogelijk, dus

$$K_i^* = \left[\frac{1}{4}\pi\sqrt{\frac{\sigma(L)}{\lambda_i}}\right] + 1. \tag{5.22}$$

Ten koste van wat meer rekenwerk kan ook de optimale waarde voor K_i^* gevonden worden, d.w.z. de waarde voor K_i^* welke bij reeds gevonden $K_1^*, K_2^*, \ldots, K_{i-1}^*$, de gemiddelde convergentie-snelheid $R(K + K_1^* + \ldots + K_i^*)$ zo groot mogelijk maakt. We geven hier deze optimale waarde voor de gevallen waarin K groot is:

Stelling 5.4

De waarden voor K_i^*, welke de uitdrukking

$$\left|2\sqrt{\frac{a}{\sigma(L)}} - \ln\frac{\sigma(E_{K_i^*}(\lambda_i, L))}{\sigma(E_{K_{i+1}}(\lambda_i, L))}\right|$$

minimaliseren, maximaliseren voor $K >> 1$ de convergentie-snelheid

$$R(K + K_1^* + K_2^* + \ldots + K_n^*),$$

waarin n het aantal te elimineren eigenfuncties is.

Deze stelling wordt bewezen in v.d. Houwen [4].

In de volgende paragraaf zal het een en ander geïllustreerd worden aan de hand van voorbeelden.

5.5 Proces van Richardson t.o.v. de methode van Jacobi

Als voorbeeld kiezen we het Dirichlet-probleem voor de Laplace-vergelijking op een vierkant met zijden π. Op een rooster met vierkante cellen (h,h) definiëren we de differentie-operator

$$D = \frac{Y_+ + \alpha + Y_-}{2 + \alpha} \cdot \frac{X_+ - 2 + X_-}{h^2} + \frac{X_+ + \beta + X_-}{2 + \beta} \cdot \frac{Y_+ - 2 + Y_-}{h^2},$$

waarin $X_\pm$ en $Y_\pm$ translaties over $\pm$ h in de x- respectievelijk de y-richting voorstellen en waarin α en β positieve gewichtsparameters zijn. De operator D is consistent met de Laplace-operator. Kiezen we voor T de operator $\frac{\partial}{\partial t}$ en discretiseren we deze als voorwaartse differentie met tijdstap τ_k, dan ontstaat het volgende schema voor de fout v_k

$$v_{k+1} = v_k - \tau_k(-D)v_k. \tag{5.23}$$

Indien $\tau_k = \frac{1}{2} h^2$ is dit equivalent met de methode van Jacobi.
Dit schema is van de vorm (5.7) met

$$\omega_k = \tau_k, \quad L = - D.$$

De eigenfuncties van -D, welke nul zijn op de rand worden gegeven door:

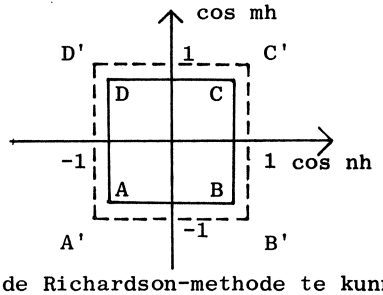
$$e(n,m) = \sin njh \sin mlh, \quad n,m = 1, 2, \ldots, \frac{\pi}{h} - 1.$$

Hierin stelt (jh, lh) een netpunt van het rooster voor. De eigenwaarden van -D behorend bij e(n,m) worden gegeven door

$$\lambda(n,m) = - 2h^{-2} \left[- \left(\frac{\alpha}{2+\alpha} + \frac{\beta}{2+\beta}\right) + \left(\frac{\alpha}{2+\alpha} - \frac{2}{2+\beta}\right) \cos nh \right.$$
$$\left. + \left(\frac{\beta}{2+\beta} - \frac{2}{2+\alpha}\right) \cos mh + \left(\frac{2}{2+\alpha} + \frac{2}{2+\beta}\right) \cos nh \cos mh \right].$$

Nu is $\lambda(n,m)$ een harmonische functie in cos nh en cos mh en is langs de randen van het (cos nh, cos mh)-gebied in benadering lineair. Dus de grootste en kleinste eigenwaarden treden op in de hoekpunten van het (cos nh, cos mh)-gebied. Deze waarden (in de figuur zijn dit $\lambda(A)$, $\lambda(B)$, $\lambda(C)$ en $\lambda(D)$) worden begrensd door $\lambda(A')$, $\lambda(B')$, $\lambda(C')$ en $\lambda(D')$;

voor kleine h geldt:



$$\lambda(A') = 4h^{-2}\left(\frac{\alpha-2}{\alpha+2} + \frac{\beta-2}{\beta+2}\right),$$

$$\lambda(B') = \lambda(D') = 4h^{-2},$$

$$\lambda(C') \simeq 2.$$

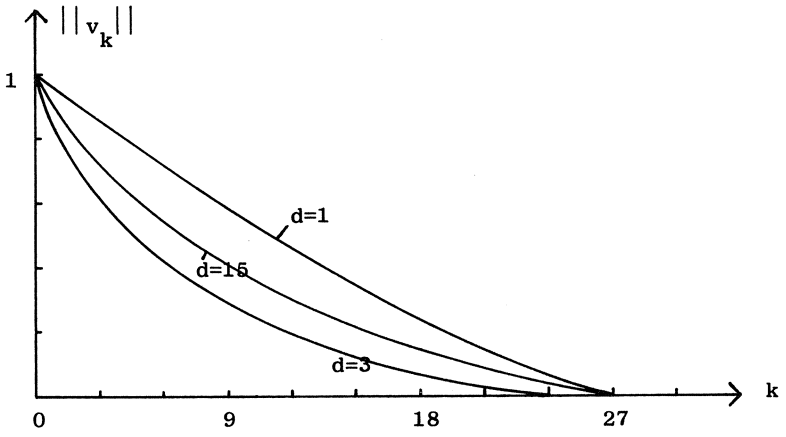
Om de Richardson-methode te kunnen toepassen moet $L = -D$ positief zijn en is het wenselijk dat λ_1 zo groot en $\sigma(L)$ zo klein mogelijk is. We stellen daarom

$$\frac{1}{2}h^2 \leq \frac{\alpha-2}{\alpha+2} + \frac{\beta-2}{\beta+2} \leq 1, \tag{5.24}$$

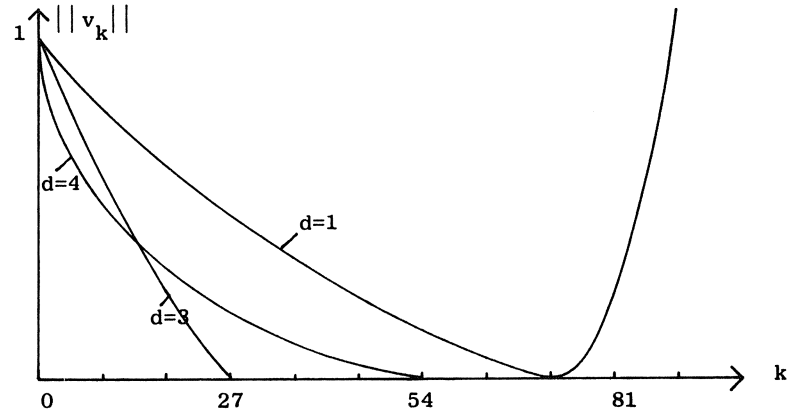
dan is $(\lambda_1,\sigma(L)) \sim (2,4h^{-2})$.
Blijkbaar is de gebruikelijke manier om de Laplace-operator te discretiseren ($\alpha = \beta = \infty$) niet de beste, want dan vinden we $(\lambda_1,\sigma(L)) \sim (2,8h^{-2})$.
Voor de convergentiesnelheden wordt volgens (5.14) verkregen:

$$\begin{cases} R(K) = \sqrt{2}\,h - \dfrac{\ln 2}{K} & \text{als (5.24) geldt,} \\ R(K) = \quad h - \dfrac{\ln 2}{K} & \text{als } \alpha = \beta = \infty. \end{cases} \tag{5.25}$$

Voor grote K betekent het eerste geval een aanzienlijke verbetering.
Young [2] heeft geëxperimenteerd met de Richardson-methode toegepast op het schema $(h,\alpha,\beta) = (\frac{\pi}{20}, \infty, \infty)$; de ordening (5.15), (5.16) bleek stabiel te zijn voor K = 40. We zullen hier het schema $(h,\alpha,\beta) = (\frac{\pi}{20}, 2, \infty)$ onderzoeken. Voor de beginapproximatie kiezen we een functie v_0 met norm ongeveer 1 en we passen de Richardson-methode met (K,a,b) = (27,2,162) toe.



In bovenstaande figuur is de kwadraatnorm van v_k als functie van k uitgezet voor verschillende waarden van d (vergelijk stelling 5.3). De relaxatie-parameters in de factoroperatoren zijn volgens Young geordend ((5.15) en (5.16)). De ordening d = 1 is dus in feite de ordening van Young voor de operator $P_{27}(L)$. We zien dat er een duidelijke voorkeur is voor de ordening d = 3. Vervolgens beschouwen we de Richardsonmethode met (K,a,b) = (81,2,162). Men mag verwachten dat de ordening d = 3 stabiel is op grond van het feit dat $\frac{81}{3} = 27$ relaxatieparameters van de operator $P_{27}(L)$ stabiel geordend kunnen worden.



In bovenstaande figuur zijn de gevallen d = 1, d = 3 en d = 41 getekend. De ordening van Young (d = 1) is instabiel, terwijl de ordening d = 3 inderdaad wel stabiel is. Ook de ordening d = 41 blijkt nog stabiel te zijn.

We besluiten dit hoofdstuk met enige numerieke gegevens over de convergentiesnelheid van de in 5.4 beschreven eliminatiemethode. De eerste (kleinste) eigenwaarden van L = -D worden in benadering gegeven door de formule

$$\lambda(n,m) \sim n^2 + m^2,$$

dus achtereenvolgens 2, 5, 8, 10, 13, ...
We berekenen de convergentiesnelheid volgens formule (5.17). Daartoe moeten we K_i^* kennen.
Volgens formule (5.22) vindt men direct

$$K_1^* = 8 \ , \quad K_2^* = 5 \ , \quad K_3^* = 4 \ , \quad K_4^* = 4.$$

Past men stelling 5.4 toe, dan hebben we de volgende tabel nodig:

K_i^*	$1/\sigma(E_{K_1^*})$	$\frac{\sigma(E_{K_1^*})}{\sigma(E_{K_1^*+1})}$	$1/\sigma(E_{K_2^*})$	$\frac{\sigma(E_{K_2^*})}{\sigma(E_{K_2^*+1})}$	$1/\sigma(E_{K_3^*})$	$\frac{\sigma(E_{K_3^*})}{\sigma(E_{K_3^*+1})}$	$1/\sigma(E_{K_4^*})$	$\frac{\sigma(E_{K_4^*})}{\sigma(E_{K_4^*+1})}$
1	0,013	4,6	0,032	5	0,052		0,066	
2	0,061	2,5	0,16	2,4	0,27		0,34	
3	0,15	1,8	0,39	1,97	0,68	2,05	0,90	2,11
4	0,27	1,6	0,77	1,69	1,4	1,85	1,9	1,94
5	0,44	1,5	1,3	1,68	2,6	1,77	3,7	1,83
6	0,68	1,44	2,2	1,59	4,6	1,69	6,8	1,76
7	0,98	1,42	3,5	1,51	7,8	1,66	12	1,75
8	1,4	1,36	5,3	1,50	13	1,61	21	1,71

Volgens stelling 5.4 moet gelden

$$\exp\left(2\sqrt{\frac{a}{\sigma(L)}}\right) \approx \frac{\sigma(E_{K_i^*})}{\sigma(E_{K_i^*+1})} \ ,$$

dus zodra a gekozen is kan K_i afgelezen worden, waarbij we K_i^* naar boven zullen afronden, omdat de schatting van stelling 5.4 een ondergrens is.

B.v. a = 8 levert $\exp(2\sqrt{\frac{a}{\sigma(L)}}) \simeq 1{,}56$ en dus $K_1^* = 5$ en $K_2^* = 7$ met

$$\sigma(E_{K^*}) \leq \sigma(E_{K_1^*}) \cdot \sigma(E_{K_2^*}) = (0{,}44)^{-1} \cdot (3{,}5)^{-1} \simeq 0{,}62.$$

Aldus verkrijgt men de volgende tabel

a	λ_1	λ_2	λ_3	λ_4	K_1^*	K_2^*	K_3^*	K_4^*	$\sigma(E_{K^*})$	$R(K + K^*)$
5	2				8				0,7	0,351 - 3,144/(K + 8)
5	2				8				0,7	0,351 - 3,144/(K + 8)
8	2	5			8	5			0,55	0,444 - 5,867/(K + 13)
8	2	5			5	7			0,62	0,444 - 5,543/(K + 12)
10	2	5	8		8	5	4		0,4	0,497 - 8,226/(K + 17)
10	2	5	8		4	6	8		0,13	0,497 - 7,599/(K + 18)
13	2	5	8	10	8	5	4	4	0,22	0,566 -11,065/(K + 21)
13	2	5	8	10	4	4	5	7	0,15	0,566 -10,142/(K + 20)

De eliminatie-operator is afwisselend volgens (5.22) en stelling 5.4 berekend; we zien dat beide manieren ongeveer op hetzelfde neerkomen. Om te illustreren dat de eliminatiemethode werkelijk een aanzienlijke versnelling betekent, berekenen we het aantal slagen dat nodig is om hetzelfde resultaat te verkrijgen als met de Richardson-methode (a,b,K) = (2,162,81) verkregen wordt:
Stel

$$R(K + K^*) = A - \frac{B}{K + K^*},$$

dan is

$$(K + K^*)R(K + K^*) = A(K + K^*) - B = 81R_0,$$

waarin R_0 de snelheid van het Richardson-proces is; het aantal benodigde iteraties is dus

$$K + K^* = \frac{81R_0 + B}{A}.$$

Nu is $R_0 = 0{,}222 - \frac{0{,}693}{81} \sim 0{,}214$, zodat we achtereenvolgens uit bovenstaande tabel vinden:

$$K + K^{*} = 60{,}60;\ 52{,}51;\ 51{,}50;\ 50{,}48.$$

Door de eliminatiemethode toe te passen en $\alpha = 2$, $\beta = \infty$ te kiezen is een besparing van ongeveer 60% verkregen op het door Young toegepaste iteratieproces.

Literatuur

[1] Rjabenki, V.S. und A.F. Filippow, Über die Stabilität von Differenzen-gleichungen. VEB, Deutscher Verlag der Wissenschaften. Berlin (1960).

[2] Young, D., On Richardson's method for solving linear systems with positive definite matrices. J. of Math. and Phys. Vol. XXXII (1953).

[3] Forsythe, G.E. and W.R. Wasow, Finite difference methods for partial differential equations. John Wiley & Sons, Inc., New York-London (1960).

[4] v.d. Houwen, P.J., On the acceleration of Richardson's method I, TW 104, Math. Centre Amsterdam (1967).

[5] v.d. Houwen, P.J., On the acceleration of Richardson's method II, TW 107, Math. Centre Amsterdam (1967).