Numerieke Analyse 2003–2004. Test 12 maart 2004.

NAAM : Katrijn Standaert

1. Beschouw de hypothetische computer met basis $B = 2$ en mantisselengte $t = 24$ (32-bit machine). Onderstel dat

$$x = 2^N(0.d_{-1}d_{-2}\cdots d_{-24}d_{-25}d_{-26}\cdots)_2 ,$$

met $d_{-1} = 1$. De twee dichtstbijzijnde machinegetallen zijn :

$$x' = 2^N(0.d_{-1}d_{-2}\cdots d_{-24})_2 , x'' = 2^N((0.d_{-1}d_{-2}\cdots d_{-24})_2 + 2^{-24}) .$$

Het getal x' of x'', dat het dichtst bij x ligt, wordt gekozen voor de vlottende-puntvoorstelling $fl(x)$. Als $fl(x) = x'$, geldt

$$|x - x'| \leq \frac{1}{2}|x'' - x'| = \frac{1}{2}2^{-24}2^N = 2^{N-25} ; \qquad (1)$$

in het geval $fl(x) = x''$ geldt eveneens $|x - x''| \leq 2^{N-25}$. De absolute fout is dus $|x - fl(x)| \leq 2^{N-25}$; de relatieve fout is

$$\left|\frac{x - fl(x)}{x}\right| \leq \frac{2^{N-25}}{2^N(0.d_{-1}d_{-2}\cdots)} \leq \frac{2^{-25}}{1/2} = 2^{-24} = \epsilon . \qquad (2)$$

Gevraagd :

- Verklaar de ongelijkheid in (1).

het getal x is gelegen tussen x' en x'', dus is $|x - x'| \leq \frac{1}{2}|x'' - x'|$. Aangezien $fl(x) = x'$ ligt x tussen x' en de helft van de afstand ~~van~~ tussen x' en x''.

$(a \leq x < a + \frac{B}{2} < a + B = b \Rightarrow |x - a| \leq \frac{1}{2}|b - a| = \frac{B}{2})$

✓

- Verklaar de tweede ongelijkheid in (2).

$$\frac{2^{N-25}}{2^N(0.d_{-1}d_{-2}\cdots)} = \frac{2^{-25}}{(0.d_{-1}d_{-2}\cdots)} \leq \frac{2^{-25}}{\frac{1}{2}}$$

aangezien $0.d_{-1}d_{-2}\cdots \geq \frac{1}{2}$,

✓

- Hoeveel verschillende positieve machinegetallen kunnen voorgesteld worden op deze machine?

$2^{~~24~~\,23}$ getallen per waarde 2^N

De machine werkt tot 2^{-24} nauwkeurig

2^{31} 24 mantisse 32 totaal → $N = 8 \Rightarrow 2^{23} \cdot 2^8 = 2^{31}$

$\Rightarrow 2^{31}$

2. Zij V een lineaire ruimte over $\mathbb{C}$. De afbeelding $\langle\cdot,\cdot\rangle: V \times V \to \mathbb{C}$ is een inproduct als

i) $\langle f+g,h\rangle = \langle f,h\rangle + \langle g,h\rangle$	(lineariteit)
ii) $\langle \alpha f, g\rangle = \alpha\langle f,g\rangle$	(homogeniteit)
iii) $\langle g,f\rangle = \overline{\langle f,g\rangle}$ (complexe toevoeging)	(symmetrie)
iv) $\langle f,f\rangle > 0 \quad \forall f \neq 0$.	(positiviteit)

Uit de eigenschappen van $\langle\cdot,\cdot\rangle$ volgt dan dat $(V, \|\cdot\|)$ met $\|f\| = \sqrt{\langle f,f\rangle}$ een genormeerde ruimte is. De driehoeksongelijkheid volgt uit de Schwarz-ongelijkheid :

$$|\langle f,g\rangle| \leq \|f\|\,\|g\| .$$

Om dit laatste te bewijzen, onderstel dat $f \neq 0$ en $g \neq 0$. Zij $\lambda \in \mathbb{C}$, dan :

$$\|\lambda f + g\|^2 = \lambda\overline{\lambda}\langle f,f\rangle + \lambda\langle f,g\rangle + \overline{\lambda}\langle g,f\rangle + \langle g,g\rangle \geq 0 .$$

Kiezen we $\lambda = -\langle g,f\rangle/\langle f,f\rangle$, dan is $\overline{\lambda} = -\langle f,g\rangle/\langle f,f\rangle$ en komt er :

$$|\langle f,g\rangle|^2 \leq \langle f,f\rangle\langle g,g\rangle .$$

Bijgevolg is :

$$\begin{aligned} \|f+g\|^2 &= \langle f+g, f+g\rangle = \|f\|^2 + \|g\|^2 + \langle f,g\rangle + \langle g,f\rangle \\ &\leq \|f\|^2 + \|g\|^2 + 2|\langle f,g\rangle| \leq (\|f\| + \|g\|)^2 . \end{aligned}$$

Hieruit zien we dat de gelijkheid $\|f+g\| = \|f\| + \|g\|$ slechts kan optreden indien ook beide leden van de Schwarz-ongelijkheid gelijk zijn; dit impliceert dat $\|\lambda f + g\|^2 = 0$, dus dat f en g lineair afhankelijk zijn. M.a.w. de norm geïnduceerd door een inproduct is strikt.
Gevraagd :

- Toon aan dat de bovenstaande definitie voor $\|f\|$ voldoet aan : $\|f\| = 0 \Leftrightarrow f = 0$. Op welke van de vier bovenstaande eigenschappen steun je?

$\sqrt{\langle f,f\rangle} = 0 \Rightarrow \langle f,f\rangle = 0$ en $\langle f,f\rangle = \overline{\langle f,f\rangle}$, dus moet het reële deel van f gelijk zijn aan 0 (eigenschap iii)

$f = 0 \Rightarrow \langle f,f\rangle = \langle 0f, f\rangle = 0\langle f,f\rangle = 0$ dus $\|f\| = 0$ iii)

$\|f\| = 0 \Rightarrow f$ kan wegens iv niet $\neq$ zijn van 0

- Waarom is $\langle f,g\rangle + \langle g,f\rangle \leq 2|\langle f,g\rangle|$?

$\langle f,g\rangle + \langle g,f\rangle = \langle f,g\rangle + \overline{\langle f,g\rangle}$

$= 2 \times$ reëel deel van $\langle f,g\rangle$ (imaginair deel valt weg)

en $2|\langle f,g\rangle| = 2 \times$ ~~een deel van~~ $\langle f,g\rangle$ (abs. waarde van $\langle f,g\rangle$ = modulus van $\langle f,g\rangle$ $\Rightarrow |a+ib| = \sqrt{a^2+b^2}$)

$\Rightarrow \langle f,g\rangle + \langle g,f\rangle = \langle f,g\rangle + \overline{\langle f,g\rangle} = 2\,\mathrm{Re}\langle f,g\rangle \leq 2\sqrt{\mathrm{Re}(\langle f,g\rangle)^2} \leq 2\sqrt{\mathrm{Re}(\langle f,g\rangle)^2 + \mathrm{Im}(\langle f,g\rangle)^2} = 2|\langle f,g\rangle|$

enkel gelijkheid als ... reëel is

- Als $\|f+g\| = \|f\| + \|g\|$, wat kan je dan zeggen over het reëel gedeelte van $\langle f, g\rangle$?

$\|f+g\|^2 = \langle f+g, f+g\rangle = \langle f,f\rangle + \langle f,g\rangle + \overline{\langle f,g\rangle} + \langle g,g\rangle$

$(\|f\|+\|g\|)^2 = \langle f,f\rangle + \langle g,g\rangle + 2\sqrt{\langle f,f\rangle\langle g,g\rangle}$

$\Rightarrow \langle f,g\rangle + \overline{\langle f,g\rangle} = 2\sqrt{\langle f,f\rangle\langle g,g\rangle} = 2\,Re(\langle f,g\rangle) \Rightarrow$ reëel deel van $\langle f,g\rangle = \|f\|\|g\|$

3. Neem aan dat we werken op een machine met basis $B = 10$ en mantisselengte $t = 2$. Als we Gauss-eliminatie rechtstreeks toepassen op het stelsel

$$\begin{pmatrix} 0.005 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}, \tag{3}$$

dan vinden we

$$\begin{matrix} R_1 \\ R_2 \end{matrix} \left(\begin{array}{cc|c} 0.005 & 1 & 0.5 \\ 1 & 1 & 1 \end{array}\right) \rightarrow \left(\begin{array}{cc|c} 0.005 & 1 & 0.5 \\ 0 & -200 & -99 \end{array}\right), \tag{4}$$

met als oplossing $x_2 = 0.50$ en $x_1 = 0.00$.
Gevraagd :

- Leg uit hoe de waarde -200 in (4) werd berekend.

Rij 2 wordt vervangen door (rij 2 - 200x rij 1), om 20 een nul te krijgen onder 0,005. Normaal krijg je dan -199, maar aangezien de mantisselengte 2 is, moet je afronden naar het dichtstbijzijnde veelvoud van 10 (in dit geval dus naar -200)

- Waarom werd een oplossing gevonden die sterk afwijkt van de echte oplossing van het stelsel? Hoe kan men (in het algemeen) dit probleem voorkomen?

Omdat het element 0,005 (relatief gezien) erg klein is in verhouding met de andere getallen. Door op deze manier verder te werken wordt de afrondingsfout met een grote factor vermenigvuldigd, wat kan vermeden worden door de rijen te pivoteren (in dit geval rij 2 en rij 1 omwisselen)

- Geef L en U expliciet voor het bovenstaande voorbeeld (cfr. LU-decompositie), zoals ze berekend worden op deze machine. Bereken eveneens het product LU (ook op

deze machine). $\ell_{ij} = a_{ij}^{(j)} / a_{jj}^{(j)}$

$L = \begin{pmatrix} 1 & 0 \\ 1000 & 1 \end{pmatrix}$ $\quad U = \begin{pmatrix} 0,005 & 1 \\ 0 & -999 \end{pmatrix}$

$LU = \begin{pmatrix} 0,005 & 1 \\ 5 & 0 \end{pmatrix}$

4. Voor gegeven $x \in \mathbb{R}^n$ kan men steeds een Householder matrix $H \in \mathbb{R}^{n \times n}$ berekenen zodanig dat $Hx = \sigma e_1$, nl. als volgt :

$$\begin{array}{ll} \text{(i)} & u = (\operatorname{sgn}(x_1)(|x_1| + \|x\|_2), x_2, \ldots, x_n) \ , \\ \text{(ii)} & \beta = 1/(\|x\|_2^2 + |x_1| \, \|x\|_2) \ , \\ \text{(iii)} & H = I - \beta u u^T \ . \end{array} \tag{5}$$

Zulke matrices kan men dan gebruiken om voor een matrix $A \in \mathbb{R}^{n \times n}$ een QR-decompositie te bepalen. Als A niet-singulier is, kan men de QR-decompositie gebruiken om $Ax = b$ op te lossen. Dit geeft aanleiding tot het volgende algoritme, met als input n en $C = (A|b) \in \mathbb{R}^{n \times (n+1)}$:

voor $k = 1$ tot $n - 1$, stel

$$\begin{aligned} s &= \left(\sum_{j=k}^{n} c_{jk}^2 \right)^{1/2} \\ \beta &= (s(|c_{kk}| + s))^{-1} \\ u &= (0, 0, \ldots, 0, c_{kk} + \operatorname{sgn}(c_{kk}) s, c_{k+1,k}, \ldots, c_{nk})^T \\ H^{(k)} &= I - \beta u u^T \\ C &= H^{(k)} C \end{aligned}$$

In een programma zal men een aantal vereenvoudigingen aanbrengen. Om $H^{(k)}C$ te berekenen, stelt men $v = C^T u$. Dan wordt $H^{(k)}C = C - \beta u v^T$. We kunnen dan het volgende algoritme opstellen.

Algoritme *Householder*

for $k = 1 : n - 1$
  $s = \mathbf{sqrt}(C(k:n,k)C(k:n,k))$;
  $\beta = 1/(s(\mathbf{abs}(C(k,k)) + s))$;
  $u(k) = C(k,k) + \mathbf{sgn}(C(k,k))s$;
  $u(k+1:n) = C(k+1:n,k)$;
  for $j = k : n+1$
    $v(j) = C(k:n,j)u(k:n)$
  end;
  for $i = k : n$
    for $j = k : n+1$
      $C(i,j) = C(i,j) - \beta u(i) v(j)$

end

end

end;

Gevraagd :

- Wat stelt de grootheid s voor?

$C=(A|b)$

$s = \left(\sum_{j=k}^{n} c_{jk}^2\right)^{1/2}$ is de norm $\|.\|_2$ van de vector a waarbij a de vector is die bestaat uit het diagonaalelement van A in de k-de kolom en alle elementen van onder

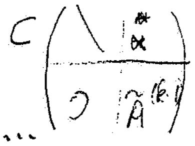
V

- Waarom herleidt men de berekening van $H^{(k)}C$ tot $C - \beta u v^T$ i.p.v. dit expliciet te berekenen?

omdat er minder berekeningen nodig zijn om $\beta u v^T$ te berekenen en dit van C af te trekken dan om beide matrices te vermenigvuldigen, zeker voor grote n

V

- In het Algoritme berekent men de componenten van v slechts voor de indices $j = k : n+1$, want $v_1 = \ldots = v_{k-1} = 0$. Leg uit waarom deze eerste $(k-1)$ componenten nul zijn.



$v = C^T u$ waarbij in de vorige stap steeds $u(k-1)$ nul werd en de elementen $C(k-1,j)$ en $C(j,k-1)$ ook, zodat in het product $v(k-1)$ nul wordt

om v_i met $i = 1, \ldots, k-1$ te berekenen worden de elementen van de i-de rij in C vermenigvuldigd met elementen van de vector u, maar dit is dus 0 omdat in C de elementen van de eerste k-kolommen onder de diagonaal nul zijn

- Welke componenten van $C(i,j)$ worden 0 in stap k?

$C(k,j)$ voor $j>k$

en $C(j,k)$ voor $j>k$

—

$C(k+1:n, k)$

5. Zijn de volgende uitspraken waar of vals? Geef telkens een uitleg voor jouw antwoord.

- Als men rekent op een 32-bit machine (cfr. eerste vraag) is $(x+y)+z$ niet noodzakelijk gelijk aan $(z+y)+x$ (hierbij stellen x, y en z machinegetallen voor).

vals : als x, y en z machinegetallen zijn, dan kunnen we problemen krijgen in de ..., omdat dat ... Hierdoor krijg je bij $(x+y)+z$ en $(z+y)+x$ ... resultaten. (*)

- Voor $A \in \mathbb{R}^{n\times n}$ geldt $\|A\|_1 = \|A^T\|_\infty$.

waar : $\|A\|_1 = \max_j \sum_i |a_{ij}|$ geeft het maximum van de kolomsommen

$\|A\|_\infty$ geeft het maximum van de rijsommen $= \max_i \sum_j |a_{ij}|$

Door te transponeren worden de kolommen rijen, dus de uitspraak

- Als $A \in \mathbb{R}^{n\times n}$ niet-singulier is, bezit ze een unieke rechtstreekse LU-factorisatie.

waar als A niet singulier is, zijn $\forall i \in \{0,1,\dots,n\} : a_{ii} \neq 0$

dus geeft de LU-factorisatie geen problemen en

alle principale minors moeten ... $\neq 0$ zijn $\quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

- Tridiagonale matrices bezitten altijd een rechtstreekse LU-decompositie.

vals tridiagonale matrices bezitten geen rechtstreekse LU-decompositie. (~~1 ... diagonaal element ... a_{ii} nul is.~~)

een α_i nul is, ~~als~~ met $\alpha_i = a_i - \beta_i c_{i-1}$ en $\beta_i = \frac{b_i}{\alpha_{i-1}}$

$(\alpha_1 = a_1)$

(*) waar omdat ~~...~~ $x+y$ en $z+y$ misschien geen machinegetallen zijn en dan rond hij af. Als daarbij de waarde z resp. x opgeteld is, is het geen exact getal meer en blijft de afronding ... behouden

Numerieke Analyse 2003–2004. Eerste examenperiode.

OEFENINGEN

1. Beschouw de matrix $A \in \mathbb{R}^{n\times n}$ met $a_{ij} = 1$ als $i \neq j$ en $a_{ii} = 1 + a$, waarbij a een reëel getal is. Bepaal de 2-norm $\|A\|_2$.

2. Zij $A \in \mathbb{R}^{n\times n}$ een reële, symmetrische, tridiagonale en positief definiete matrix van de vorm

$$A = \begin{pmatrix} a_1 & b_2 & 0 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & 0 & & \vdots \\ 0 & b_3 & a_3 & b_4 & & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & b_{n-1} & a_{n-1} & b_n \\ 0 & \cdots & \cdots & 0 & b_n & a_n \end{pmatrix}.$$

   (a) Beschouw de Cholesky-decompositie GG^T van A: bepaal de vorm van G en hoe je de elementen van G berekent. Beschouw vervolgens het stelsel $Ax = f$ ($x, f \in \mathbb{R}^n$), met gegeven rechterlid f, en bepaal hoe je dit stelsel oplost als de Cholesky-decompositie van A gekend is.

   (b) Vertaal de berekeningen uit (a) in een **Algoritme** in pseudocode. Gebruik voor de voorstelling van de matrix A slechts twee rijen (arrays), nl. $a(i)$ en $b(i)$ ($i = 1, \ldots, n$), en voer voor de matrix G geen nieuwe rijen in (m.a.w. stop de waarden in de gepaste 'geheugenplaatsen' $a(i)$ en $b(i)$). Voorzie voor het oplossen van het stelsel ook slechts één bijkomende array $f(i)$ ($i = 1, \ldots, n$), m.a.w. de uiteindelijke oplossing x vindt men na afloop van het algoritme terug in de array f.

   (c) Hoeveel elementaire bewerkingen vraagt deze factorisatie en het oplossen van het stelsel? (Reken hierbij de vierkantswortel nemen als één bewerking.)

3. Bepaal de monische vijfdegraadsveelterm(en) $p(x)$ waarvoor

$$M = \max_{x\in[-1,0]} |\frac{1}{5}p'(x) + \frac{1}{128}| \qquad (p'(x) = \frac{dp}{dx})$$

   minimaal is. Wat is de waarde van M?

4. Beschouw de volgende kwadratuurformule :

$$\int_{-1}^{1} f(x)dx \approx Af(1) + Bf'(0) + Cf''(0).$$

   a) Bepaal A, B en C zodat de GVAN zo hoog mogelijk wordt. Wat is de GVAN?

   b) Bereken de Peano kern $K(t)$ voor de aldus bekomen kwadratuurformule.

   c) Indien $K(t)$ een vast teken heeft in $[-1, 1]$, stel dan een expliciete uitdrukking op voor de procesfout van de kwadratuurformule.