TECHNISCHE HOGESCHOOL EINDHOVEN
Afdeling Algemene Wetenschappen
Onderafdeling der Wiskunde

OPGAVEN

NUMERIEKE WISKUNDE II

Najaar 1972-1973

Oefeningen Numerieke Wiskunde II najaar 1973

1. Het oplossen van stelsels lineaire vergelijkingen.

In deze opgave zal met $\| \ \|$ steeds $\| \ \|_\infty$ bedoeld worden.
Bij het praktische gedeelte moet gebruik gemaakt worden van de Algolprocedures uit RC-informatie nr. 5, uitgezonderd de procedure CROUTDECOMPOSITION. Hoe deze procedures gehanteerd moeten worden op de Burroughs en wat de veranderingen zijn t.o.v. de X-8 staat beschreven in RC-informatie nr. 39. Bovendien moet gebruik gemaakt worden van de procedures INPROD en DLINPROD; zie voor het gebruik van deze procedures ook de RC-informatie nr. 39. Bij de praktikumleiding is een gewijzigde procedure CROUTDECOMPOSITION te verkrijgen, die daarin van de bibliotheekversie verschilt dat de pivotkeuze niet gebaseerd is op de geschaalde coefficiëntenmatrix. Verder is een procedure beschikbaar om matrices met gewenst (slecht) conditiegetal te genereren.

Praktisch gedeelte

A. Maak een Algolprogramma,dat bij gegeven n×n-matrix A, L en U berekent zodanig dat $PA + E_{LU} = LU$ (P is een permutatiematrix). Tevens moet berekend worden: $\|A\|$, $\|L\|$, $\|U\|$, $\| |L| |U| \|$, $\|A^{-1}\|$, $\| |(PA)^{-1}| |L| |U| \|$, E_{LU} in dubbellengte, en $\|E_{LU}\|$.

Verder moet voor een aantal verschillende rechterleden b berekend worden:
een benadering x_0 voor de oplossing van $Ax = b$,
$r := b - Ax_0$ zowel in enkellengte als in dubbellengte,
y uit $Ay = r$ (uit de in dubbellengte berekende r),
en tenslotte $x_1 := x_0 + y$.
Test de correctheid van Uw programma met enkele eenvoudige voorbeelden.

B. Behandel met Uw programma voor een aantal waarden van n, waarbij ook grotere waarden bijv. n = 50 of n = 100 voorkomen, de volgende gevallen:

a) $c(A) := \|A\|\|A^{-1}\|$ is klein, d.w.z. $c(A) \approx 1$.

b) $c(A)$ is groot, d.w.z. $c(A) \approx 10^4$, 10^6.

Neem in beide gevallen een vaste matrix A, een aantal vektoren b waarvoor geldt $\frac{\|x\|}{\|b\|} \sim \|A^{-1}\|$ en een aantal vektoren b waarvoor geldt $\frac{\|x\|}{\|b\|} \sim \frac{1}{\|A\|}$.

Opmerking. Er geldt

$$\frac{1}{\|A\|} \le \frac{\|x\|}{\|b\|} \le \|A^{-1}\| \,.$$

Bewijs dit.

De keuze van b, zodat de rechter ongelijkheid scherp is, is eenvoudig als A^{-1} bekend is. Voor willekeurige b zal bijna altijd ook gelden $\frac{\|x\|}{\|b\|} \sim \|A^{-1}\|$. Ga dit na.

Bedenk zelf hoe U b kunt kiezen zodat $\frac{\|x\|}{\|b\|} \sim \frac{1}{\|A\|}$.

Bespreek de numerieke resultaten aan de hand van de volgende opmerkingen en vragen.

1) Er geldt

$$\frac{\|E_{LU}\|}{\|A\|}^{(1)} \le n\eta \frac{\| \, |L|\,|U| \, \|}{\|A\|}^{(2)} \le n\eta \frac{\|L\| \, \|U\|}{\|A\|}^{(3)} \le n^2 \eta \frac{\|U\|}{\|A\|}^{(4)} \le n^2 2^{n-1} \eta^{(5)} \,,$$

waarbij η afhangt van de arithmetiek van de machine.

Verzamel de berekende waarden van (1), (2), (3), (4) en (5) van een aantal praktikanten. Ga na welke ongelijkheden in de praktijk ongeveer scherp zijn en welke niet. Geef een verklaring voor Uw waarnemingen. Wat is een in de praktijk bruikbare schatting voor $\frac{\|E_{LU}\|}{\|A\|}$?

2) De (voorwaartse) fout δx_0 in x_0 voldoet aan:

$$\frac{\|\delta x_0\|}{\|x_0\|}^{(1)} \le f(n)\eta \, \| \, |(PA)^{-1}|\,|L|\,|U| \, \|^{(2)} \le f(n)\eta \; c(A) \frac{\| \, |L|\,|U| \, \|}{\|A\|}^{(3)} \,.$$

De berekende y is een benadering voor δx_0 (zie vraag 4). Neem daarom als benadering voor $\frac{\|\delta x_0\|}{\|x_0\|}$: $\frac{\|y\|}{\|x_0\|}$

Verzamel de waarden van (1), (2) en (3) van een aantal praktikanten. Ga na welke ongelijkheden ongeveer scherp zijn en welke niet.

Hoe gedraagt $\frac{\|\delta x_0\|}{\|x_0\|} \cdot \frac{1}{\eta \| \, |(PA)^{-1}|\,|L|\,|U| \, \|}$ zich in de praktijk als functie van n?

Hoe $\frac{\|\delta x_0\|}{\|x_0\|} \cdot \frac{1}{\eta c(A)}$?

3) In het geval dat c(A) groot is blijkt vaak dat de fouten δx_0, behorende bij heel verschillende rechterleden b, ongeveer dezelfde richting hebben al ontstaan deze door afrondingsfouten met een willekeurig karakter.
Ga dit na.
Geef een verklaring (geen bewijs) voor het geconstateerde gedrag van δx_0.

4) Ga na hoeveel significante cijfers de residuvektor r heeft in het geval dat deze is berekend in enkellengte en in het geval dat deze is berekend in dubbellengte.
Hoeveel significante cijfers verwacht U in y als benadering voor δx_0?

5) Beschouw de volgende situaties:

a) x_0 is de afronding van de exacte x;
b) x_0 is verkregen door numeriek oplossen van het stelsel.

Noem in beide gevallen $\delta x_0 = x - x_0$, $r = -A\delta x_0$.

(i) Ga na dat in beide gevallen als regel geldt

$$\|r\| \sim n \quad \|A\| \, \|x\| \, ,$$

zij het op zeer verschillende gronden!

(ii) Bewijs dat altijd geldt

$$\frac{1}{\|A\|} \le \frac{\|\delta x_0\|}{\|r\|} \le \|A^{-1}\| \, .$$

Ga na dat als regel in geval a) de linkerongelijkheid scherp is en in geval b) de rechterongelijkheid.

(iii) Bewijs dat altijd geldt

$$\frac{1}{c(A)} \; \frac{\|r\|}{\|b\|} \le \frac{\|\delta x_0\|}{\|x\|} \le \frac{\|b\| \, \|A^{-1}\|}{\|x\|} \; \frac{\|r\|}{\|b\|} \le c(A) \; \frac{\|r\|}{\|b\|} \, .$$

Ga na in welke gevallen als regel $\frac{\|\delta x_0\|}{\|x\|} \sim \frac{\|r\|}{\|b\|}$ zal zijn.

Theoretisch gedeelte

C. Construeer een voorbeeld van een matrix (3 × 3 met A_{11} variabel) waarmee U kunt aantonen, dat bij het oplossen zonder rijverwisselingen van een stelsel lineaire vergelijkingen veel nauwkeurigheid (onnodig) verloren kan gaan. Geef aan op welke plaats in het rekenproces het verlies van nauwkeurigheid optreedt, en waarom.
Simuleer (op een tafelrekenmachine) het oplossen van het door U geconstrueerde stelsel met en zonder rijverwisselingen door een decimale machine (mantisse lengte 5).

D. Zij δb een storing in b en δx de hierdoor ontstane storing in x. Zij verder g zodanig dat

$$\frac{\|\delta x\|}{\|x\|} = g\,\frac{\|\delta b\|}{\|b\|} .$$

Zoek de best mogelijke onder- en bovengrens voor g en de waarden van b en δb waarvoor die grenzen bereikt worden.
Maak ook een getallenvoorbeeld waarin g de grenswaarden aanneemt.

E. Voor de invloed van een storing E in de matrix A op de waarde van x geldt

$$\frac{\|\delta x\|}{\|x\|} \le \frac{c(A)\,\frac{\|E\|}{\|A\|}}{1 - c(A)\,\frac{\|E\|}{\|A\|}}$$

Bewijs dit.

F. Er geldt: bij iedere A en x is er een storingsmatrix $E \neq 0$ zodanig dat

$$\frac{c(A)\,\frac{\|E\|}{\|A\|}}{1 + c(A)\,\frac{\|E\|}{\|A\|}} \le \frac{\|\delta x\|}{\|x\|} \le \frac{c(A)\,\frac{\|E\|}{\|A\|}}{1 - c(A)\,\frac{\|E\|}{\|A\|}} .$$

Bewijs dit in het geval van de ∞-norm, bijv. als volgt:

a) $$\max_{x\neq 0} \frac{|v^T x|}{\|x\|_\infty} = \|v\|_1 .$$

b) $$\|uv^T\|_\infty = \|u\|_\infty \|v\|_1 .$$

Bewijs dit op twee manieren: met behulp van $\|A\|_\infty = \max_i \sum_j |A_{ij}|$

en met behulp van $\|A\|_\infty = \max_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty}$.

c) Bij iedere vektor x is er een vektor y zodat $|y^T x| = \|y\|_1 \|x\|_\infty$.

d) Bij iedere matrix A en iedere vektor x is er een matrix B van de vorm $B = uv^T$ zodanig dat

$$\|ABx\|_\infty = \|A\|_\infty \|B\|_\infty \|x\|_\infty .$$

e) Zij δx de storing in x t.g.v. E, dan is

$$(A + E)\delta x = -Ex .$$

Zij Δx zodanig dat

$$A\Delta x = -Ex ,$$

dan geldt

$$\frac{\|\Delta x\|_\infty}{1 + \|A^{-1}\|_\infty \|E\|_\infty} \le \|\delta x\|_\infty \le \frac{\|\Delta x\|_\infty}{1 - \|A^{-1}\|_\infty \|E\|_\infty} .$$

f) De gevraagde ongelijkheid.

<u>Opmerking</u>. Hieruit volgt dat de bovengrens $\dfrac{c(A)\frac{\|E\|}{\|A\|}}{1 - c(A)\frac{\|E\|}{\|A\|}}$ in bovenstaande ongelijkheid praktisch de best mogelijke is.

G. Zij de matrix A tridiagonaal:

$$A = \begin{pmatrix} a_1 & c_1 & & & \\ b_2 & a_2 & c_2 & & \bigcirc \\ & * & * & * & \\ & & b_{n-1} & a_{n-1} & c_{n-1} \\ \bigcirc & & & b_n & a_n \end{pmatrix}$$

Zij bovendien gegeven dat A kolomsgewijs strikt diagonaal dominant is.

a) Bewijs dat er een LU-decompositie voor A bestaat, waarbij L en U voldoen aan

$$L = \begin{pmatrix} 1 & & & & \bigcirc \\ \ell_2 & 1 & & & \\ & * & * & & \\ \bigcirc & & * & * & \\ & & & \ell_n & 1 \end{pmatrix}, \quad U = \begin{pmatrix} u_1 & c_1 & & & \bigcirc \\ & u_2 & c_2 & & \\ & & * & * & \\ \bigcirc & & & u_{n-1} & c_{n-1} \\ & & & & u_n \end{pmatrix}$$

en bewijs dat $|\ell_i| \le 1$ en $\|U\| \le 2\|A\|$.

b) Geef de algoritme voor de LU-decompositie van A.

c) Indien de algoritme door de machine wordt uitgevoerd, dan geldt voor de verkregen L en U

$$LU = A + E_{LU} .$$

Bewijs door de algoritme in detail te analyseren dat

$$\|E_{LU}\| \le 4\ 2^{-t}\|A\| .$$

d) Beschouw nu het stelsel $Ax = b$.
Bepaal een schatting voor de afrondingsfout in de met de oplossingsmethode van Crout berekende oplossing.

Oefeningen Numerieke Wiskunde II | Najaar 1971

Opgave 2. Het bepalen van eigenwaarden en eigenvectoren.

Praktische gedeelte.

A. Schrijf een procedure om alle eigenwaarden en eigenvectoren te bepalen van een symmetrische matrix A met behulp van de methode van Jacobi. Zorg ervoor, dat een gewenste nauwkeurigheid voor de eigenwaarden bereikt wordt.

B. Zij μ een benadering voor een enkelvoudige eigenwaarde λ van de symmetrische matrix A. Schrijf een procedure om de bij λ behorende eigenvector met een gewenste nauwkeurigheid te berekenen.
Doe dit met behulp van de volgende algoritme (inverse Wielandt iteratie):
Zij $\underline{y}_0$ willekeurig en zodanig, dat $\|\underline{y}_0\|_2 = 1$.
Los $\underline{x}_n$ op uit:

$$\left.\begin{array}{l} (A - \mu I)\underline{x}_n = \underline{y}_{n-1} \\ \text{bepaal vervolgens:} \\ \underline{y}_n := \mathrm{sign}(\underline{x}_n, \underline{y}_{n-1}) \dfrac{\underline{x}_n}{\|\underline{x}_n\|_2} \end{array}\right\} \quad n = 1, 2, \ldots$$

C. Test uw procedures met enkele matrices, waarvan de eigenwaarden en eigenvectoren bekend zijn.
Neem ook een geval met een meervoudige eigenwaarde.
Ga in ieder geval na hoe de convergentie verloopt (zorg voor de uitvoer van voldoend veel significante tussenresultaten).

Theoretische gedeelte.

A. Zij $(\lambda, \underline{u})$ een eigenpaar van de symmetrische matrix A en zij B een willekeurige matrix (λ enkelvoudig).
Bewijs, dat A + εB een eigenwaarde $\mu(\varepsilon)$ heeft, waarvoor geldt:

$$\mu(\varepsilon) = \lambda + \varepsilon \frac{\underline{u}^T B \underline{u}}{\underline{u}^T \underline{u}} + O(\varepsilon^2) \; , \; \varepsilon \to 0 \; . \qquad (*)$$

Wat is het analogon, als λ een meervoudige eigenwaarde is?
Gebruik uw programma om de juistheid van (*) te toetsen aan de hand van een voorbeeld.

B. Zij Λ een diagonaalmatrix en E een symmetrische matrix, $|E_{ij}| \leq \varepsilon$.
Het $\Lambda + E$ heeft de eigenwaarden:

$$\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_n$$

en Λ heeft de eigenwaarden:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \; .$$

Zij λ_k een enkelvoudige eigenwaarde en stel

$$d_k = \min_{j \neq k} |\lambda_j - \lambda_k| > 2n\varepsilon \; .$$

Stel $\underline{u}_k$ is de bij α_k behorende eigenvector van $\Lambda + E$.
Bewijs de volgende uitspraken:

1) $$|\alpha_j - \lambda_j| \leq n\,\varepsilon \, , \quad 1 \leq j \leq n \; .$$

2) $$\min_{\gamma} \frac{\|\underline{u}_k - \gamma \underline{e}_k\|_2}{\|\underline{u}_k\|_2} \leq \frac{2n\varepsilon}{d_k} \; .$$

3) $$|\alpha_k - \lambda_k - E_{kk}| \leq \frac{2n\varepsilon^2}{d_k} \; .$$

C. Wat is de betekenis van onderdeel B voor de nauwkeurigheid van uw numerieke resultaten?

D. Zij A een symmetrische matrix met eigenwaarden $\lambda_1,\ldots,\lambda_n$ en genormeerde eigenvectoren $\underline{u}_1,\ldots,\underline{u}_n$.
Stel

$$0 < |\mu - \lambda_1| < \min_{2 \leq i \leq n} |\mu - \lambda_i| \; .$$

Stel

$$(\underline{y}_0, \underline{u}_1) \neq 0 \quad \text{en} \quad \|\underline{y}_0\|_2 = 1 \; .$$

Zij $\{\underline{y}_i\}$ de rij verkregen met de inverse Wielandt iteratie, uitgevoerd zonder afrondingsfouten.
Bewijs dat deze rij convergeert naar $\sigma \underline{u}_1$, waarbij $\sigma = \pm 1$.

3. Bepaling van de rang van een matrix.

A. Inleiding

$\| \ \|$ betekent in het vervolg steeds $\| \ \|_E$. $\mathfrak{M}$ is de verzameling van alle $m\times n$-matrices, m en n zijn constant. $\mathfrak{M}^{(k)}$ is de verzameling van alle $m\times n$- matrices met rang $\le k$. Op het college is de ε-rang van $A \in \mathfrak{M}$ bij gegeven $\varepsilon > 0$ gedefinieerd als

$$k_\varepsilon := \min\{k \mid \exists_{B\in\mathfrak{M}^{(k)}} \|A-B\| \le \varepsilon\|A\|\} .$$

Het bij gegeven matrix A bepalen van de ε-rang en van B impliceert het bepalen van de singular value decompositie van A, hetgeen nogal veel rekenwerk oplevert. Een eenvoudigere methode die een benadering voor k_ε en voor B oplevert wordt hierna beschreven.

B. Bepaling van de benadering voor k_ε en B

Zij $A \in \mathfrak{M}$ en $\varepsilon > 0$ gegeven.
Zij de rij $\{A_k\}$, $k = 0,1,2,\ldots$ als volgt gedefinieerd:

$$A_0 := A$$

$$A_k := Q_k A_{k-1} P_k ,$$

waarbij P_k een permutatiematrix is die de k-de kolom van A_{k-1} verwisselt met een kolom met index $\ge k$ en Q_k een elementaire Householdertransformatie is die de eerste k-1 kolommen van A_{k-1} invariant laat,
zodanig dat

$$A_k = \left(\begin{array}{c|c} R_k & S_k \\ \hline O & T_k \end{array}\right) , \qquad (1)$$

waarin R_k een $k\times k$ bovendriehoeksmatrix is en $\|T_k\|$ minimaal is.
Zij $r := \min\{k \mid \|T_k\| \le \varepsilon\|A\|\}$ dan nemen we r als benadering voor k_ε.
Welke matrix kunnen we dan nemen als benadering voor B?

C. De algoritme

De matrix A_k kan, bij gegeven A_{k-1}, als volgt geconstrueerd worden.
Noem de kolommen van T_{k-1} : $t_k, t_{k+1}, \ldots, t_n$.
Zij

$$\tau(j) := \sum_{i=k}^{n} \frac{(t_i^T t_j)^2}{(t_j^T t_j)} \tag{2}$$

voor $j = k, k+1, \ldots, n$, $t_j \neq 0$
maximaal voor $j = j_k$.
Zij P_k de permutatiematrix die de k-de en de j_k-de kolom van A_{k-1} verwisselt en zij Q_k zodanig dat A_k voldoet aan (1) dan is $\|T_k\|$ minimaal.
Bewijs dit.
Ga na hoe de inproducten uit (2) voor T_k eenvoudig berekend kunnen worden met behulp van de overeenkomstige inproducten voor T_{k-1}.

D. Practisch gedeelte

Schrijf een procedure met als procedure heading

```
integer procedure RANG(m,n,A,ε,Ar,C,γ,p);
value m,n; integer m,n; real ε;
real array A,Ar,C,γ; integer array p;
```

die bij een gegeven $m \times n$-matrix A en bij gegeven $\varepsilon > 0$ de benadering voor k_ε bepaalt.
Na afloop van de procedure bevat het $m \times n$-array A_r de matrix A_r; de k-de kolom van het $m \times n$-array C bevat c_k, die samen met γ_k de Householdertransformatie Q_k bepaalt; het integer array p bevat de nodige informatie over de kolomverwisselingen.

Schrijf een procedure

```
procedure TRANSFORM(m,n,Ar,C,γ,p,B);
```

die bij gegeven m,n,Ar,C, γ en p de matrix B bepaalt waarvoor geldt $\|A-B\| \leq \varepsilon \|A\|$.
Schrijf een programma om de procedures te testen.
Bepaal vervolgens voor een aantal matrices de benadering voor k_ε en voor B.

Oefeningen Numerieke Wiskunde II

Najaar 1970

Opgave 2 : Lineaire Chebyshev approximatie.

A. Het algemene lineaire approximatie probleem.

Zij R een genormeerde lineaire ruimte.

Zij V een n-dimensionale deelruimte van R.

Dan is het algemene lineaire approximatie probleem:

Bepaal bij een gegeven $f \in R$ een $g \in V$ zodanig dat

$$\forall_{h \in V} \quad \|f-g\| \leq \|f-h\| . \tag{1}$$

Definitie. De afstand $\rho(f)$ van f tot V is

$$\rho(f) := \inf_{h \in V} \|f-h\| . \tag{2}$$

Stelling 1. Bij elke $f \in R$ bestaat een $g \in V$ zodanig dat voldaan is aan (1). Bewijs deze stelling.

Opmerking. Men kan in dit geval formule (2) vervangen door

$$\rho(f) := \min_{h \in V} \|f-h\| . \tag{2'}$$

Men noemt een $g \in V$ waarvoor geldt $\rho(f) = \|f-g\|$ een minimum oplossing (beste benadering) in V voor f.

B. Lineaire Chebyshev approximatie.

Zij B een compacte ruimte en zij R := C(B), de lineaire ruimte van de op B reële continue functies.

Zij V de n-dimensionale deelruimte van R opgespannen door de elementen $\varphi_1, \varphi_2, \ldots, \varphi_n$.

Zij de norm in R de Chebyshev-norm, die gedefinieerd is door

$$\|f\| := \max_{x \in B} |f(x)| . \tag{3}$$

De volgende stelling geeft een mogelijke ondergrens voor $\rho(f)$.

Stelling 2. Zij $h_o \in V$ en zij $D \subset B$ zodanig dat

$$\forall_{h \in V} \inf_{x \in D} \left((f(x) - h_o(x)) . h(x)\right) \leq 0. \tag{4}$$

Dan geldt

$$\inf_{x \in D} |f(x) - h_o(x)| \le \rho(f). \tag{5}$$

Bewijs deze stelling.

Gevolg. Voor bovenstaande h_o en D krijgen we de schatting

$$\inf_{x \in D} |f(x) - h_o(x)| \le \rho(f) \le \|f-h_o\| , \tag{6}$$

die aangeeft of h_o dicht bij een minimum oplossing ligt.

De volgende stelling geeft een karakterisering van een minimum oplossing.

Stelling 3. Zij $h_o \in V$ en zij $D := \{x \in B \mid |f(x) - h_o(x)| = \|f-h_o\| \}$. Dan is h_o dan en slechts dan een minimum oplossing in V voor f als voldaan is aan (4).

Bewijs deze stelling.

We voeren nu de voorwaarde in waardoor het lineaire Chebyshev approximatie probleem een eenduidige oplossing heeft.

Definitie. V voldoet aan de Haar conditie als elk element van V, dat niet identiek nul is, ten hoogste n-1 nulpunten in B heeft.

Voorbeelden.

1. Zij R := C[a,b], de verzameling van de continue functies op het gesloten interval [a,b]. Zij V de verzameling van de polynomen waarvan de graad ten hoogste n-1 is. Dan is V een n-dimensionale deelruimte van R die voldoet aan de Haar conditie.

2. B is een verzameling van m punten, m > n. R := C(B) is dan de m-dimensionale cartesische vectorruimte. Zij V de n-dimensionale deelruimte van R opgespannen door de vectoren $a_j := (A_{1j}, A_{2j}, \ldots, A_{mj})$, $j = 1,2,\ldots,n$.

   Zij A de matrix met als kolommen de vectoren a_j, dus

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ & & \\ & & \\ A_{m1} & \cdots & A_{mn} \end{pmatrix}$$

Dan voldoet V aan de Haar conditie als ieder n-tal rijen uit A lineair onafhankelijk is.

Stelling 4. De volgende drie uitspraken zijn equivalent.

a) V voldoet aan de Haar conditie

b) voor elk n-tal punten $x_i \in B$, $i = 1,2,\ldots,n$, geldt

$$\det (\varphi_j(x_i)) \neq 0$$

c) bij elk n-tal punten $x_i \in B$ en elk n-tal reële getallen η_i, $i = 1,2,\ldots,n$, is er precies één $h \in V$ zodanig dat

$$h(x_i) = \eta_i \qquad i = 1,2,\ldots,n\ .$$

Bewijs deze stelling

Stelling 5. Veronderstel dat V voldoet aan de Haar conditie. Zij $g \in V$ een minimum oplossing in V voor $f \in R$. Dan heeft $f-g$ tenminste $n+1$ (absolute) extrema in B.

Bewijs deze stelling.

Stelling 6. Veronderstel dat V voldoet aan de Haar conditie. Dan is er bij elke $f \in R$ precies één minimum oplossing g in V.

Bewijs deze stelling.

Toepassingen.

1. Stelling van De la Vallée Poussin.

   $R := C[a,b]$. V voldoet aan de Haar conditie.

   $D := \{x_i \mid a \le x_0 < x_1 < \ldots < x_n \le b\}$.

   Zij $h \in V$ zodanig dat geldt

   $$f(x_i) - h(x_i) = (-1)^i \lambda_i \qquad\qquad i = 0,1,2,\ldots,n$$

   met $\lambda_i > 0$. Dan geldt

   $$\min_i \lambda_i \le \rho(f).$$

   Ga dit na.

2. Stelling van Chebyshev.

   R, V en D als in 1.

   Zij $g \in V$ zodanig dat geldt

   $$f(x_i) - g(x_i) = \sigma(-1)^i \,\|f-g\| \qquad i = 0,1,\ldots,n$$

met $\sigma = \pm 1$. Dan is g de (eenduidig bepaalde) minimum oplossing voor f.

C. Constructie van de minimum oplossing.

$R := C(B)$ met Chebyshev-norm. V is een lineaire deelruimte opgespannen door de elementen $\varphi_1, \varphi_2, \ldots, \varphi_n$. V voldoet aan de Haar conditie.
Voor iedere $f \in R$ en de bijbehorende (eenduidige bepaalde) minimum oplossing $g \in V$ geldt dat f-g tenminste n+1 extrema in B heeft.
Op grond hiervan en van stelling 3 komen we tot het volgende vermoeden.
Neem n+1 punten $x_i \in B$, $i = 0,1,\ldots,n$.
Zij $D := \{x_i \mid 0 \leq i \leq n\}$. Zij $h \in V$ en het reële getal μ zodanig dat voldaan is aan (4) en $|f(x_i - h(x_i)| = |\mu|$, $i = 0,1,\ldots,n$. Zoek de deelverzameling D waarvoor $|\mu|$ maximaal is.
Dan zal gelden: $|\mu| = \rho(f)$ en de bijbehorende h is de minimum oplossing.

We voeren de volgende notaties in.
Het punt $x := (x_o, x_1, \ldots, x_n)$ met $x_i \in B$.
De open verzameling X van de punten x waarvoor geldt $x_i \neq x_j$ als $i \neq j$.
De rand ∂X van X; dit zijn de punten x waarvoor geldt $x_i = x_j$ voor tenminste één paar $i \neq j$.
De afsluiting $\overline{X}$ van X is $X + \partial X$.

Zij A de $(n+1) \times n$ matrix met elementen $A_{ij} := \varphi_j(x_i)$ voor $i = 0,1,2,\ldots,n$ en $j = 1,2,\ldots,n$.
Zij A_i de matrix die ontstaat uit A door weglaten van de i-de rij, d.i. de rij met elementen $\varphi_j(x_i)$. Zij $\Delta_i := \det(A_i)$.

<u>Stelling 7</u>. Bij elke $x \in X$ bestaat precies één $h \in V$ en een reëel getal μ zodanig dat

$$\forall_{k \in V} \inf_i \left((f(x_i) - h(x_i)) \cdot k(x_i) \right) \leq 0 \qquad (4')$$

en

$$|f(x_i) - h(x_i)| = |\mu| \qquad i = 0,1,\ldots,n .$$

Bewijs deze stelling door te bewijzen dat voor elke $x \in X$ het stelsel vergelijkingen

$$f(x_i) - h(x_i) = \sigma_i \mu \qquad i = 0,1,\ldots,n , \qquad (7)$$

met $\sigma_i = (-1)^i$ sign (Δ_i), een éénduidige oplossing heeft.

<u>Gevolg 1</u>. Voor iedere $x \in X$ voldoen de bijbehorende $\mu(x)$ en $h_x \in V$ aan de ongelijkheid

$$|\mu(x)| \leq \rho(f) \leq \|f-h_x\| . \tag{8}$$

<u>Gevolg 2</u>. Bij gegeven waarde $x \in X$ definiëren we

$$\lambda_i := \frac{(-1)^i \, \Delta_i}{\sum\limits_{j=0}^{n} |\Delta_j|} , \qquad i = 0,1,\ldots,n . \tag{9}$$

Dan geldt

$$\mu(x) = \sum_{i=0}^{n} \lambda_i \, f(x_i) \tag{10}$$

$$\forall_{k \in V} \; \sum_{i=0}^{n} \lambda_i \, k(x_i) = 0 . \tag{11}$$

Verder geldt dat $\{\lambda_i\}$ op een factor $\sigma = \pm 1$ na éénduidig bepaald is door het stelsel vergelijkingen

$$\left.\begin{aligned} &\sum_{i=0}^{n} |\lambda_i| = 1 \\ &\sum_{i=0}^{n} \lambda_i \, \varphi_j(x_i) = 0 \qquad j = 1,2,\ldots,n . \end{aligned}\right. \tag{12}$$

<u>Gevolg 3</u>. Als we definiëren: $\mu(x) = 0$ voor $x \in \partial X$, dan geldt voor elke $x \in \overline{X}$

$$|\mu(x)| = \min_{k \in V} \; \max_{i} \; |f(x_i) - k(x_i)| \tag{13}$$

en dit minimum wordt aangenomen voor $k = h_x$. Met andere woorden h_x is de minimum oplossing voor f als we ons beperken tot $B = \{x_i \mid 0 \leq i \leq n\}$. Ga dit alles na.

<u>Stelling 8</u>. $|\mu|$ is een continue functie van x in $\overline{X}$, die zijn maximum aanneemt in het inwendige, d.i. X.
Bewijs deze stelling.

<u>Stelling 9</u>. Zij $x \in X$. Veronderstel dat voor de bijbehorende $\mu(x)$ en h_x geldt

$X+\partial X$ is compact, staat niet in hun anal. of alg en anal !

$$|\mu(x)| < \|f-h_x\| .$$

Dan is er een $y \in X$ zodanig dat

$$|\mu(x)| < |\mu(y)| .$$

Bewijs dit door het volgende te bewijzen.
Zij $\eta \in B$ zodanig dat $|f(\eta) - h_x(\eta)| > |\mu(x)|$. Dan is er een component x_j van x waarvoor geldt dat het punt y, dat uit x ontstaat door vervanging van x_j door η, het gezochte punt is. Er geldt namelijk

$$|\mu(y)| = |\mu(x)| + \sum_{i=0}^{n} |\lambda_i| \left(|d(y_i)| - |\mu(x)|\right) \tag{14}$$

waarbij $d := f-h_x$ en λ_i de bij y behorende waarden uit (9) zijn.
Bewijs deze stelling.

<u>Gevolg</u>. Zij $x \in X$ een punt waarin $|\mu|$ maximaal is. Dan is de bijbehorende h_x de minimum oplossing en $|\mu(x)| = \rho(f)$.
Ga dit na.

<u>Toepassing</u>. Zij $B := [a,b]$.
Zij $X := \{x \mid a \le x_0 < x_1 < \dots < x_n \le b\}$. Dan geldt voor elke $x \in X$

$$\text{sign } (\Delta_i) = \text{sign } (\Delta_j) \qquad i,j = 0,1,\dots,n . \tag{15}$$

Zij $x \in X$ een punt waarin $|\mu(x)|$ niet maximaal is.
Zij $\eta \in [a,b]$ zodanig dat $|d(\eta)| > |\mu(x)|$ en zij $x_i < \eta < x_{i+1}$.
Dan wordt het punt y uit x verkregen door hetzij x_i, hetzij x_{i+1} te vervangen door η.
Ga dit na.

De volgende stelling geeft een nieuwe karakterisering van de minimum oplossing.

<u>Stelling 10</u>. Het element $g \in V$ is dan en slechts dan de minimum oplossing in V voor $f \in R$ als er een $x \in X$ bestaat zodanig dat

$$f(x_i) - g(x_i) = \sigma \cdot \sigma_i \; \|f-g\| \qquad i = 0,1,2,\dots,n \tag{16}$$

met $\sigma = \pm 1$ en $\sigma_i = (-1)^i \text{ sign } (\Delta_i)$.
Bewijs deze stelling.

Speciaal geval. $B := [a,b]$. $X := \{x \mid a \le x_o < x_1 < \ldots < x_n \le b\}$.
In dit geval is de minimum oplossing gekarakteriseerd door de eigenschap: er is een $x \in X$ waarvoor geldt

$$f(x_i) - g(x_i) = \sigma(-1)^i \;\|f-g\|$$

(vergelijk toepassing 2 op pag. 3). Men noemt dit de alternerende eigenschap van de foutfunctie $d := f-g$. Het punt x wordt een alternante van f genoemd.

D. De Remes algortihme.

We formuleren de (tweede) Remes algorithme voor het bepalen van de beste Chebyshev approximatie in V van een functie f, die gedefinieerd is op het gesloten interval $[a,b]$.

Zij gegeven een $(n+1)$-tal punten $a \le x_o^{(m)} < x_1^{(m)} < \ldots < x_n^{(m)} \le b$.

a. Bepaal $h_m \in V$ en μ_m zodanig dat geldt

$$h_m(x_i^{(m)}) + (-1)^i \mu_m = f(x_i^{(m)}) \qquad i = 0,1,\ldots,n \tag{7'}$$

b. Zij $d_m := f-h_m$.
Bepaal een $(n+1)$-tal punten $a \le x_o^{(m+1)} < x_1^{(m+1)} < \ldots < x_n^{(m+1)} \le b$ zodanig dat geldt

$$\text{sign } (d_m(x_i^{(m+1)})) = - \text{ sign } (d_m(x_{i+1}^{(m+1)})) \qquad i = 0,1,\ldots,n-1$$

$$|d_m(x_i^{(m+1)})| \ge |\mu_m| \qquad i = 0,1,\ldots,n$$

en voor tenminste één indèx k geldt

$$|d_m(x_k^{(m+1)})| \ge |\mu_m| + \theta(\|d_m\| - |\mu_m|).$$

Hierin is $0 < \theta \le 1$, θ onafhankelijk van k en m.

c. Als $\|d_m\| - |\mu_m| > \varepsilon$ dan terug naar a, met $\{x_i^{(m+1)}\}$ in plaats van $\{x_i^{(m)}\}$.

Ga na dat het volgende geldt

1) $\lim_{m\to\infty} |\mu_m|$ bestaat

2) als $\|d_m\| = |\mu_m|$, dan breekt de algorithme af, maar dan is h_m ook de beste benadering.

Stelling 11. De tweede Remes algorithme convergeert voor elke keuze van het startpunt $x^{(0)}$ naar de gewenste oplossing.
Dat wil zeggen: de rij functie $\{h_m\}$ convergeert uniform in [a,b] naar de minimum oplossing g en $\lim_{m\to\infty} |\mu_m| = \rho(f)$.

Voor het bewijs van deze stelling hebben we het volgende lemma nodig.

Lemma. Bij elke $\gamma > 0$ is er een $\delta > 0$ zodanig dat voor elke $x \in X$ geldt: als $|\mu(x)| \geq \gamma$ dan is $|\Delta_i(x)| \geq \delta$ voor $i = 0,1,\ldots,n$.
Bewijs dit lemma.
Bewijs vervolgens de stelling.

De laatste stelling zegt iets over de convergentie snelheid van de algorithme.

Stelling 12. Zij f en φ_j voor $j = 1,2,\ldots,n$ continu differentieerbaar in [a,b] en tweemaal continu differentieerbaar in het open interval (a,b).
Zij g de minimum oplossing voor f. Zij $d := f-g$.
Veronderstel dat f precies één alternante heeft, $x = (x_o,x_1,\ldots,x_n)$ en dat voor de componenten x_i van de alternante geldt:
1. als $x_i \in (a,b)$ dan is $d''(x_i) \neq 0$
2. als $x_i = a$ of $x_i = b$ dan is $d'(x_i) \neq 0$.

Dan is de Remes algorithme kwadratisch convergent, als de componenten van $x^{(m+1)}$ de extrema van d_m zijn, dus

$$d'_m\,(x_i^{(m+1)}) = 0 \text{ tenzij } x_i^{(m+1)} = a \text{ of } x_i^{(m+1)} = b.$$

Bewijs deze stelling.

Opmerking. In het algemeen geldt ook nog dat, onder de voorwaarden van stelling 12, de rij $\{x^{(m)}\}$ kwadratisch convergeert naar de alternante x. Ga dit na.

E. Practisch gedeelte.

Schrijf een procedure

REMES(x, fx, a, b, n, eps, lambda, mu)

voor het bepalen van de beste polynoom benadering van de graad $\leq n$ in de zin van Chebyshev van een functie f gedefinieerd op [a,b], met behulp van de beschreven Remes algorithme.
Neem als basisfuncties voor V: $\varphi_j(x) = x^{j-1}$ of $\varphi_j(x) = T_{j-1}(x)$ het Chebyshev polynoom van de graad j-1.

Schrijf een programma waarmee U Uw procedure kunt testen.
Verifieer hiermee de kwadratische convergentie uit stelling 12 en de daarop volgende opmerking.

Literatuur:

G. Meinardus, Approximation von Funktionen und ihre numerische Behandlung, Springer Verlag, Berlin etc. 1964.

2. Orthogonalisatie volgens Gram Schmidt.

A. Schrijf een ALGOL programma om de QR decompositie van een $m \times n$ matrix A te berekenen.

Doe dit op twee manieren. Gebruik de op college behandelde algoritme 1 (klassiek Gram Schmidt) en de gewijzigde vorm hiervan algoritme 2.
Noem de met algoritme 1 verkregen Q en R $Q^{(1)}$ en $R^{(1)}$ en de met algoritme 2 verkregen Q en R $Q^{(2)}$ en $R^{(2)}$.
Bereken bovendien met behulp van $Q^{(2)}$ de matrix $Q^{(3)}$ als volgt:

```
for j := 1 step 1 until m do
for i := 1 step 1 until n do
begin Q(3)ji := qiT ej; ej := ej - Q(3)ji × qi end
```

$$\underline{\textbf{for}}\ j := 1\ \underline{\textbf{step}}\ 1\ \underline{\textbf{until}}\ m\ \underline{\textbf{do}}$$
$$\underline{\textbf{for}}\ i := 1\ \underline{\textbf{step}}\ 1\ \underline{\textbf{until}}\ n\ \underline{\textbf{do}}$$
$$\underline{\textbf{begin}}\ Q_{ji}^{(3)} := q_i^T e_j;\ e_j := e_j - Q_{ji}^{(3)} \times q_i\ \underline{\textbf{end}}$$

hierin is q_i de i-de kolom van $Q^{(2)}$ en e_j initieel de j-de kolom van de $m \times m$ matrix I.
Bereken ook

$$E_1^{(i)} := A - Q^{(i)}R^{(i)}\ ,\quad E_2^{(i)} := Q^{(i)^T}A - R^{(i)}\ ,\quad E_3^{(i)} := Q^{(i)^T}Q^{(i)} - I$$

voor $i = 1,2$

en

$$E_1^{(3)} := A - Q^{(3)}R^{(2)}\ ,\quad E_2^{(3)} := Q^{(3)^T}A - R^{(2)}\ ,\quad E_3^{(3)} := Q^{(3)^T}Q^{(3)} - I\ .$$

Bereken tenslotte c(A).
Test de correctheid van Uw programma met enkele voorbeelden.

B. Behandel met Uw programma matrices met een conditiegetal variërend van klein naar groot.
Vergelijk de waarden van

$$\frac{\|E_1^{(i)}\|}{\|A\|} \qquad \text{voor } i = 1,2,3$$

met elkaar en waar mogelijk met een theoretische schatting.
Doe hetzelfde voor

$$\frac{\|E_2^{(i)}\|}{\|A\|} \quad \text{en} \quad \|E_3^{(i)}\|\ .$$

C. Beschouw de matrix

$$A = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix},$$

met ε zodanig dat $\varepsilon^2 \le 2^{-t}$.

Bereken voor dit geval $Q^{(i)}$ en $E_3^{(i)}$ voor $i = 1,2,3$.

Oefeningen Numerieke Wiskunde II   najaar 1973

2. Kleinste kwadraten aanpassing.

Inleiding.

Stel $\mathfrak{M}^n_{m,n}$ is de verzameling van alle $m \times n$ matrices met rang n. Zij $A \in \mathfrak{M}^n_{m,n}$ en $b \in \mathbb{R}^m$. Zij x de bij A en b behorende kleinste kwadratenoplossing (afgekort: kk-oplossing).

In het volgende zullen we een aantal methoden onderzoeken om x numeriek te benaderen (de benadering noemen we $\bar{x}$).

We zullen een methode stabiel noemen als er een ΔA en een Δb bestaan zodanig dat

(i) $\bar{x}$ is de exacte kk-oplossing behorende bij $A + \Delta A$ en $b + \Delta b$,

(ii) $\|\Delta A\| \le \eta\|A\|$,

(iii) $\|\Delta b\| \le \eta\|b\|$,

waarbij $\eta = f(m,n)2^{-t}$.

Tenzij anders vermeld betekent $\|\ \|$ steeds $\|\ \|_2$.

In het volgende zal steeds gelden: $A \in \mathfrak{M}^n_{m,n}$, $Q \in \mathfrak{M}^n_{m,n}$, $R \in \mathfrak{M}^n_{n,n}$.

Theoretisch gedeelte.

1. Stelling 1.

Zij

$$R = Q^TA + E$$

$$Rx = Q^Tb + e\ ,$$

dan is x de kk-oplossing behorende bij

$$\bar{A} := A + (Q^+)^TE - (I - QQ^+)A$$

en

$$\bar{b} := b + (Q^+)^Te\ .$$

Bewijs deze stelling.

Lemma.

Zij

$$Q^TQ = I + E\ , \quad \|E\| < 1\ ,$$

dan geldt

$$\|Q^+\| \le (1 - \|E\|)^{-\frac{1}{2}} .$$

Bewijs dit.

<u>Stelling 2</u>.

Zij

$$R = Q^T A + E_1$$

$$A = QC + E_2$$

$$Q^T Q = I + E_3$$

$$Rx = Q^T b + e ,$$

dan is x de kk-oplossing behorende bij $A + \Delta A$ en $b + \Delta b$, waarbij

$$\|\Delta A\| \le \|E_1\|(1 - \|E_3\|)^{-\frac{1}{2}} + \|(I - QQ^+)E_2\|$$

$$\|\Delta b\| \le \|e\|(1 - \|E_3\|)^{-\frac{1}{2}} .$$

Bewijs deze stelling.

Welke relatie bestaat er tussen E_1, E_2 en E_3 ?

2. Beschouw de volgende algoritme:

$C := A^T A$

$R^T R := C$ (via Choleski)

$c := A^T b$

los d op uit $R^T d = c$

los x op uit $Rx = d$.

Onderzoek de stabiliteit van deze algoritme.
(Hint: Geef een schatting voor de voorwaartse fout in x en vergelijk dit met de schatting voor de fout in x zoals die geldt voor een stabiele algoritme.)

Praktisch gedeelte.

3. Op grond van stelling 2 loont het de moeite om bij A een QR-decompositie te maken, waarbij E_1 klein is, waarbij de kolommenruimten van A en Q numeriek goed overeenstemmen (E_2 klein) en waarbij bovendien $\|E_3\|$ niet te dicht bij 1 ligt.
Schrijf twee procedures om van een gegeven matrix A een QR-decompositie te berekenen, één gebaseerd op het klassieke Gram-Schmidt proces en de andere gebaseerd op het gemodificeerde Gram-Schmidt proces.
Gebruik deze procedures in een programma om bij gegeven A en b een benadering voor de kk-oplossing te berekenen.
Uw programma moet bovendien in beide gevallen berekenen:
$A\bar{x}$; $Q^TA - R$; $A - QR$; $I - Q^TQ$; $Q^Tb - Rx$; $\|A\|_\infty$ en een schatting voor $c_2(A)$.
Behandel met Uw programma matrices met een conditiegetal variërend van klein naar groot; U kunt hierbij gebruik maken van de procedure ILLCONDMATRIX.
Neem bij iedere A een aantal rechterleden b, die goed in de kolommenruimte van A zitten ("goede aanpassing mogelijk") en een aantal rechterleden waarbij dit niet het geval is. Neem bijvoorbeeld n 5 à 10 en m ~ 3n.
Ga met behulp van stelling 2 na of de beide methoden stabiel zijn.

Foutenanalyse.

4. Ga na dat bij het klassieke Gram-Schmidt proces matrices Q en R worden geproduceerd waarvoor geldt:

$$Q^TA - R = L + E_1$$

$$A = QR + E_2$$

$$Q^TQ = I + U + U^T + E_3 \,,$$

waarbij $\|E_1\| \leq \eta\|A\|$, $\|E_2\| \leq \eta\|A\|$ en $\|E_3\| \leq \eta$, L linksonder en U strikt rechtsboven is.
Ga na dat L en U nul zijn, als E_1, E_2 en E_3 nul zijn. Laat vervolgens zien dat, als E_1, E_2 en E_3 niet nul zijn, L en U minstens een factor $c_2(A)$ zullen bevatten.
Vergelijk deze uitspraken met Uw numerieke resultaten.

5. Het gemodificeerde Gram-Schmidt proces produceert getallen R_{kj} en vectoren q_k en $a_j^{(k)}$ zodanig dat

$$a_j^{(0)} = a_j$$

$$R_{kj} = q_k^T a_j^{(k-1)} + \rho_{kj} , \quad k \le j$$

$$a_j^{(k)} = a_j^{(k-1)} - R_{kj} q_k + e_j^{(k)} , \quad k < j$$

$$0 = a_j^{(j-1)} - R_{jj} q_j + e_j^{(j)} .$$

Geef schattingen voor ρ_{kj} en $e_j^{(k)}$.
Als $A = QR + E_2$, geef dan een schatting voor $\|E_2\|$.
Definieer

$$P_k := I - q_k q_k^T$$

$$\tilde{q}_k := P_1 P_2 \ldots P_{k-1} q_k .$$

Als $\tilde{Q}^T A = R + E_1$, geef dan een schatting voor $\|E_1\|$.
Bewijs dat $Q = \tilde{Q}(I + U)$, waarbij U strikt rechtsboven is. Geef een schatting voor U.
Geef vervolgens schattingen voor $\|Q^T A - R\|$, $\|I - Q^T Q\|$ en $\|Q^T b - Rx\|$ en verifieer de uitspraak over de stabiliteit van de algoritme gebaseerd op het gemodificeerde Gram-Schmidt proces.
Geef schattingen voor $\|I - \tilde{Q}^T\tilde{Q}\|$ en $\|A - \tilde{Q}R\|$.
Ga na hoe $\tilde{Q}^T b$ berekend kan worden.
Laat met behulp van het voorafgaande zien dat de algoritme stabiel is, als in plaats van $Q^T b$ met $\tilde{Q}^T b$ gewerkt wordt.

6. Implementeer de boven gevonden wijziging.
Bereken bovendien $\tilde{Q}$, $\|\tilde{Q}^T A - R\|$, $\|A - \tilde{Q}R\|$ en $\|I - \tilde{Q}^T\tilde{Q}\|$.
Vergelijk Uw uitkomsten met de theoretische schattingen.