Seppo Linnainmaa

ALGORITMIN KUMULATIIVINEN PYÖRISTYSVIRHE YKSITTÄISTEN PYÖRISTYSVIRHEIDEN TAYLOR-KEHITELMÄNÄ

Pro gradu-tutkielma · ohjaaja professori M.Tienari

Tapiola 1970

YHTEENVETO

Algoritmien kumulatiivista pyöristysvirhettä liukuvan pilkun aritmetiikassa pyritään analysoimaan kehittämällä tämä yksittäisten pyöristysvirheiden Taylorkehitelmäksi. Tähän tarkoitukseen esitetään sekä analyyttinen että tietokoneelle soveltuva menetelmä. Yksittäiselle suhteelliselle pyöristysvirheelle konstruoidaan tilastollinen malli. Sovellutuksina käsitellään pienalgoritmia $a^2 - b^2 = (a+b)\cdot(a-b)$ sekä Hornershemaa ja Gauss-Jordanin matriisinkääntöalgoritmia. Sovellutuksiin liittyvät ohjelmat on ajettu IBM 7094-tietokoneella.

SISÄLLYS

1. JOHDANTO

Tietokoneiden mahdollistaman suuren laskentanopeuden johdosta on tullut välttämättömäksi pyrkiä selvittämään pyöristysvirheiden vaikutuksia, koska niiden merkitys kasvaa nopeasti peräkkäisten laskutoimitusten määrän kasvaessa [1].

Kysymystä lähestyttäessä on lähtökohtana tosiasia, että algoritmissa tapahtuva kumulatiivinen pyöristysvirhe on peräisin kussakin erillisessä laskutoimituksessa syntyvästä yksittäisestä pyöristysvirheestä. Näiden avulla on pyritty laskemaan ylärajoja kumulatiiviselle pyöristysvirheelle. Laajoissa algoritmeissa on ollut pakko tyytyä melko karkeisiin arvioihin tälle ylärajalle [10].

Useissa tapauksissa tieto keskimääräisestä virheestä ja sen vaihtelualttiudesta olisi hyödyllisempi kuin virheen yläraja. Tässä tarkoituksessa on pyöristysvirheitä pyritty selittämään tilastollisten jakautumien avulla. Professori Henrici on mm. teoksessaan 'Elements of Numerical Analysis' [4] laajalti lähestynyt kysymystä tältä kannalta. Hän käsittelee pääasiassa kiinteän pilkun aritmetiikkaa. Seuraavassa esityksessä pyritään tekemään vastaavia huomioita liukuvan pilkun aritmetiikasta.

2. PYÖRISTYSVIRHEEN KÄSITE

Mielivaltainen nollasta eroava kymmenjärjestelmän reaaliluku voidaan esittää muodossa

$$z = \pm 0.d_{-1}d_{-2}d_{-3}\ldots \cdot 10^{p} , \qquad (2.1)$$

missä $d_{-1}, d_{-2}, \ldots$ ovat numeroita, $d_{-1} \neq 0$ ja p on kokonaisluku. Osaa

$$m = \pm 0.d_{-1}d_{-2}d_{-3}\ldots \qquad (2.2)$$

kutsutaan luvun z mantissaosaksi ja sillä tarkoitetaan lukua

$$m = \pm(d_{-1}\cdot 10^{-1}+d_{-2}\cdot 10^{-2}+d_{-3}\cdot 10^{-3}+ \ldots). \qquad (2.3)$$

Osaa 10^{p} kutsutaan eksponenttiosaksi, 10 on kantaluku ja p on eksponentti.

Tietokoneissa kantalukuna b on luvun 10 sijasta useimmin jokin kahden potenssi (esim. $2^1 = 2$, $2^3 = 8$ tai $2^4 = 16$). Myös tällöin luku voidaan esittää vastaavassa muodossa [4],[7]

$$z = m\cdot b^{p} = \pm 0.d_{-1}d_{-2}d_{-3}\ldots \cdot b^{p}, \quad d_{-1} \neq 0, \qquad (2.4)$$

missä $d_{-1}, d_{-2}, \ldots$ ovat b-kantaisen lukujärjestelmän numeroita. Tällöin mantissalla m tarkoitetaan lukua

$$m = \pm(d_{-1}\cdot b^{-1}+d_{-2}\cdot b^{-2}+d_{-3}\cdot b^{-3}+ \ldots). \qquad (2.5)$$

Ehdosta $d_{-1} \neq 0$ seuraa

$$b^{-1} \le |m| < 1 . \qquad (2.6)$$

Mantissaosaan mahtuu tietokoneessa rajoitettu määrä, t kpl, numeroita, jolloin luku z koneesta riippuen joko katkaistaan tai pyöristetään 'oikein' muotoon

$$z^* = \pm 0.d_{-1}d_{-2}\ldots d_{-t}\cdot b^p \qquad (2.7)$$

Tätä muotoa kutsutaan liukuvan pilkun esitykseksi ja lukua z^* liukuluvuksi.

Pyöristetyn ja tarkan luvun erotusta

$$r = z^*-z \qquad (2.8)$$

kutsutaan absoluuttiseksi pyöristysvirheeksi. Koska absoluuttisen pyöristysvirheen suuruus liukulukujen yhteydessä riippuu ratkaisevasti itse luvun z suuruudesta, on usein käytännöllisempää tarkastella suhteellista pyöristysvirhettä

$$e = \frac{z^*-z}{z}\ , \qquad (2.9)$$

jolloin

$$z^* = z(1+e)\ . \qquad (2.10)$$

3. SUHTEELLISEN PYÖRISTYSVIRHEEN JAKAUTUMA

Mantissaosan pyöristysvirhe

Suhteellisen pyöristysvirheen tilastollisen jakautuman selvittämiseksi on syytä tarkastella aluksi pelkästään mantissaosaa. Merkitsemme pyöristetyn luvun z^* mantissaosaa m^*:lla, jolloin

$$\varepsilon = m^* - m \tag{3.1}$$

on mantissaosan absoluuttinen pyöristysvirhe.

Olkoon

$$u = b^{-t} \tag{3.2}$$

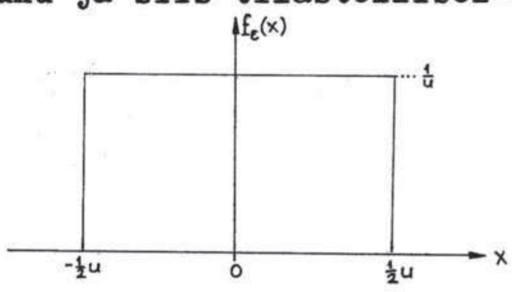
eli

$$u = 0.00\ldots01 \cdot b^{0} \tag{3.3}$$

esitettynä muodossa (2.7). Mikäli pyöristys on suoritettu 'oikein', on voimassa

$$-\tfrac{1}{2}u \le \varepsilon \le \tfrac{1}{2}u \; . \tag{3.4}$$

Intuitiivisesti pidetään selvänä, että pyöristysvirhe on satunnainen luku ja siis tilastolliselta jakautumaltaan tasaisesti jakautunut välillä $[-\frac{1}{2}u, \frac{1}{2}u]$. Satunnaisuus on tosin näennäistä: jos otamme uudelleen alkuperäisen luvun ja pyöristämme sen, saamme uudel-



kuva 1

leen saman pyöristysvirheen [4]. Tasainen jakautuma antaa kuitenkin hyvän tilastollisen mallin, jonka avulla voidaan tarkastella pyöristysvirheiden käyttäytymistä pitkissä laskutoimitussarjoissa.

Tilastollisen jakautuman tiheysfunktiolla f(x) tarkoitetaan Δx:llä jaettua todennäköisyyttä sille, että muuttujan arvo on x:n ja $x+\Delta x$:n välillä, kun Δx on äärettömän pieni [2]. Pyöristysvirheen ε jakautuman tiheysfunktio on muotoa

$$f_\varepsilon(x) = \begin{cases} 0 , & x < -\frac{1}{2}u \\ c , & -\frac{1}{2}u \le x \le \frac{1}{2}u \\ 0 , & x > \frac{1}{2}u , \end{cases} \qquad (3.5)$$

missä c on eräs vakio.

Kokonaistodennäköisyys on yksi, joten

$$\int_{-\infty}^{\infty} f_\varepsilon(x)dx = cu = 1 , \qquad (3.6)$$

mistä

$$c = \frac{1}{u} . \qquad (3.7)$$

Näin saatu virheen tiheysfunktio on havainnollistettu kuvassa 1.

Virheen itseisarvolle saamme tiheysfunktioksi

$$f_{|\varepsilon|}(x) = \begin{cases} 0 , & x < 0 \\ \frac{2}{u} , & 0 \le x \le \frac{1}{2}u \\ 0 , & x > \frac{1}{2}u . \end{cases} \qquad (3.8)$$

Useissa tietokoneissa käytetään oikean pyöristyksen sijasta katkaisevaa aritmetiikkaa [6], jolloin pyöristysvirhe

$$\varepsilon' = -|m^*-m| = (m^*-m)\cdot \text{sign } m \qquad (3.9)$$

on tasaisesti jakautunut välillä [-u,0]. Funktio sign x määritellään

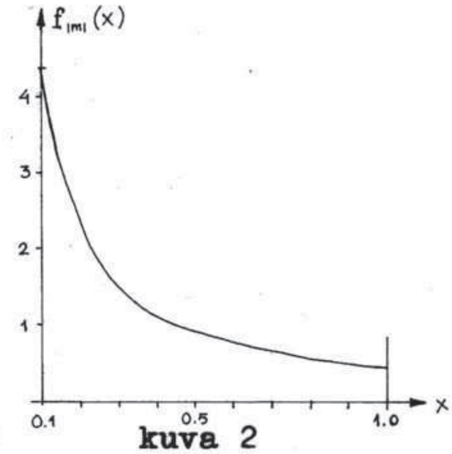
$$\operatorname{sign} x = \begin{cases} -1\ , & x < 0 \\ 0\ , & x = 0 \\ 1\ , & x > 0\ . \end{cases} \tag{3.10}$$

Pyöristysvirheen ε' tiheysfunktioksi saamme

$$f_{\varepsilon'}(x) = \begin{cases} 0\ , & x < -u \\ \frac{1}{u}\ , & -u \le x \le 0 \\ 0\ , & x > 0\ . \end{cases} \tag{3.11}$$

Tarvitsemme vielä itse mantissan jakautuman pystyäksemme laskemaan suhteellisen pyöristysvirheen jakautuman. Tuntuisi luonnolliselta olettaa, että $|m|$ olisi tasaisesti jakautunut välillä $[b^{-1},1)$. On kuitenkin todettu, ettei tämä pidä paikkaansa [3]. Jos tarkastelemme suurta joukkoa reaalimaailman kymmenjärjestelmän lukuja, esimerkiksi fysikaalisia vakioita, voimme todeta, että niistä on noin 30% ykkösellä alkavia.

Jonkinlaisen teoreettisen selvityksen tälle tilanteelle antaa samoin ilmeisenä pitämämme tosiasia, että kun kaikki reaalimaailman luvut kerrotaan tietyllä vakiolla, niiden ensimmäisten numeroiden jakautuma ei muutu. Tähän nojaamalla voidaan m :n tiheysfunktiolle johtaa kuvassa 2 havainnollistettu kaava [8]



kuva 2

$$f_{|m|}(x) = \begin{cases} 0\ , & x < b^{-1} \\ \dfrac{1}{x \cdot \ln b}\ , & b^{-1} \le x < 1 \\ 0\ , & x \ge 1\ . \end{cases} \tag{3.12}$$

Tämä tiheysfunktio antaa ykkösellä alkavien lukujen esiintymistodennäköisyydeksi $\int_{0.1}^{0.2} f_{|m|}(x)dx = \log_{10} 2 = 0.30$, mikä vastaa kokeellista tulosta.

Suhteellisen pyöristysvirheen jakautuma pyöristävässä aritmetiikassa

Kun kirjoitamme suhteellisen pyöristysvirheen e lausekkeen (2.9) mantissa- ja eksponenttiosan avulla, saamme kaavojen (2.4) ja (3.1) perusteella

$$e = \frac{z^*-z}{z} = \frac{(m^*-m)\cdot b^p}{m\cdot b^p} = \frac{\varepsilon\cdot b^p}{m\cdot b^p} = \frac{\varepsilon}{m}\,. \tag{3.13}$$

Johdamme nyt e:n itseisarvon $|e| = |\varepsilon/m|$ tiheysfunktion.

Kahden toisistaan riippumattoman satunnaismuuttujan η ja ξ suhteen $\zeta = \eta/\xi$ tiheysfunktio saadaan kaavasta [2]

$$f_\zeta(z) = \int_0^\infty x f_\eta(zx) f_\xi(x)dx\,, \tag{3.14}$$

kun $\xi > 0$. Mikäli oletamme, että ε ja m ovat riippumattomia, mikä tuntuu luonnolliselta ainakin suurilla t:n arvoilla, saadaan $|e|$:n tiheysfunktioksi kaavan (3.14) perusteella

$$f_{|e|}(z) = \begin{cases} 0\,, & z < 0 \\ \int_{b^{-1}}^{1} x\cdot\frac{2}{u}\cdot\frac{1}{x\cdot\ln b}dx = \frac{2\cdot(b-1)}{ub\cdot\ln b}\,, & 0 \le z \le \frac{1}{2}u \\ \int_{b^{-1}}^{u/2z} x\cdot\frac{2}{u}\cdot\frac{1}{x\cdot\ln b}dx = \frac{1}{\ln b}\left[\frac{1}{z}-\frac{2}{ub}\right], & \\ & \frac{1}{2}u < z \le \frac{1}{2}ub \\ 0\,, & z > \frac{1}{2}ub\,. \end{cases} \tag{3.15}$$

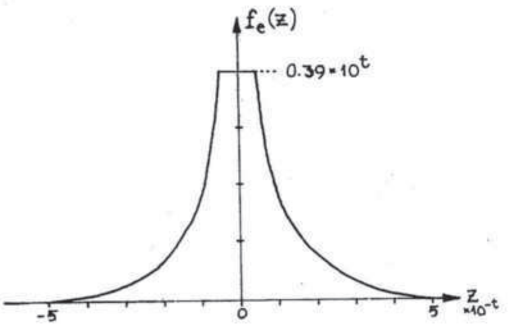
Saamme tästä $|e|$:n ylärajaksi [10]

$$|e| \le \tfrac{1}{2}ub\,. \tag{3.16}$$

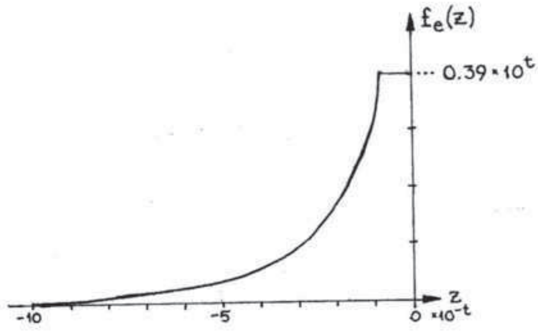
Kun vielä oletamme e:n jakautuman olevan symmetrinen origon suhteen, mikä myös tuntuu ilmeiseltä, saamme e:n tiheysfunktioksi

$$
f_e(z) = \begin{cases} \dfrac{b-1}{ub \cdot \ln b}, & |z| \le \frac{1}{2}u \\[2mm] \dfrac{1}{\ln b}\left[\dfrac{1}{2|z|} - \dfrac{1}{ub}\right], & \frac{1}{2}u < |z| \le \frac{1}{2}ub \\[2mm] 0, & |z| > \frac{1}{2}ub . \end{cases} \qquad (3.17)
$$

Tämä on havainnollistettu kuvassa 3a kymmenjärjestelmän tapauksessa.



kuva 3a



kuva 3b

Satunnaismuuttujan keski- eli odotusarvo saadaan kaavasta [2]

$$E(\xi) = \int_{-\infty}^{\infty} x f_\xi(x)\,dx \qquad (3.18)$$

ja varianssi kaavasta

$$D^2(\xi) = \int_{-\infty}^{\infty} [x - E(\xi)]^2 f_\xi(x)\,dx. \qquad (3.19)$$

Varianssin neliöjuuri, keskihajonta $D(\xi)$ kuvaa muuttujan arvojen keskimääräistä poikkeamaa keskiarvosta.

Saamme kaavojen (3.18) ja (3.19) avulla e:n odotusarvoksi ja varianssiksi

$$\mu_A = E(e) = 0, \quad \sigma_A^2 = D^2(e) = \frac{u^2}{24 \cdot \ln b}(b^2 - 1). \qquad (3.20)$$

Seuraavassa taulukossa on varianssin arvoja kantaluvun b eri arvoilla.

b	σ_A^2/u^2
2	0.1803
4	
8	1.262
10	1.791
16	3.832
32	
64	41.02

Suhteellisen pyöristysvirheen jakautuma katkaisevassa aritmetiikassa

Katkaisevassa aritmetiikassa saamme yhtälöstä (2.9) kaavojen (2.4) ja (3.11) perusteella suhteelliselle pyöristysvirheelle lausekkeen

$$e = \frac{z^*-z}{z} = \frac{(m^*-m)\cdot b^p}{m\cdot b^p} = \frac{m^*-m}{|m|\cdot \mathrm{sign}\ m} = \frac{\varepsilon'}{|m|}. \quad (3.21)$$

Kun oletamme ε':n ja $|m|$:n olevan riippumattomia, saamme kaavan (3.14) perusteella e:n tiheysfunktioksi (kuva 3b)

$$f_e(z) = \begin{cases} 0\ , & z < -ub \\ \int_{b^{-1}}^{-u/z} x\cdot\frac{1}{u}\cdot\frac{1}{x\cdot\ln b}dx = \frac{1}{\ln b}\left[-\frac{1}{z}-\frac{1}{ub}\right], & -ub \le z < -u \\ \int_{b^{-1}}^{1} x\cdot\frac{1}{u}\cdot\frac{1}{x\cdot\ln b}dx = \frac{b-1}{ub\cdot\ln b}, & -u \le z \le 0 \\ 0\ , & z > 0 \end{cases} \quad (3.22)$$

Kaavojen (3.18) ja (3.19) perusteella saamme odotusarvoksi ja varianssiksi

$$\mu_A = E(e) = \frac{u(1-b)}{2\cdot\ln b}\ ,$$
$$\sigma_A^2 = D^2(e) = -E(e)\left[\frac{u(b+1)}{3} + E(e)\right]. \quad (3.23)$$

Seuraavassa taulukossa on odotusarvoja ja variansseja kantaluvun eri arvoilla.

b	μ_A/u	σ_A^2/u^2
2	-0.7213	0.2010
4	-1.082	
8	-1.683	2.216
10	-1.954	3.346
16	-2.705	8.011
32	-4.472	
64	-7.574	106.7

4. PYÖRISTYSVIRHE LASKUTOIMITUSTEN YHTEYDESSÄ

Yhteen- ja vähennyslasku

Kahden liukuluvun yhteenlaskun tarkkuuteen vaikuttaa ratkaisevasti yhteenlaskettavien lukujen eksponenttien p_1 ja p_2 keskinäinen ero $|p_1-p_2|$ [10].

Mikäli $p_1 = p_2$, on yhteenlasku varsin suurella todennäköisyydellä tarkka. Ainoan poikkeuksen muodostaa tapaus, jossa saadaan muistinumero mantissojen vasemmanpuoleisimpia numeroita yhteenlaskettaessa. Tällöin joudutaan vähiten merkitsevästä numerosta luopumaan, koska mantissa edelleen saa sisältää korkeintaan t numeroa. Samaan tilanteeseen saatetaan joutua myös, kun p_1 ja p_2 ovat likimain yhtäsuuret. Tämä on kuitenkin sitä epätodennäköisempää mitä suurempi $|p_1-p_2|$ on.

Kun $|p_1-p_2| = 1$, kohdistuu mielenkiinto lähinnä tapauksiin, joissa luvut ovat erimerkkiset ja niiden itseisarvot niin lähellä toisiaan, että tulosta joudutaan 'siirtämään vasemmalle', jolloin päästään tarkkaan tulokseen.

D.W. Sweeneyn suorittamassa tutkimuksessa [8] $|p_1-p_2|$ oli käytännön yhteenlaskuissa nolla 33-56%:n ja yksi 12-27%:n todennäköisyydellä kantaluvun saadessa arvot 2-64. Oikealle siirtojen todennäköisyys oli vastaavasti 20-2% ja vasemmalle siirtojen 20-11%.

Eksponenttien eron kasvaessa vähenee tarkan laskutoimituksen todennäköisyys ja edellä johdetut mallit pyöristävälle ja katkaisevalle aritmetiikalle pitävät varsin hyvin paikkansa. Tämä paikkansapitävyys loppuu $|p_1-p_2|$:n ylittäessä arvon t. Tällöin luvuista itseisarvoltaan pienempi muodostaa sellaisenaan pyöristysvirheen. Näiden tapausten suhteellinen osuus kaikista yhteenlaskuista on kuitenkin niin pieni, ettei niillä ole suhteellisen pyöristysvirheen jakautumaan merkittävää vaikutusta, kunhan huomioimme tapaukset, joissa toinen yhteenlaskettava on nolla, jolloin laskutoimitus on tarkka. Toinen operandi on edellä mainitun tutkimuksen mukaan tarkka noin 8% todennäköisyydellä.

Koska suhteellisen pyöristysvirheen jakautuma riippuu $|p_1 - p_2|$:n jakautumasta, sille on mahdotonta johtaa täysin yleispätevää mallia. Jonkinlaisen keskivertomallin muodostaminen tosin on mahdollista. Tällöin tullee kysymykseen lähinnä professori Tienarin ehdottama malli: kun

$$(z_1 + z_2)^* = (z_1 + z_2) \cdot (1+e) \;, \qquad (4.1)$$

missä e on suhteellinen pyöristysvirhe, kirjoitetaan e muotoon

$$e = \delta \cdot e' \;, \qquad (4.2)$$

missä e' noudattaa edellä johdettua jakautumaa ja satunnaismuuttuja δ saa arvon 0 todennäköisyydellä p sekä arvon 1 todennäköisyydellä 1-p. Luku p edustaa tällöin todennäköisyyttä tarkalle laskutoimitukselle. Merkitsemme seuraavassa näin saatavan e:n jakautuman odotusarvoa μ_s:llä ja varianssia σ_s^2:llä.

Virheen e yläraja voidaan määrätä yleispätevästi. Perustuen epäyhtälöihin (2.6) ja (3.4) saadaan yhtälöstä (3.13) pyöristävälle aritmetiikalle [7]

$$|e| = \left|\frac{\varepsilon}{m}\right| \le \tfrac{1}{2}ub. \qquad (4.3)$$

Katkaisevalle aritmetiikalle saadaan vastaavasti

$$|e| \le ub. \qquad (4.4)$$

Eräissä koneissa joudutaan jo ennen laskutoimitusta pyöristämään eksponentiltaan pienempää operandia. Tällöin saadaan ylärajoiksi

$$|e| \leq \tfrac{1}{2}u(b+1) \qquad (4.5)$$

pyöristävässä ja

$$|e| \leq u(b+1) \qquad (4.6)$$

katkaisevassa aritmetiikassa [10].

Kaikki yhteenlaskua koskeva koskee myös vähennyslaskua, koska näiden ero voidaan tulkita lukujen etumerkkien vaihteluksi.

Kerto- ja jakolasku

Kahden liukuluvun välinen kertolasku suoritetaan laskemalla eksponentit yhteen ja kertomalla mantissaosat keskenään. Kun

$$(z_1 \cdot z_2)^* = z_1 \cdot z_2 (1+e) , \qquad (4.7)$$

voidaan todeta pyöristysvirheen e noudattavan johdettua jakautumaa, so. pyöristävässä aritmetiikassa jakautumaa (3.17) ja katkaisevassa jakautumaa (3.22).

Jakolaskussa vastaavasti vähennetään eksponentit toisistaan ja mantissaosat jaetaan keskenään. Myös osamäärän pyöristysvirhe e noudattaa annettua jakautumaa, kun merkitsemme

$$\left(\frac{z_1}{z_2}\right)^* = \frac{z_1}{z_2}(1+e) . \qquad (4.8)$$

Merkitsemme seuraavassa kerto- ja jakolaskun pyöristysvirheen odotusarvoa μ_T:llä ja varianssia σ_T^2:llä. Tällöin

$$\mu_T = \mu_A , \quad \sigma_T^2 = \sigma_A^2 . \qquad (4.9)$$

Johdetun mallin pitävyys kerto- ja jakolaskussa testattiin algoritmin S avulla.

Algoritmi S (Satunnaisalgoritmi). Algoritmi suorittaa satunnaisia laskutoimituksia 2t numeron tarkkuudella katkaisten tulokset t numeron mittaisiksi ja luetteloiden syntyneet pyöristysvirheet. Jos tuloksen eksponentti ylittää itseisarvoltaan puolet suurimmasta mahdollisesta eksponentista max|p|, sen itseisarvoa pienennetään $\frac{1}{2}$·max|p|:llä ylivuotojen ehkäisemiseksi. Algoritmi käyttää vektoria LUKU(1),LUKU(2),...,LUKU(100). Merkintä A←SAT{B, C,...,Z} tarkoittaa, että A:n arvoksi sijoitetaan joukon {B,C,...,Z} satunnaisesti valitun alkion arvo.

S1. [Alkuasetukset.] Vektorin LUKU kymmenen ensimmäisen alkion arvoksi sijoitetaan tunnettuja vakioita: π, Neperin luku, Eulerin vakio, kultainen suhde jne. N←10 (N on käytössä olevan taulukon osan suurin indeksi).

S2. [Laskutoimituksen valinta.] I←SAT{1,2,...,N}, J←SAT{1,2,...,N}, SX←SAT{S3,S4,S5,S6}, →SX.

S3. [Yhteenlasku.] TULOS←LUKU(I)+LUKU(J), →S7.

S4. [Vähennyslasku.] TULOS←LUKU(I)-LUKU(J), →S7.

S5. [Kertolasku.] TULOS←LUKU(I)·LUKU(J), →S7.

S6. [Jakolasku.] Jos LUKU(J) = 0, →S2. Muuten TULOS←LUKU(I)/LUKU(J).

S7. [Etumerkin valinta.] SIGN←SAT{-1,1}, TULOS←SIGN·TULOS.

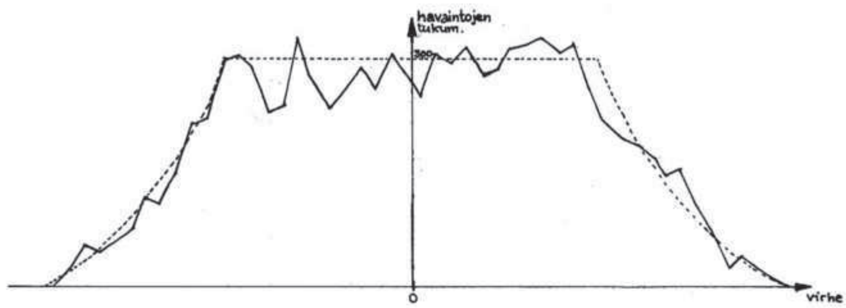
S8. [Pyöristysvirheen kirjaus.] Pyöristä mantissa t numeron mittaiseksi, luetteloi pyöristysvirhe ja pienennä tarvittaessa eksponenttia.

S9. [Tulosalkion valinta.] Jos N < 100, N←N+1,K←N, muuten K←SAT{1,2,..,N}.

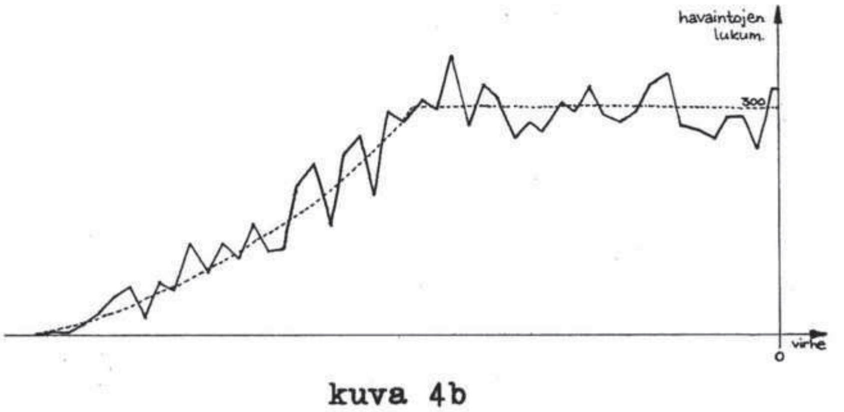
S10. [Tuloksen talletus.] LUKU(K)←TULOS. Jos testituloksia halutaan lisää, →S2, muuten algoritmi päättyy. ▌

Testitulosten lukumääräksi määrättiin 10000 sekä pyöristävälle että katkaisevalle aritmetiikalle, jolloin saatiin riittävä kuva pyöristysvirheiden jakautumasta. Summan ja erotuksen pyöristysvirheiden jakautumalle algoritmi S ei antanut mielekästä mallia, koska eksponenttien jakautumaan ei kiinnitetty erityistä huomiota. Sen sijaan tulon ja osamäärän pyöristysvirheiden jakautumalle saatiin malli, joka vastaa teoreettista mallia. Tulokset on esitetty pyöristävälle aritmetiikalle kuvassa 4a ja katkaisevalle kuvassa 4b.

Testikoneena oli IBM 7094, jonka aritmetiikan kantalukuna on kaksi ja t = 27 (simuloitaessa pyöristävää aritmetiikkaa jouduttiin asettamaan t = 26). Mallin hyvä pitävyys osoittaa omalta osaltaan myös jakautuman (3.12) pätevyyttä kaksijärjestelmänkin yhteydessä. Tätä jakautumaahan käytettiin pyöristysvirheen teoreettista jakautumaa johdettaessa.



kuva 4a



kuva 4b

5. PYÖRISTYSVIRHEEN TAYLOR-KEHITELMÄ

Suoritettaessa peräkkäisiä laskutoimituksia liittyy jokaiseen laskutoimitukseen oma pyöristysvirheensä. Näiden yksittäisten pyöristysvirheiden yhteisvaikutusta kutsutaan kumulatiiviseksi pyöristysvirheeksi. Jos tarkastelemme esimerkiksi kahden pyöristetyn luvun a^* ja b^* jakolaskua, saamme

$$
\begin{aligned}
\left(\frac{a^*}{b^*}\right)^* &= \frac{a(1+e_a)}{b(1+e_b)}(1+e_c) \\
&= \frac{a}{b}(1+e_a)(1+e_c)(1-e_b+e_b^2-e_b^3+\ldots) \\
&= \frac{a}{b}(1+e_a-e_b+e_c-e_ae_b+e_ae_c-e_be_c \\
&\qquad -e_ae_be_c+e_ae_b^2+e_ce_b^2-e_b^3+\ldots) \\
&= \frac{a}{b}(1+E_c),
\end{aligned}
\tag{5.1}
$$

jolloin E_c edustaa kumulatiivista suhteellista pyöristysvirhettä. Kuten tässä esimerkissä voidaan kumulatiivinen pyöristysvirhe E_n aina lausua yksittäisten pyöristysvirheiden Taylor-kehitelmänä, so. muodossa

$$E_n = \sum_i a_{n.i}\, e_i + \sum_{i,j} a_{n.ij}\, e_i\, e_j + \langle e^3\rangle , \tag{5.2}$$

missä $\langle e^3\rangle$ tarkoittaa e_i:den suhteen vähintään kolmannen asteen termejä ja a:t algoritmin alkuarvoista riipuvia vakioita. Kutsumme sarjaa (5.2) seuraavassa (E,e)-sarjaksi.

Kun merkitsemme (E,e)-sarjan k:nnetta osasummaa, so. e_i:den suhteen k:nnen asteen summaa $E_n^{(k)}$:lla, saamme yhtälön (5.2) muotoon

$$E_n = E_n^{(1)} + E_n^{(2)} + \langle e^3\rangle . \tag{5.3}$$

Toisen ja korkeamman asteen termien vaikutus kumulatiiviseen pyöristysvirheeseen on yleensä merkityksetön yksittäisten pyöristysvirheiden e_i suuruusluokan pienuuden johdosta, joten käytännössä voidaan E_n:n lauseke ilmoittaa ensimmäisen asteen termiensä avulla [9], so.

$$E_n \approx E_n^{(1)} . \tag{5.4}$$

Kumulatiivisen pyöristysvirheen jakautuma voidaan johtaa yksittäisten pyöristysvirheiden jakautumien avulla. Jos satunnaisluvut $\xi_1, \xi_2, \ldots, \xi_m$ ovat toisistaan riippumattomia ja niiden odotusarvot ovat $\mu_1, \mu_2, \ldots, \mu_m$ ja varianssit $\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2$, on satunnaismuuttujan

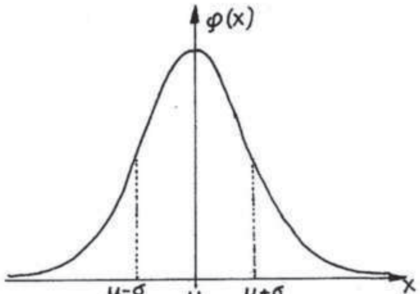
$$\xi = a_1\xi_1 + a_2\xi_2 + \ldots + a_m\xi_m \tag{5.5}$$

odotusarvolle ja varianssille voimassa [4]

$$E(\xi) = a_1\mu_1 + a_2\mu_2 + \ldots + a_m\mu_m, \tag{5.6}$$
$$D^2(\xi) \approx a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_m^2\sigma_m^2. \tag{5.7}$$

Todennäköisyyslaskennan keskeisen raja-arvoväittämän nojalla ξ:n jakautuma lisäksi lähestyy ns. normaalijakautumaa (kuva 5) m:n lähestyessä ääretöntä [2].

Normaalijakautuman määrittävät yksikäsitteisesti sen odotusarvo ja varianssi. Sille on ominaista mm., että 68.3% havaituista satunnaisluvuista poikkeaa vähemmän kuin varianssin neliöjuuren eli keskihajonnan verran odotusarvosta.



kuva 5

Pyöristysvirheet eivät yleensä ole täysin riippumattomia toisistaan, mutta niiden väliset riippuvuudet ovat niin vähäisiä, että kaavoja (5.6) ja (5.7) käyttämällä saadaan riittäviä arvioita kumulatiivisten pyöristysvirheiden jakautumille yksittäisten pyöristysvirheiden jakautumien avulla [4]. Oikein pyöristävässä aritmetiikassa on kaavan (5.6) perusteella myös kumulatiivisen pyöristysvirheen odotusarvo nolla.

Pitkissä laskusarjoissa voidaan nojata normaalijakautumaan myös määritettäessä käytännöllisiä ylärajoja pyöristysvirheille. Normaalijakautumassa sijoittuu 99% kaikista havainnoista lähemmäksi odotusarvoa kuin 2.576σ, missä σ on normaalijakautuman keskihajonta. Selkeänä esimerkkinä voidaan tarkastella peräkkäisten kertolaskujen kumulatiivista pyöristysvirhettä oikein pyöristävässä aritmetiikassa.

Tuloa

$$P = \prod_{i=0}^{N} h_i$$

vastaa liukuvan pilkun aritmetiikkaa käyttävässä koneessa, kun alkuarvot h_i, $i = 0,\ldots,N$ oletetaan tarkoiksi,

$$P^* = P(1+E_N) = \prod_{i=0}^{N} h_i \prod_{i=1}^{N}(1+e_i) \approx \prod_{i=0}^{N} h_i (1+\sum_{i=1}^{N} e_i) ,$$

missä kukin e_i on yksittäisessä kertolaskussa syntynyt pyöristysvirhe.

Jos kertolaskut on suoritettu binääriaritmetiikassa, on virheiden e_i, $i = 1,\ldots,N$ odotusarvo 0 ja varianssi $0.1803u^2$ kaavan (3.20) perusteella. Saamme E_N:n odotusarvoksi ja varianssiksi kaavojen (5.6) ja (5.7) avulla

$$E(E_N) = 0 , \quad D^2(E_N) = N \cdot 0.1803u^2 ,$$

jolloin siis 99% varmuudella

$$|E_N| < 2.576 \cdot \sqrt{N \cdot 0.1803u^2} = 1.09\sqrt{N}u .$$

Tri Wilkinson [10] on päätynyt käytännön ylärajana tulokseen $\sqrt{N}u$, mikä vastaa varsin tarkkaan edellä johdettua tulosta. Teoreettiseksi ylärajaksi saadaan

$$|E_N| < \tfrac{1}{2}Nub\ ,$$

joten varsinkin suurilla N:n arvoilla saataisiin varsin epärealistinen ylärajan arvo ilman tilastollista jakautumaa tai empiirisiä kokeita.

Käsitellyssä esimerkissä syntyivät kaikki pyöristysvirheet kertolaskujen yhteydessä. Pyöristysvirheet voidaan kuitenkin jakautumansa ja syntytapansa perusteella jakaa esimerkiksi tyyppeihin

1. alkuarvojen pyöristysvirheet (keskiarvo μ_A, varianssi σ_A^2)
2. summan ja erotuksen pyöristysvirheet (keskiarvo μ_S, varianssi σ_S^2)
3. tulon ja osamäärän pyöristysvirheet (keskiarvo μ_T, varianssi σ_T^2).

Olkoot yhtälössä (5.2) $e_{i_1}, e_{i_2}, \ldots, e_{i_k}$ tyyppiä 1, $e_{i_{k+1}}, \ldots, e_{i_l}$ tyyppiä 2 ja $e_{i_{l+1}}, \ldots, e_{i_m}$ tyyppiä 3 olevia pyöristysvirheitä. Tällöin E_n:n odotusarvo ja varianssi voidaan kaavojen (5.4), (5.6) ja (5.7) mukaan lausua muodossa

$$E(E_n) \approx E(E_n^{(1)}) = \sum_{j=1}^{k} a_{n.ij}\mu_A + \sum_{j=k+1}^{l} a_{n.ij}\mu_S + \sum_{j=l+1}^{m} a_{n.ij}\mu_T\ , \qquad (5.8)$$

$$D^2(E_n) \approx D^2(E_n^{(1)}) = \sum_{j=1}^{k} a_{n.ij}^2\sigma_A^2 + \sum_{j=k+1}^{l} a_{n.ij}^2\sigma_S^2 + \sum_{j=l+1}^{m} a_{n.ij}^2\sigma_T^2\ . \qquad (5.9)$$

Esimerkiksi jakolaskun (5.1) pyöristysvirheelle E_c

$$E(E_c) \approx \mu_T\ ,\ D^2(E_c) \approx 2\sigma_A^2 + \sigma_T^2\ .$$

Mikäli alkuarvot oletetaan tarkoiksi, on yhtälössä (5.8) $\mu_A = 0$ ja yhtälössä (5.9) $\sigma_A^2 = 0$. Pyöristävässä aritmetiikassa $\mu_A = \mu_S = \mu_T = 0$.

6. TAYLOR-SARJAN ANALYYTTINEN KEHITTÄMINEN

Algoritmin laskutoimitusten lukumäärän kasvaessa on yhä vaikeampaa saada selville Taylor-kehitelmän kertoimia. Algoritmin rakennetta analysoimalla on kuitenkin mahdollista laskea analyyttisesti yleisiä lausekkeita kertoimille.

Kaikki algoritmit muodostuvat alkeislaskutoimituksista, joissa jokin operaatio kohdistetaan korkeintaan kahteen operandiin. Olkoon q_n erään laskutoimituksen tulos. Tällöin

$$q_n = Q_n(q_i, q_j) , \tag{6.1}$$

missä Q_n tarkoittaa yleensä yhteen-, vähennys-, kerto- tai jakolaskua. Se saattaa tarkoittaa myös potenssiin korotusta, logaritmin ottoa tms.

Tietokoneessa vastaa yhtälöä (6.1) pyöristysten vaikutukseata yhtälö

$$q_n^* = Q_n(q_i^*, q_j^*) \cdot (1+e_n) , \tag{6.2}$$

missä e_n on laskutoimituksen Q_n yksittäinen suhteellinen pyöristysvirhe.

Merkitsemme absoluuttista kumulatiivista pyöristysvirhettä

$$R_n = q_n^* - q_n , \tag{6.3}$$

jolloin suhteellinen kumulatiivinen pyöristysvirhe voidaan esittää muodossa

$$E_n = \frac{R_n}{q_n} . \tag{6.4}$$

Jos q_n on erityisesti alkuarvo, jolloin $q_n^* = q_n(1+e_n)$, on

$$R_n = q_n e_n \tag{6.5}$$

ja

$$E_n = e_n \quad . \tag{6.6}$$

Myös R_n voidaan lausua yksittäisten suhteellisten pyöristysvirheiden Taylor-kehitelmänä

$$R_n = \sum_i c_{n.i} e_i + \sum c_{n.ij} e_i e_j + \langle e^3 \rangle \quad . \tag{6.7}$$

Nimitämme sarjaa (6.7) (R,e)-sarjaksi. Taylor-kehitelmän yksikäsitteisyyden ja yhtälön (6.4) perusteella saamme E_n:n ja R_n:n yhtälöissä (5.2) ja (6.7) esiintyville kertoimille kaikilla kysymykseen tulevilla i:n ja j:n arvoilla

$$a_{n.i} = \frac{c_{n.i}}{q_n} \tag{6.8}$$

$$a_{n.ij} = \frac{c_{n.ij}}{q_n} \quad . \tag{6.9}$$

R_n voidaan E_n:n tapaan esittää eriasteisten osasummiensa avulla muodossa

$$R_n = R_n^{(1)} + R_n^{(2)} + \langle e^3 \rangle \quad . \tag{6.10}$$

Lausekkeen $Q_n(q_i^*, q_j^*)$ saamme väliarvolauseen ja määritelmän (6.3) nojalla muotoon [4]

$$Q_n(q_i^*, q_j^*) = q_n + D_{n.i} R_i + D_{n.j} R_j + \tfrac{1}{2} D_{n.ii} R_i R_i + D_{n.ij} R_i R_j + \tfrac{1}{2} D_{n.jj} R_j R_j + \langle R^3 \rangle \quad , \tag{6.11}$$

missä $D_{n.i}$ on $Q_n(q_i, q_j)$:n osittaisderivaatta q_i:n suhteen pisteessä q_n ja $D_{n.ij}$ vastaavasti toinen osittaisderivaatta samassa pisteessä. Yhtälöiden (6.2), (6.3),(6.10) ja (6.11) avulla saamme

$$R_n^{(1)} = D_{n.i}R_i^{(1)} + D_{n.j}R_j^{(1)} + q_n e_n \tag{6.12}$$

ja

$$\begin{aligned} R_n^{(2)} = {} & D_{n.i}R_i^{(2)} + D_{n.j}R_j^{(2)} + \tfrac{1}{2}D_{n.ii}R_i^{(1)}R_i^{(1)} + D_{n.ij}R_i^{(1)}R_j^{(1)} \\ & + \tfrac{1}{2}D_{n.jj}R_j^{(1)}R_j^{(1)} + D_{n.i}R_i^{(1)}e_n + D_{n.j}R_j^{(1)}e_n \ . \end{aligned} \tag{6.13}$$

Tarkastelemme saatuja tuloksia kahden pienen esimerkin valossa. Varsinaisissa algoritmeissa päädymme differenssiyhtälöihin, esimerkkinä tällaisista sovellamme teoriaa myöhemmin Horner-shemaan.

<u>Esimerkki 1.</u> Suoritetaan laskutoimitus $c = a/b$, missä a on tarkka, so. sitä ei tarvitse koneessa pyöristää, ja b:tä vastaa koneessa $b^* = b(1+e_b)$. Voimme siis yhtälön (6.5) perusteella merkitä $R_a^{(1)} = R_a^{(2)} = 0$, $R_b^{(1)} = be_b$, $R_b^{(2)} = 0$. Kaavojen (6.12) ja (6.13) perusteella

$$R_c^{(1)} = \frac{1}{b}R_a^{(1)} - \frac{a}{b^2}R_b^{(1)} + ce_c = -ce_b + ce_c$$

ja

$$\begin{aligned} R_c^{(2)} &= \frac{1}{b}R_a^{(2)} - \frac{a}{b^2}R_b^{(2)} + 0 - \frac{1}{b^2}R_a^{(1)}R_b^{(1)} + \frac{a}{b^3}R_b^{(1)}R_b^{(1)} + \frac{1}{b}R_a^{(1)}e_c - \frac{a}{b^2}R_b^{(1)}e_c \\ &= ce_be_b - ce_be_c \ . \end{aligned}$$

Kaavojen (6.8), (6.9) ja (6.10) perusteella c:n (E,e)-sarja on

$$E_c = -e_b + e_c + e_b^2 - e_b e_c + \ldots \ .$$

Samaan tulokseen päädyimme aikaisemmin yhtälöissä (5.1), kun otamme huomioon, että $e_a = 0$.

<u>Esimerkki 2.</u> Olkoon $b^* = (\ln a^*)^*$ kun $a^* = a(1+e_a)$. Edellä esitetyt kaavat johdettiin kahdelle operandille, mutta niitä voidaan soveltaa yhden operandin operaatioihin, kunhan osittaisderivaatat 'toisen' operandin suhteen ajatellaan nolliksi, mikä on luonnollista, koska tätä ei esiinny laskutoimi-

tuksessa. Myös tätä toista operandia vastaava virhetermi ajatellaan nollaksi. Tällöin saamme

$$R_b^{(1)} = \frac{1}{a}ae_a + \ln a \cdot e_b = e_a + be_b$$

ja

$$R_b^{(2)} = \frac{1}{a}0 - \frac{1}{2}\cdot\frac{1}{a^2}a^2 e_a^2 + \frac{1}{a}e_a e_b = -\frac{1}{2}e_a^2 + e_a e_b \ ,$$

missä e_b on logaritmin otossa tapahtuva suhteellinen virhe. Jälleen pääsemme samaan tulokseen yhtälöstä $R_b = [\ln(a+ae_a)](1+e_b)-b$, kun käytämme hyväksi kaavaa $\ln(1+e_a) = e_a - e_a^2/2 + \langle e_a^3\rangle$.

Edellä esitetyn analyysin perusteella voimme myös verrata suhteellisten ja absoluuttisten yksittäisten pyöristysvirheiden e_n ja r_n yhteyttä kumulatiivisen pyöristysvirheen Taylor-kehitelmässä.

Operaation Q_n aiheuttama yksittäinen absoluuttinen pyöristysvirhe voidaan kirjoittaa yhtälöiden (6.2) ja (6.11) mukaan muotoon

$$\begin{aligned} r_n &= Q_n(q_i^*, q_j^*)e_n \\ &= q_n e_n + D_{n.i}R_i e_n + D_{n.j}R_j e_n + \ldots \ . \end{aligned} \tag{6.14}$$

Täten R_n:n esityksestä pyöristysvirheiden r_i avulla, ns. (R,r)-sarjasta

$$R_n = \sum_i d_{n.i}r_i + \sum_{i,j} d_{n.ij}r_i r_j + \langle r^3\rangle \tag{6.15}$$

saadaan

$$R_n = \sum_i d_{n.i}q_i e_i + \langle e^2\rangle \ . \tag{6.16}$$

Kun vertaamme yhtälöitä (6.7) ja (6.16), saamme Taylor-kehitelmän yksikäsitteisyyden nojalla i:n kaikilla arvoilla

$$c_{n.i} = q_i d_{n.i} \ . \tag{6.17}$$

Kaavan (6.8) perusteella

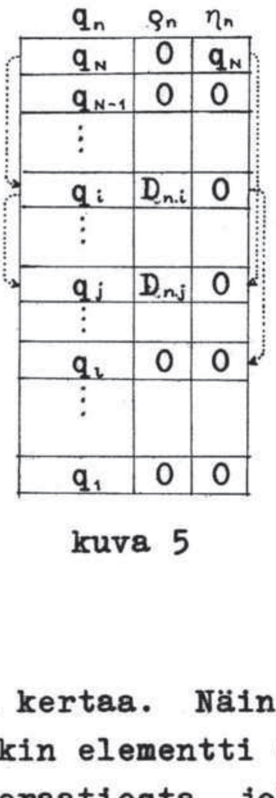
$$a_{n.i} = \frac{q_i}{q_N} d_{m.i} \, . \tag{6.18}$$

Näin olemme selvittäneet (E,e)-, (R,e)- ja (R,r)-sarjojen ensimmäisen asteen termien väliset yhteydet.

7. TAYLOR-SARJAN KERTOIMIEN MÄÄRITTÄMINEN TIETOKONEOHJELMALLA

Kertoimienlaskualgoritmi L

Laskualgoritmien laajetessa alkaa myös Taylor-sarjan analyyttinen kehittäminen tuottaa vaikeuksia, koska saadut differenssiyhtälöt monimutkaistuvat. Voimme kuitenkin soveltaa analyyttistä teoriaa tietokonealgoritmiin, jonka avulla Taylor-sarjan ensimmäisen asteen kertoimet voidaan laskea kullekin alkuarvojoukolle erikseen varsin vaikeissakin algoritmeissa.



kuva 5

Kertoimienlaskualgoritmi L perustuu yhtälöön (6.12). Algoritmin lähtökohtana on kaikkien alkuarvojen ja yksittäisten laskutoimitusten tulosten q_n vieminen pinoon niiden esiintyessä ensimmäistä kertaa. Näin syntyneen pinon (q_1, q_2,..., q_N) kukin elementti on siis joko alkuarvo tai tulos operaatiosta, joka on kohdistunut yhteen tai kahteen pinossa alempana olevaan elementtiin. Ajattelemme kuhunkin elementtiin q_n, $n = 1,...,N$, liittyvän R_n:n ja e_n:n kertoimet ϱ_n ja η_n siten, että ehto

$$R_N^{(1)} = \sum_{n=1}^{N} \varrho_n R_n + \sum_{n=1}^{N} \eta_n e_n \qquad (7.1)$$

on voimassa.

Aluksi toteutamme yhtälön (7.1) asettamalla $\varrho_N = 1$, $\varrho_n = 0$, $n \neq N$ ja $\eta_n = 0$, $n = 1,2,\ldots,N$. Olkoon $q_N = Q_N(q_i, q_j)$, missä $i > j$. Yhtälön (6.12) mukaan täytämme yhtälön (7.1) ehdon myös sijoittamalla nyt $\varrho_N \leftarrow 0$, $\varrho_i \leftarrow D_{n.i}$, $\varrho_j \leftarrow D_{n.j}$ ja $\eta_N \leftarrow q_N$. Tämä tilanne on esitetty kuvassa 5.

Seuraavaksi nollaamme pinossa ylimpänä olevan nollasta poikkeavan ϱ_n:n eli ϱ_i:n. Jos esimerkiksi $q_i = Q_i(q_j, q_l)$, on tämä nollaus kaavan (6.12) mukaan mahdollista asettamalla $\varrho_j \leftarrow \varrho_j + \varrho_i D_{i.j}$, $\varrho_l \leftarrow \varrho_i D_{i.l}$ ja $\eta_i \leftarrow \varrho_i q_i$. Näin jatketaan ϱ_n:ien nollaamista edeten pinossa alaspäin, kunnes koko pino on käyty läpi, so. $\varrho_n = 0$, $n = 1,2,\ldots,N$. Ehto (7.1) on koko ajan voimassa, joten lopputuloksena saadaan yhtälöiden (6.7) ja (7.1) sekä Taylor-kehitelmän yksikäsitteisyyden nojalla η-sarakkeeseen (R,e)-sarjan ensimmäisen asteen kertoimet.

Voimme todeta, että algoritmin edistyessä η_n pysyy nollana, kunnes ϱ_n otetaan nollattavaksi. Tällöin η_n:n arvoksi tulee $\varrho_n q_n$. Mikäli jätämme tässä vaiheessa kertomatta q_n:llä, saamme yhtälön (6.17) perusteella η-sarakkeeseen (R,r)-sarjan kertoimet. Voimme siis poistaa operaatiot $\varrho_n \leftarrow 0$, $\eta_n \leftarrow \varrho_n q_n$ ja samalla koko η-sarakkeen, jolloin algoritmin päättyessä ϱ-sarakkeessa on (R,r)-sarjan kertoimet. Algoritmissa L vastaa ϱ_n-kenttää kenttä COEFF(n).

Algoritmin L 1.osassa luotavan pinon I:s tietue on muotoa

VALUE(I)	TYPE(I)	OPER1(I)	OPER2(I)	COEFF(I)

.

Kentässä VALUE on algoritmin alkuarvon tai laskutoimituksen tuloksen arvo q_I, TYPE ilmoittaa operaation Q_I laadun viereisen taulukon mukaan, OPER1 ja OPER2-kentissä on linkit operandeja vastaaviin tietueisiin ja COEFF-kent-

TYPE	operaation laatu
1	alkuarvo
2	negatointi
3	yhteenlasku
4	vähennyslasku
5	kertolasku
6	jakolasku

tään lasketaan kertoimet kuten edellä esitettiin. TYPE- ja OPER-kenttien avulla pystytään laskemaan tarvittavat osittaisderivaattojen arvot.

Kun q_i liittyy laskualgoritmissa laskutoimitukseen, siihen viitataan algoritmissa L kentän QI avulla, joka sisältää linkin q_i:n viimeksi laskettua arvoa vastaavaan tietueeseen (ko. tietueen pinoindeksin). Seuraavassa kutsutaan QI:tä muuttujan q_i nimikkeeksi.

<u>Algoritmi L</u> (Kumulatiivisen pyöristysvirheen Taylor-sarjan kertoimien lasku). Algoritmi L jakautuu kahteen osaan. Mielivaltainen laskualgoritmi, 'algoritmi A', suoritetaan kokonaisuudessaan käyttäen osaa 1, joka luo tarvittavan tietuepinon. Pinolle on varattava tilaa vähintään niin monelle tietueelle kuin algoritmissa A on alkuarvoja ja laskutoimituksia yhteensä. Pinoindeksi I on pinon pohjalta lukien ensimmäisen vapaan tietueen järjestysnumero. Osassa 2 lasketaan tuloksen q_N (vastaava pinoindeksi N) kumulatiivisen pyöristysvirheen Taylor-kehitelmän ensimmäisen asteen kertoimet osassa 1 muodostetun tietuepinon avulla. D1 on Q_K:n osittaisderivaatta ensimmäisen ja D2 toisen operandin suhteen voimassaolevalla pinoindeksin K arvolla. Huom. q_N:n ei välttämättä tarvitse olla algoritmin A lopputulos.

<u>Algoritmi L, osa 1.</u>

L1. [Pinoindeksin alkuasetus.] I←1.

L2. [Algoritmin A seuraava alkeistoimitus.] Jos algoritmi A on päättynyt, algoritmin L osa 1 päättyy,→L12. Jos algoritmissa A otetaan käyttöön alkuarvo,→L3. Jos siinä suoritetaan negatointi,→L4, jos yhteenlasku,→L5, jos vähennyslasku,→L6, jos kertolasku,→L7 ja jos jakolasku, →L8. (Tämä algoritmi ei huomioi muita toimituksia, mutta algoritmia voidaan tarvittaessa laajentaa.)

L3. [Alkuarvo.] (Algoritmissa A otetaan käyttöön alkuarvo q_i.) TYPE(I)←1, VALUE(I)←q_i,→L11.

L4. [Negatointi.] (Algoritmissa A '$q_i = -q_j$'.) TYPE(I)←2, VALUE(I)←-VALUE(QJ), →L10.

L5. [Yhteenlasku.] (Algoritmissa A '$q_i = q_j+q_k$'.) TYPE(I)←3, VALUE(I)←VALUE(QJ)+VALUE(QK),→L9.

L6. [Vähennyslasku.] (Algoritmissa A '$q_i = q_j-q_k$'.) TYPE(I)←4, VALUE(I)←VALUE(QJ)-VALUE(QK),→L9.

L7. [Kertolasku.] (Algoritmissa A '$q_i = q_j \cdot q_k$'.) TYPE(I)←5, VALUE(I)←VALUE(QJ)*VALUE(QK),→L9.

L8. [Jakolasku.] (Algoritmissa A '$q_i = q_j/q_k$'.) TYPE(I)←6, VALUE(I)←VALUE(QJ)/VALUE(QK).

L9. [Linkki 2. operandiin.] OPER2(I)←QK.

L10. [Linkki 1. operandiin.] OPER1(I)←QJ.

L11. [Nimike kuntoon.] QI←I, I←I+1,→L2.

<u>Algoritmi L, osa 2.</u>

L12. [COEFF-kenttien alkuasetus.] Nollaa kentät COEFF(K), K = 1,2,...,N-1. COEFF(N)←1, K←N.

L13. [Tyyppivalinta.] Mene askeleeseen LX, missä X = 13+TYPE(K).

L14. [Alkuarvo.] →L21.

L15. [Negatointi.] COEFF(OPER1(K))←COEFF(OPER1(K)) -COEFF(K), COEFF(K)←0 (koska negatointi on tarkka toimitus), →L21.

L16. [Yhteenlasku.] D1←1, D2←1, →L20.

L17. [Vähennyslasku.] D1←1, D2← -1, →L20.

L18. [Kertolasku.] D1←VALUE(OPER2(K)), D2←VALUE(OPER1(K)), →L20.

L19. [Jakolasku.] D1←1/VALUE(OPER2(K)), D2← -VALUE(OPER1(K))/VALUE(OPER2(K))**2.

L20. [COEFF-kenttien käsittely.]
COEFF(OPER1(K))←COEFF(OPER1(K))+D1*COEFF(K),
COEFF(OPER2(K))←COEFF(OPER2(K))+D2*COEFF(K).

L21. [Indeksin vähennys.] K←K-1. Jos K > 0, →L13.

L22. [Sarjatyypin valinta.] (Jos halutaan (R,r)-sarjan kertoimet, algoritmi loppuu tähän.) COEFF(K)←COEFF(K)*VALUE(K), K = 1,2,...,N. (Jos halutaan (R,e)-sarjan kertoimet, algoritmi päättyy tähän.) COEFF(K)←COEFF(K)/VALUE(N), K = 1,2,...,N. Algoritmi L päättyy tähän (COEFF-kentissä on (E,e)-sarjan kertoimet). ▮

Algoritmia L vastaava FORTRAN IV - aliohjelmaryhmä

Ohjelmoitaessa algoritmia L IBM 7094:lle FORTRAN IV-kielellä on algoritmiin vielä tehty lisäys: mikäli laskutoimitus on varmasti tarkka, se huomioidaan ohjelmassa vaihtamalla TYPE- kentän etumerkki miinukseksi. Tällöin tulevat kysymykseen tapaukset, joissa

- operaatio on negatointi (ei merkitystä, koska kerroin nollautuu joka tapauksessa)
- yhteen- tai vähennyslaskun operandi on nolla
- kertolaskun operandi tai jakaja on itseisarvoltaan yksi
- lopputuloksen arvo on nolla.

Tämä tieto saattaa olla tarpeellinen määrättäessä kumulatiivisen pyöristysvirheen jakautumaa, koska eräissä algoritmeissa (esimerkiksi matriisin käännössä) näitä tapauksia on ratkaisevasti enemmän kuin satunnaisuuden perusteella voitaisiin olettaa.

Lisäksi ohjelman avulla voidaan 'määrätä' tietty laskutoimitus tarkaksi tai epätarkaksi, jolloin edellä luetelluilla tapauksilla ei ole vaikutusta TYPE-kentän etumerkkiin.

FORTRAN-ohjelma on ryhmä aliohjelmia, jotka on liitetty yhdeksi monihaaraiseksi function-aliohjelmaksi, jotta tietuepinon tilanvarausta ei tarvitsisi määritellä uudelleen jokaisen algoritmin A alkeistoimituksen aikana. Seuraavassa esitellään tähän aliohjelmaryhmään, 'ryhmään L' kuuluvat aliohjelmat.

Nimikkeet QI, QJ, QK, ja QN ovat kokonaismuuttujia. Eräissä kutsuissa esiintyvällä kokonaismuuttujalla K ei ole merkitystä; sen arvoksi tulee nolla.

Ryhmän L aliohjelmat

LBEGIN   Kutsu: K = LBEGIN(VALUE,TYPE,ØPER1,ØPER2, CØEFF,M)

Kutsun on esiinnyttävä pääohjelmassa ennen muiden ryhmän L aliohjelmien kutsuja. Aliohjelmaryhmä saa tiedon taulukoiden VALUE(M),TYPE(M), ...,CØEFF(M) sijainnista. Pinoindeksille I annetaan alkuarvo 1.

TYPE, ØPER1 ja ØPER2 ovat pääohjelmassa määriteltyjä kokonais-, VALUE ja CØEFF reaalilukutaulukoita. Kokonaisluku M ilmoittaa kaikkien taulukoiden ulottuvuuden.

LNAME   Kutsu: QI = LNAME(VAL)

Kaikki algoritmin A alkuarvot on ilmoitettava tällä kutsulla ryhmälle L. Alkuarvon, reaaliluvun VAL, nimikkeeksi asetetaan QI. Alkuarvo katsotaan epätarkaksi, so. siihen liittyy pyöristysvirhe (TYPE(QI):ksi tulee +1).

LNAMEX   Kutsu: QI = LNAMEX(VAL)

Kuten LNAME, mutta alkuarvo katsotaan tarkaksi (TYPE(QI):ksi tulee -1).

LNEG   Kutsu: QI = LNEG(QJ)

Muuttuja, jonka nimike on QJ, negatoidaan. Negaation nimikkeeksi asetetaan QI.

LADD Kutsu: QI = LADD(QJ,QK)

Muuttujat, joiden nimikkeet ovat QJ ja QK, lasketaan yhteen. Tuloksen nimike on QI.

LADDX Kutsu: QI = LADDX(QJ,QK)

Kuten LADD, mutta laskutoimitus katsotaan aina tarkaksi.

LADDN Kutsu: QI = LADDN(QJ,QK)

Kuten LADD, mutta laskutoimitus katsotaan aina epätarkaksi.

LSUB Kutsu: QI = LSUB(QJ,QK)

Muuttuja, jonka nimike on QK, vähennetään muuttujasta, jonka nimike on QJ. Erotuksen nimike on QI.

LSUBX Kutsu: QI = LSUBX(QJ,QK)

Kuten LSUB, mutta laskutoimitus katsotaan aina tarkaksi.

LSUBN Kutsu: QI = LSUBN(QJ,QK)

Kuten LSUB, mutta laskutoimitus katsotaan aina epätarkaksi.

LMUL Kutsu: QI = LMUL(QJ,QK)

Muuttujat, joiden nimikkeet ovat QJ ja QK, kerrotaan keskenään. Tulon nimike on QI.

LMULX Kutsu: QI = LMULX(QJ,QK)

Kuten LMUL, mutta laskutoimitus katsotaan aina tarkaksi.

LMULN Kutsu: QI = LMULN(QJ,QK)

Kuten LMUL, mutta laskutoimitus katsotaan aina epätarkaksi.

LDIV Kutsu: QI = LDIV(QJ,QK)

Suoritetaan jakolasku. QJ on jaettavan, QK jakajan ja QI osamäärän nimike.

LDIVX Kutsu: QI = LDIVX(QJ,QK)

Kuten LDIV, mutta laskutoimitus katsotaan aina tarkaksi.

<u>LDIVN</u> Kutsu: QI = LDIVN(QJ,QK)

Kuten LDIV, mutta laskutoimitus katsotaan aina epätarkaksi.

<u>LEND</u> Kutsu: K = LEND(QN)

Muuttujan, jonka nimike on QN, (R,r)-sarjan kertoimet lasketaan kenttiin CØEFF(1),CØEFF(2), ...,CØEFF(QN). CØEFF(QN+1) =...= CØEFF(M) = 0.

<u>LREL</u> Kutsu: K = LREL(CØEFIC,M)

Edeltävässä LEND-käskyssä mainitun muuttujan (nimike QN) (E,e)-sarjan kertoimet lasketaan kenttiin CØEFIC(1),CØEFIC(2),...,CØEFIC(QN).

CØEFIC on pääohjelmassa määritelty reaalilukutaulukko, jonka ulottuvuus on M. Voi olla myös CØEFIC = CØEFF.

<u>LABS</u> Kutsu: K = LABS(CØEFIC,M)

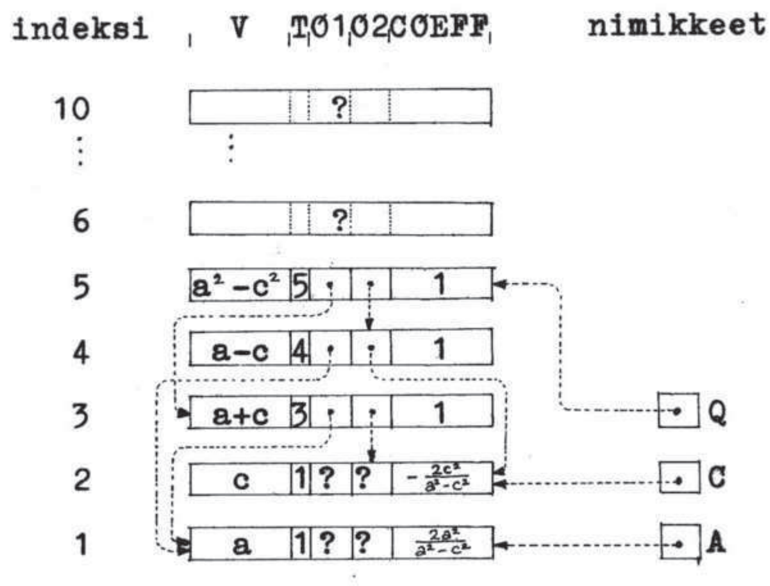
Kuten LREL, mutta CØEFIC-taulukkoon tulevat (R,e)-sarjan kertoimet.

Jos LEND-kutsua seuraa sekä LREL- että LABS-kutsu, ei näistä ensimmäiselle saa olla CØEFIC = CØEFF.

Itse aliohjelmaryhmä L on esitetty liitteessä. Esimerkkinä sen käytöstä ohjelmoimme pienalgoritmin $a^2-c^2 = (a+c)\cdot(a-c)$. Olkoon muuttujan VA arvo a ja muuttujan VC arvo c. Tällöin saamme a^2-c^2:n (E,e)-sarjan kertoimet kenttiin CØEFF(1),CØEFF(2), ...,CØEFF(Q) esimerkiksi ohjelmalla

```
INTEGER T(10),Ø1(10),Ø2(10),A,C,Q
REAL V(10),CØEFF(10)
I = LBEGIN(V,T,Ø1,Ø2,CØEFF,10)
A = LNAME(VA)
C = LNAME(VC)
Q = LMUL(LADD(A,C),LSUB(A,C))
I = LEND(Q)+LREL(CØEFF,10)  .
```

Aliohjelmaryhmän L muodostama rakenne arvoineen ohjelman suorituksen jälkeen on esitetty kuvassa 6, kysymysmerkillä merkittyjä kenttiä ohjelma ei ole käsitellyt.



kuva 6

Yksikköhäiriön menetelmä

Algoritmin L varjopuolena on, että sen käyttö vaatii paljon muistitilaa, koska jokaiselle algoritmin välitulokselle on muodostettava oma tietueensa. Tilan säästämiseksi voidaan käyttää professori Tienarin esittämää 'yksikköhäiriön menetelmää', joka puolestaan käyttää runsaasti enemmän koneaikaa, koska algoritmi on tällöin laskettava läpi kerran kutakin yksittäistä pyöristysvirhettä kohden. Yksikköhäiriön menetelmää varten ei voida luoda ryhmää L vastaavia yleisiä aliohjelmia.

Yksikköhäiriön menetelmän periaatteena on, että algoritmin tuloksen q_N 'tarkka' arvo lasketaan kaksoistarkkaa aritmetiikkaa käyttäen, jolloin voidaan asettaa yhtälössä (5.2) $E_N = 0$. Tämän jälkeen annetaan vuorollaan kullekin yksittäiselle suhteelliselle pyöristysvirheelle suuruusluokkaa b^{-t} oleva arvo. Kun algoritmi lasketaan 'vuorossa' oleva virhe e_i huomioon ottaen uudelleen läpi, saadaan tulokseksi q_N^*. Koska muut yksittäiset virheet ovat nollia, saamme yhtälön (5.2) muotoon

$$E_N = a_{N.i}\, e_i \ . \tag{7.2}$$

Tiedämme, että $E_N = (q_N^* - q_N)/q_N$. Samoin tiedämme e_i:n suuruuden, joten kaavan (7.2) perusteella voimme ratkaista $a_{N.i}$:n kaavasta

$$a_{N.i} = \frac{q_N^* - q_N}{q_N\, e_i} \ . \tag{7.3}$$

Vastaavalla tavalla saadaan $a_{N.i}$ ratkaistuksi kaikilla i:n arvoilla.

Esimerkkinä yksikköhäiriön menetelmästä ohjelmoimme jälleen pienalgoritmin $a^2 - c^2 = (a+c)\cdot(a-c)$. ER-muuttujan arvona on suuruusluokkaa b^{-t} oleva luku, ja virhekertoimet tulevat muuttujien COEFF(1),..., COEFF(5) arvoiksi (haluttaessa ne voitaisiin tulostaa heti kun kukin niistä on laskettu). Taulukko E ajatellaan valmiiksi nollatuksi.

```
      REAL COEFF(5)
      DOUBLE PRECISION A,C,QA,QC,QX,QN,E(5)
      QX = (A+C)*(A-C)
      QXER = QX*ER
      DO 10 I = 1,5
      E(I) = ER
      QA = A*(1.+E(1))
      QC = C*(1.+E(2))
      QN = (QA+QC)*(1.+E(3))*(QA-QC)*(1.+E(4))*(1.+E(5))
      COEFF(I) = (QN-QX)/QXER
   10 E(I) = 0.
```

Algoritmin L ja yksikköhäiriön menetelmän antamat tulokset ovat likimain yhtä tarkkoja tarkkuuden vähentyessä laskutoimitusten lukumäärän kasvaessa. Mikäli tämä tarkkuus ei riitä, voidaan algoritmi L ohjelmoida käyttämään kaksoistarkkaa aritmetiikkaa, jolloin lasketut arvot ovat ratkaisevasti tarkempia.

8. PIENALGORITMIT a^2-c^2:N LASKEMISEKSI

Esimerkkinä pyöristysvirheen Taylor-sarjan käytöstä pyrimme selvittämään sen avulla, kumpi lausekkeen a^2-c^2 kahdesta mahdollisesta laskutavasta,

$$1. \quad a^2-c^2 = (a+c)\cdot(a-c) \tag{8.1}$$

vai

$$2. \quad a^2-c^2 = a\cdot a-c\cdot c \tag{8.2}$$

on edullisempi.

Algoritmien yksinkertaisuuden vuoksi voimme laskea niiden Taylor-sarjat luvussa 5 esitetyllä tavalla. Tällöin

$$\begin{aligned}
&1. \quad ((a^*+c^*)^*\cdot(a^*-c^*)^*)^* \\
&= \{[a(1+e_a)+c(1+e_c)](1+e_1)[a(1+e_a)-c(1+e_c)](1+e_2)\}(1+e_3) \\
&= a^2-c^2+2a^2e_a-2c^2e_c+(a^2-c^2)e_1+(a^2-c^2)e_2+(a^2-c^2)e_3+\langle e^2\rangle \\
&= (a^2-c^2)\left(1+\tfrac{2a^2}{a^2-c^2}\underset{A}{e_a}-\tfrac{2c^2}{a^2-c^2}\underset{A}{e_c}+\underset{S}{e_1}+\underset{S}{e_2}+\underset{T}{e_3}+\langle e^2\rangle\right) \\
&= (a-c)(1+E_1)
\end{aligned} \tag{8.3}$$

ja

$$\begin{aligned}
&2. \quad ((a^*\cdot a^*)^*-(c^*\cdot c^*)^*)^* \\
&= \{[a(1+e_a)a(1+e_a)](1+e_4)-[c(1+e_c)c(1+e_c)](1+e_5)\}(1+e_6) \\
&= a^2-c^2+2a^2e_a-2c^2e_c+a^2e_4-c^2e_5+(a^2-c^2)e_6+\langle e^2\rangle \\
&= (a^2-c^2)\left(1+\tfrac{2a^2}{a^2-c^2}\underset{A}{e_a}-\tfrac{2c^2}{a^2-c^2}\underset{A}{e_c}+\tfrac{a^2}{a^2-c^2}\underset{T}{e_4}-\tfrac{c^2}{a^2-c^2}\underset{T}{e_5}+\underset{S}{e_6}+\langle e^2\rangle\right) \\
&= (a^2-c^2)(1+E_2)\ ,
\end{aligned} \tag{8.4}$$

missä e_a ja e_c ovat alkuarvojen a ja c pyöristysvirheitä, e_3, e_4 ja e_5 tulon sekä e_1, e_2 ja e_6 summan tai erotuksen pyöristysvirheitä. Edellä (kuva 6) totesimme tapauksessa 1 algoritmin L johtavan samaan Taylor-kehitelmään kuin (8.3).

Oletamme, että $|a| > |c|$, mikä ei ole oleellinen rajoitus. Jos tarkastelemme suurimpia mahdollisia virheitä, saamme

$$\max(E_1) \approx (\frac{2a^2}{|a^2-c^2|}+\frac{2c^2}{|a^2-c^2|}+3)\cdot\max(e) , \qquad (8.5)$$

$$\max(E_2) \approx (\frac{3a^2}{|a^2-c^2|}+\frac{3c^2}{|a^2-c^2|}+1)\cdot\max(e) , \qquad (8.6)$$

missä max(e) tarkoittaa suurinta mahdollista yksittäistä pyöristysvirhettä. Esimerkiksi pyöristävässä aritmetiikassa, jossa pyöristykset suoritetaan laskutoimituksen jälkeen, on yhtälöiden (3.17) ja (4.3) perusteella $\max(e) = \frac{1}{2}ub$.

Yhtälöiden (8.5) ja (8.6) perusteella $\max(E_1) < \max(E_2)$ eli tapa 1 on edullisempi, kun

$$2|a^2-c^2| < a^2+c^2 \qquad (8.7)$$

eli, koska $|a| > |c|$, kun

$$\left|\frac{a}{c}\right| < \sqrt{3} \approx 1.7302 . \qquad (8.8)$$

Jos pyrimme selvittämään, kumpi laskutavoista on suuremmalla todennäköisyydellä edullinen, meidän on tarkasteltava virheiden odotusarvoja ja variansseja. Kaavojen (5.8) ja (5.9) perusteella

$$E(E_1) \approx 2\mu_A+2\mu_S+\mu_T , \qquad (8.9)$$

$$E(E_2) \approx 2\mu_A+\mu_S+\mu_T , \qquad (8.10)$$

$$D^2(E_1) \approx 4\frac{a^4+c^4}{(a^2-c^2)^2}\sigma_A^2+2\sigma_S^2+\sigma_T^2 , \qquad (8.11)$$

$$D^2(E_2) \approx 4\frac{a^4+c^4}{(a^2-c^2)^2}\sigma_A^2+\sigma_S^2+\frac{a^4+c^4}{(a^2-c^2)^2}\sigma_T^2 . \qquad (8.12)$$

Yksinkertaistamme tehtävää olettamalla, että aritmetiikka on pyöristävä, jolloin $E(E_1) = E(E_2) = 0$. Tapa 1 on tällöin edullisempi, kun $D^2(E_1) < D^2(E_2)$ eli

$$(a^2-c^2)^2\sigma_S^2 < 2a^2c^2\sigma_T^2 . \qquad (8.13)$$

Kun $|a| \approx |c|$, on $a^2-c^2 \approx 0$ ja tapa 1 on edullisempi. Muulloin, ellei $|a| \gg |c|$, on $\sigma_S^2 \approx \sigma_T^2$. Tällöin tapa 1 on kaavan (8.13) mukaan edullisempi, mikäli

$$\left|\frac{a}{c}\right| < \sqrt{2+\sqrt{3}} \approx 1.932 \quad . \qquad (8.14)$$

Kun a:n ja c:n eksponenttien erotus on $>t$, on

$$(a^* \pm c^*)^* = a^* \quad . \qquad (8.15)$$

Myös eksponenttien erotuksen arvolla t yhtälö (8.15) on voimassa, kunhan c:n mantissan itseisarvo $< \frac{1}{2}$. Koska a:n mantissan itseisarvo on kaavan (2.6) mukaan < 1, voimme todeta yhtälön (8.15) olevan voimassa aina, jos

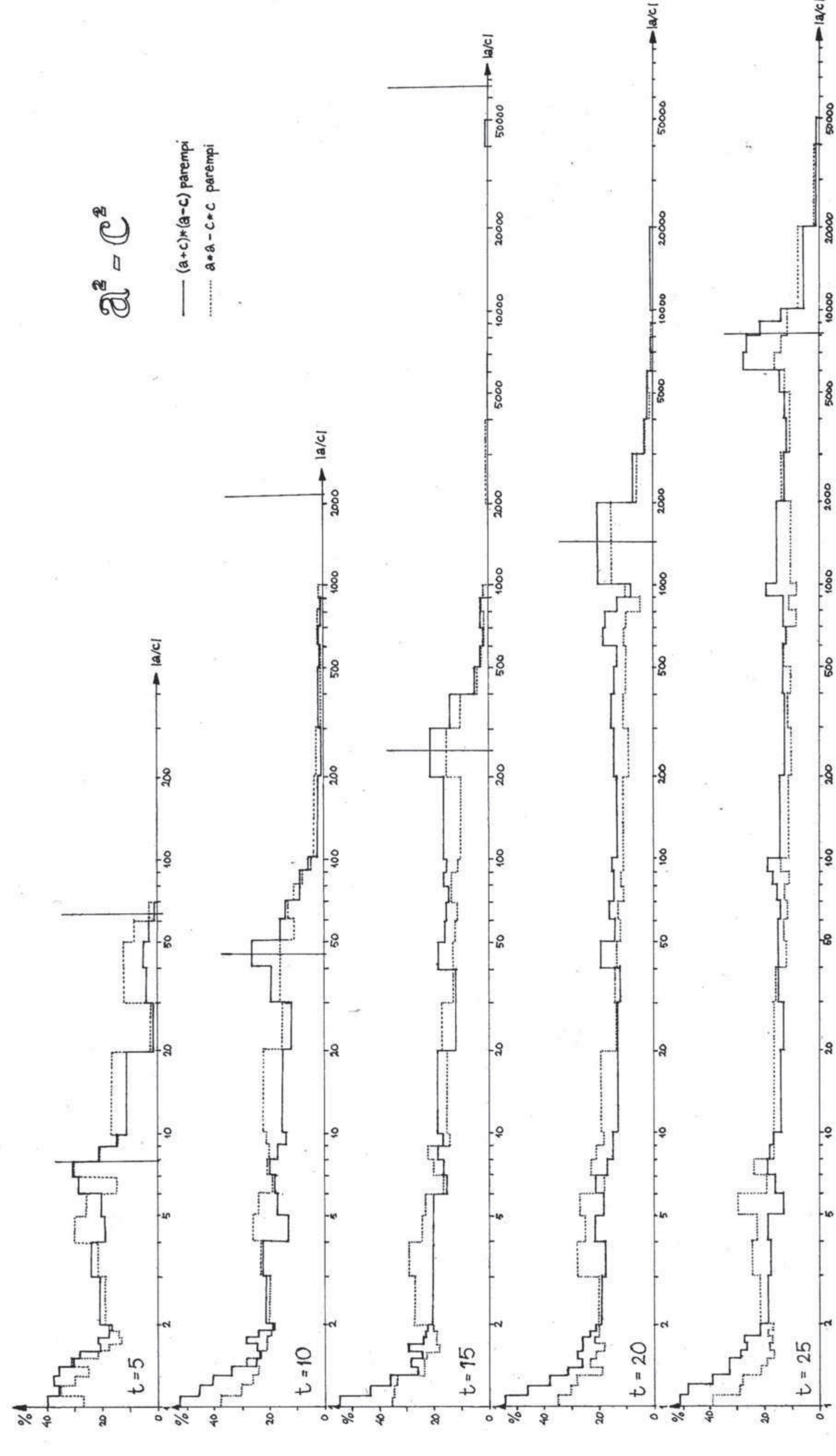
$$\left|\frac{a}{c}\right| \approx \left|\frac{a^*}{c^*}\right| > \frac{1}{\frac{1}{2}\cdot b^{-t}} = 2b^t \quad . \qquad (8.16)$$

Laskutapa 1 supistuu siis tällöin muotoon a·a. Tavassa 2 voimme vastaavasti korvata c:n nollalla, jos

$$\left|\frac{a}{c}\right| \approx \sqrt{\frac{(a^* \cdot a^*)^*}{(c^* \cdot c^*)^*}} > \sqrt{2b^t} = \sqrt{2}\cdot b^{t/2} \quad . \qquad (8.17)$$

Koska $2\cdot b^t > \sqrt{2}\cdot b^{t/2}$, kun $b \geq 2$ ja $t \geq 1$, riittää ehto (8.16) molemmissa tavoissa takaamaan, ettei c:llä ole vaikutusta lopputulokseen. Kaavoissa (8.3) ja (8.4) tämä merkitsee, että c, e_1, e_2 ja e_6 voidaan korvata nollalla, jolloin laskutavat päätyvät samaan tulokseen eikä siis kumpikaan ole toistaan edullisempi.

Saatujen tulosten testaamiseksi laskettiin a^2-c^2:n arvoja satunnaisilla a:n ja c:n arvoilla. Tulosten luetteloimiseksi ne jaettiin $|a/c|$:stä riippuviin luokkiin, välillä [1,2) toiseksi merkitsevimmän ja välillä [2,100000) merkitsevimmän (desimaali)numeron perusteella. Kussakin luokassa suoritettiin noin 300 laskutoimitusta kummallakin tavalla. Kaik-



kuva 7

ki laskutoimitukset suoritettiin binääriaritmetiikassa ($b = 2$) viidellä erilaisella numeroiden lukumäärällä ($t = 5,10,15,20$ ja 25).

Testin tulos on histogrammana kuvassa 7, jossa abskissana on $|a/c|$:n arvo ja oordinaattana prosenttinen osuus suoritetuista laskutoimituksista. Eheä viiva ilmaisee, monessako prosentissa tapa 1 oli parempi ja katkoviiva vastaavan prosenttimäärän tavasta 2. Lopuissa tapauksissa tulos oli molemmilla tavoilla laskettuna sama. $|a/c|$:n edellä mainitut arvot $\sqrt{2}\cdot b^{t/2}$ ja $2b^t$ on merkitty histogrammeihin pystyviivoilla.

Voimme todeta tavan 1 todella olevan edullisempi, kun $|a/c| \leq 2$. Mitä lähempänä $|a|$ ja $|c|$ ovat toisiaan, sitä todennäköisempää on, että laskutavat 1 ja 2 johtavat eri tuloksiin. Kun $|a/c| > 2$, ei suoritetun testin perusteella voida sanoa, että tapa 2 olisi merkittävästi parempi.

Samaan tulokseen päätymisen todennäköisyys pysyy $|a/c|$:n kasvaessa likimain vakiona aina rajaan $\sqrt{2}\cdot b^{t/2}$ asti. Tämän jälkeen alkaa olla yhä todennäköisempää, että laskutavat antavat saman tuloksen. Kuten kaavan (8.16) perusteella voitiin odottaa, ei eriäviä tuloksia saatu, kun $|a/c|$ oli $>2b^t$.

9. HORNER-SHEMA

Taylor-kehitelmän analyyttinen määrittäminen

Horner-shema on algoritmi, jolla lasketaan polynomin

$$p = a_0x^N+a_1x^{N-1}+ \dots +a_{N-1}x+a_N \ , \quad a_0 \neq 0, \qquad (9.1)$$

arvo pisteessä x. Se kirjoitetaan tavallisesti muotoon [4]

$$\begin{cases} q_0 = a_0 \\ q_n = a_n+xq_{n-1} \ , \quad n = 1,\dots,N \ . \end{cases} \qquad (9.2)$$

Kun hajoitamme algoritmin (9.2) yksittäisiin laskutoimituksiin, saamme sen muotoon

$$\begin{cases} q_{0.2} = a_0 \\ q_{n.1} = xq_{n-1.2} \\ q_{n.2} = a_n+q_{n.1} \ , \quad n = 1,\dots,N \ . \end{cases} \qquad (9.3)$$

Sovellamme tähän algoritmiin analyyttistä menetelmää kumulatiivisen pyöristysvirheen Taylor-sarjan kertoimien laskemiseksi.

Kaavan (6.12) perusteella saamme

$$\begin{cases} R^{(1)}_{n.1} = q_{n-1.2}xe_x +xR^{(1)}_{n-1.2}+q_{n.1}e_{n.1} \\ R^{(1)}_{n.2} = a_ne_n +R^{(1)}_{n.1}+q_{n.2}e_{n.2} \ , \end{cases} \qquad (9.4)$$

missä e_x on x:n ja e_n a_n:n pyöristysvirhe sekä $e_{n.1}$ ja $e_{n.2}$ n:nnen iteraatiokierroksen kerto- ja yhteenlaskun pyöristysvirheet.

Sijoittamalla yhtälöistä (9.4) edellinen jälkimmäiseen saamme ensimmäisen asteen differenssiyhtälöm

$$R^{(1)}_{n.2} = xR^{(1)}_{n-1.2}+xq_{n-1.2}e_x+a_n e_n+xq_{n-1.2}e_{n.1}+q_{n.2}e_{n.2} , \qquad (9.5)$$

jonka alkuarvona on $R^{(1)}_{0.2} = a_0 e_0$ eli a_0:n absoluuttinen pyöristysvirhe.

Muotoa

$$x_n = ax_{n-1}+b_n , \quad x_0 = b_0 \qquad (9.6)$$

olevan differenssiyhtälön ratkaisu on [4]

$$x_n = \sum_{i=0}^{n} a^{n-i} b_i , \qquad (9.7)$$

joten saamme differenssiyhtälön (9.5) ratkaisuksi

$$\begin{aligned} R^{(1)}_{n.2} = {} & \sum_{i=1}^{n} x^{n-i+1} q_{i-1.2} e_x + \sum_{i=0}^{n} x^{n-i} a_i e_i + \sum_{i=1}^{n} x^{n-i+1} q_{i-1.2} e_{i.1} \\ & + \sum_{i=1}^{n} x^{n-i} q_{i.2} e_{i.2} , \quad n = 2,\ldots,N . \end{aligned} \qquad (9.8)$$

Vastaavasti saadaan

$$\begin{cases} R^{(1)}_{1.1} = xq_{0.2}e_x + xa_0 e_0 + xq_{0.2}e_{1.1} \\ R^{(1)}_{n.1} = \sum_{i=1}^{n} x^{n-i+1} q_{i-1.2} e_x + \sum_{i=1}^{n-1} x^{n-i} a_i e_i + \sum_{i=1}^{n} x^{n-i+1} q_{i-1.2} e_{i.1} \\ \qquad + \sum_{i=1}^{n-1} x^{n-i} q_{i.2} e_{i.2} , \quad n = 2,\ldots,N . \end{cases} \qquad (9.9)$$

Tarkastelemme erityisesti algoritmin lopputuloksen $p = q_{N.2}$ pyöristysvirhettä $R_{N.2}$. Otamme käyttöön merkinnät

$$\begin{cases} \alpha_i = x^{N-i} a_i \\ \beta_i = \sum_{j=0}^{i} \alpha_j \\ \gamma_i = \sum_{j=0}^{i} (N-j)\alpha_j \end{cases} \qquad (9.10)$$

Merkinnällä α_i takoitamme siis polynomin p (N-i):nnen asteen termiä ja β_i:llä (N-i):nnen ja sitä korkeamman asteen termien summaa. Kun toteamme, että

$$q_{i.2} = \sum_{j=0}^{i} a_j x^{i-j} , \qquad (9.11)$$

saamme β_i:lle myös lausekkeen

$$\beta_i = x^{N-i} q_{i.2} . \qquad (9.12)$$

Erityisesti

$$\beta_N = p \quad . \qquad (9.13)$$

Polynomin p derivaatan p' arvo on

$$p' = Na_0 x^{N-1} + (N-1)a_1 x^{N-2} + \ldots + 2a_{N-2}x + a_{N-1} \, . \qquad (9.14)$$

Huomaamme, että γ_i on polynomin xp' (N-i):nnen ja sitä korkeamman asteen termien summa, erityisesti

$$\gamma_N = xp' \quad . \qquad (9.15)$$

Voimme lausua γ_i:n myös muodossa

$$\gamma_i = \sum_{j=0}^{i}(N-i)\alpha_j + \sum_{j=0}^{i}(i-j)\alpha_j = (N-i)\beta_j + \sum_{j=0}^{i-1}\beta_j \; , \qquad (9.16)$$

erityisesti

$$\gamma_N = \gamma_{N-1} = \sum_{i=0}^{N-1}\beta_i = \sum_{i=1}^{N-1} x^{N-i+1} q_{i-1.2} \quad . \qquad (9.17)$$

Saamme yhtälön (9.8) yhtälöiden (9.10), (9.12) ja (9.17) avulla muotoon

$$R_{N.2}^{(1)} = \gamma_N e_x + \sum_{i=0}^{N}\alpha_i e_i + \sum_{i=1}^{N}\beta_{i-1} e_{i.1} + \sum_{i=1}^{N}\beta_i e_{i.2} \; , \quad N \geq 2, \qquad (9.18)$$

josta yhtälön (6.4) perusteella saamme polynomin p (E,e)-sarjaksi

$$E_{N.2}^{(1)} = \frac{\gamma_N}{p} e_x + \sum_{i=0}^{N}\frac{\alpha_i}{p} e_i + \sum_{i=1}^{N}\frac{\beta_{i-1}}{p} e_{i.1} + \sum_{i=1}^{N}\frac{\beta_i}{p} e_{i.2} \; , \quad N \geq 2 \, . \qquad (9.19)$$

Kaavojen (5.8) ja (5.9) mukaan saamme $E_{N.2}^{(1)}$:n odotusarvon ja varianssin muotoon

$$E(E_{N.2}^{(1)}) = \frac{1}{p}(\gamma_N + \sum_{i=0}^{N}\alpha_i)\mu_A + \sum_{i=1}^{N}\frac{\beta_i}{p}\mu_S + \sum_{i=1}^{N}\frac{\beta_{i-1}}{p}\mu_T \; , \qquad (9.20)$$

$$D^2(E_{N.2}^{(1)}) = \frac{1}{p^2}(\gamma_N^2 + \sum_{i=0}^{N}\alpha_i^2)\sigma_A^2 + \sum_{i=1}^{N}\frac{\beta_i^2}{p^2}\sigma_S^2 + \sum_{i=1}^{N}\frac{\beta_{i-1}^2}{p^2}\sigma_T^2 \, . \qquad (9.21)$$

Kun oletamme alkuarvot tarkoiksi ja laskutoimitukset suoritettavaksi kiinteän pilkun aritmetiikkaa käyttäen, jolloin yhteenlaskut ovat tarkkoja, on yhtälössä (9.18) $e_x = e_i = e_{i.2} = 0$, $i = 0,\ldots,N$. Kaavojen (6.17) ja (9.12) perusteella saamme yhtälön (9.18) tällöin muotoon

$$R^{(1)}_{N.2} = \sum_{i=1}^{N} \beta_{i-1} \frac{r_{i.1}}{q_{i.1}} = \sum_{i=1}^{N} x^{N-i+1} q_{i-1.2} \frac{r_{i.1}}{x q_{i-1.2}} = \sum_{i=1}^{N} x^{N-i} r_{i.1} \,, \tag{9.22}$$

mikä on sama kuin Henricin saama tulos [4].

Toisen asteen kertoimet

Myös toisen asteen kertoimet Horner-sheman kumulatiivisen pyöristysvirheen Taylor-kehitelmässä voidaan johtaa analyyttisesti. Kaavojen (6.13) ja (9.3) perusteella

$$\begin{cases} R^{(2)}_{n.1} = q_{n-1.2} \cdot 0 + xR^{(2)}_{n-1.2} + xe_x R^{(1)}_{n-1.2} + q_{n-1.2} x e_x e_{n.1} + xR^{(1)}_{n-1.2} e_{n.1} \\ R^{(2)}_{n.2} = 1 \cdot 0 + 1 \cdot R^{(2)}_{n.1} + 0 + a_n e_n e_{n.2} + R^{(1)}_{n.1} e_{n.2} \,. \end{cases} \tag{9.21}$$

Sijoittamalla ensimmäisen yhtälön toiseen saamme ensimmäisen asteen differenssiyhtälön

$$\begin{aligned} R^{(2)}_{n.2} = {} & xR^{(2)}_{n-1.2} + xe_x R^{(1)}_{n-1.2} + xq_{n-1.2} e_x e_{n.1} + xR^{(1)}_{n-1.2} \\ & + a_n e_n e_{n.2} + R^{(1)}_{n.1} e_{n.2} \,. \end{aligned} \tag{9.22}$$

Alkuarvon $R^{(2)}_{0.2} = 0$ avulla saamme sen ratkaisuksi

$$\begin{aligned} R^{(2)}_{n.2} = {} & \sum_{i=1}^{n} x^{n-i+1} e_x R^{(1)}_{i-1.2} + \sum_{i=1}^{n} x^{n-i+1} q_{i-1.2} e_x e_{i.1} + \sum_{i=1}^{n} x^{n-i+1} R^{(1)}_{i-1.2} e_{i.1} \\ & + \sum_{i=1}^{n} x^{n-i} a_i e_i e_{i.2} + \sum_{i=1}^{n} x^{n-i} R^{(1)}_{i.1} e_{i.2} \,, \quad n = 1,\ldots,N \,. \end{aligned} \tag{9.23}$$

Sijoittamalla tähän yhtälöön kaavojen (9.8) ja (9.9) tulokset saamme

$$\begin{aligned} R^{(2)}_{n.2} = {} & a_0 x^n e_0 (e_x + e_{1.1}) + \sum_{i=2}^{n} x^{n-i+1} \Big(\sum_{j=1}^{i-1} x^{i-j} q_{j-1.2} e_x + \sum_{j=0}^{i-1} x^{i-j-1} a_j e_j \\ & + \sum_{j=1}^{i-1} x^{i-j} q_{j-1.2} e_{j.1} + \sum_{j=1}^{i-1} x^{i-j-1} q_{j.2} e_{j.2} \Big) (e_x + e_{i.1}) + \sum_{i=1}^{n} x^{n-i+1} q_{i-1.2} e_x e_{i.1} \\ & + \sum_{i=1}^{n} x^{n-i} a_i e_i e_{i.2} + \sum_{i=1}^{n} x^{n-i} \Big(\sum_{j=1}^{i} x^{i-j+1} q_{j-1.2} e_x + \sum_{j=0}^{i-1} x^{i-j} a_j e_j \\ & + \sum_{j=1}^{i} x^{i-j+1} q_{j-1.2} e_{j.1} \Big) e_{i.2} + \sum_{i=2}^{n} \sum_{j=1}^{i-1} x^{n-i} x^{i-j} q_{j.2} e_{j.2} e_{i.2} \,, \\ & n = 2,\ldots,N \,. \end{aligned} \tag{9.24}$$

Erityisesti kaavojen (9.10) ja (9.12) perusteella

$$
\begin{aligned}
R^{(2)}_{N,2} = {} & \alpha_0 e_0 (e_x + e_{1.1}) + \sum_{i=2}^{N} \Big(\sum_{j=1}^{i-1} \beta_{j-1} e_x + \sum_{j=0}^{i-1} \alpha_j e_j + \sum_{j=1}^{i-1} \beta_{j-1} e_{j.1} + \sum_{j=1}^{i-1} \beta_j e_{j.2} \Big) (e_x + e_{i.1}) \\
& + \sum_{i=1}^{N} \beta_{i-1} e_x e_{i.1} + \sum_{i=1}^{N} \alpha_i e_i e_{i.2} + \sum_{i=1}^{N} \Big(\sum_{j=1}^{i} \beta_{j-1} e_x + \sum_{j=0}^{i-1} \alpha_j e_j \\
& + \sum_{j=1}^{i} \beta_{j-1} e_{j.1} \Big) e_{i.2} + \sum_{i=2}^{N} \sum_{j=1}^{i-1} \beta_j e_{j.2} e_{i.2} \quad .
\end{aligned}
\tag{9.25}
$$

Kaavan

$$
\sum_{i=k}^{N} \sum_{j=k-1}^{i-1} c_j = \sum_{j=k-1}^{N-1} (N-j) c_j \tag{9.26}
$$

avulla saamme yhtälön (9.25) muotoon

$$
\begin{aligned}
R^{(2)}_{N.2} = {} & \sum_{j=1}^{N-1} (N-j) \beta_{j-1} e_x e_x + \sum_{j=0}^{N} (N-j) \alpha_j e_x e_j + \sum_{i=1}^{N} \Big[(N-i) \beta_{i-1} \\
& + \sum_{j=1}^{i} \beta_{j-1} \Big] e_x e_{i.1} + \sum_{i=1}^{N} \Big[(N-i) \beta_i + \sum_{j=1}^{i} \beta_{j-1} \Big] e_x e_{i.2} \\
& + \sum_{i=1}^{N} \sum_{j=0}^{i-1} \alpha_j e_j e_{i.1} + \sum_{i=1}^{N} \sum_{j=0}^{i-1} \alpha_j e_j e_{i.2} + \sum_{i=1}^{N} \alpha_i e_i e_{i.2} \\
& + \sum_{i=2}^{N} \sum_{j=1}^{i-1} \beta_{j-1} e_{i.1} e_{j.1} + \sum_{i=2}^{N} \sum_{j=1}^{i-1} \beta_j e_{i.2} e_{j.2} + \sum_{i=1}^{N} \sum_{j=1}^{i} \beta_{j-1} e_{j.1} e_{i.2} \\
& + \sum_{i=2}^{N} \sum_{j=1}^{i-1} \beta_j e_{i.1} e_{j.2} \, , \qquad N \geq 2 \, .
\end{aligned}
\tag{9.27}
$$

Kaavan (9.10) perusteella

$$
\begin{aligned}
\sum_{j=1}^{N-1} (N-j) \beta_{j-1} & = \sum_{j=1}^{N-1} \sum_{k=0}^{j-1} (N-j) \alpha_k = \sum_{k=0}^{N-2} \tfrac{1}{2} (N-k-1)(N-k) \alpha_k \\
& = \tfrac{1}{2} x^2 p'' \quad ,
\end{aligned}
\tag{9.28}
$$

missä p'' on polynomin p toinen derivaatta pisteessä x. Kun vielä käytämme kaavoja (9.16) ja (6.4), saamme

$$
\begin{aligned}
E^{(2)}_{n.2} = {} & \frac{1}{2p} x^2 p'' e_x e_x + \sum_{j=0}^{N} (N-j) \frac{\alpha_j}{p} e_x e_j + \sum_{i=1}^{N} \frac{\gamma_{i-1}}{p} e_x e_{i.1} + \sum_{i=1}^{N} \frac{\gamma_i}{p} e_x e_{i.2} \\
& + \sum_{i=1}^{N} \sum_{j=0}^{i-1} \frac{\alpha_j}{p} e_j e_{i.1} + \sum_{i=1}^{N} \sum_{j=0}^{i} \frac{\alpha_j}{p} e_j . e_{i.2} + \sum_{i=2}^{N} \sum_{j=1}^{i-1} \frac{\beta_{j-1}}{p} e_{i.1} e_{j.1} \\
& + \sum_{i=2}^{N} \sum_{j=1}^{i-1} \frac{\beta_j}{p} e_{i.1} e_{j.2} + \sum_{i=1}^{N} \sum_{j=1}^{i} \frac{\beta_{j-1}}{p} e_{j.1} e_{i.2} + \sum_{i=2}^{N} \sum_{j=1}^{i-1} \frac{\beta_j}{p} e_{i.2} e_{j.2} \, , \quad N \geq 2 \, .
\end{aligned}
\tag{9.29}
$$

Kaavojen (9.19) ja (9.29) antamat tulokset p:n (E,e)-sarjan ensimmäisen ja toisen asteen kertoimille on esitetty alla olevassa taulukossa, kun polynomin asteluku $N \geq 1$.

tekijä	kerroin*p	ehto tekijän indekseille	ao. termien lukumäärä
e_x	γ_N	-	1
e_i	α_i	$i \geq 0$	$N+1$
$e_{i.1}$	β_{i-1}	$i \geq 1$	N
$e_{i.2}$	β_i	$i \geq 1$	N
$e_x e_x$	$\frac{1}{2}x^2p''$	-	1 (0, jos $N=1$)
$e_x e_i$	$(N-i)\alpha_i$	$i \geq 0$	$N+1$
$e_x e_{i.1}$	γ_{i-1}	$i \geq 1$	N
$e_x e_{i.2}$	γ_i	$i \geq 1$	N
$e_i e_{j.1}$	α_i	$j > i \geq 0$	$\frac{1}{2}(N^2+N)$
$e_i e_{j.2}$	α_i	$j>i\geq 0$ tai $i=j\geq 1$	$\frac{1}{2}(N^2+3N)$
$e_{i.1} e_{j.1}$	β_{i-1}	$j > i \geq 1$	$\frac{1}{2}(N^2-N)$
$e_{i.1} e_{j.2}$	β_j	$i > j \geq 1$	$\frac{1}{2}(N^2-N)$
$e_{i.1} e_{j.2}$	β_{i-1}	$j \geq i \geq 1$	$\frac{1}{2}(N^2+N)$
$e_{i.2} e_{j.2}$	β_i	$j > i \geq 1$	$\frac{1}{2}(N^2-N)$

Voimme todeta ensimmäisen asteen termejä olevan kaikkiaan $3N+2$ ja toisen asteen termejä $3N^2+4N+2$ kappaletta eli toisen asteen termien lukumäärä on suurilla N:n arvoilla noin N-kertainen verrattuna ensimmäisen asteen termien lukumäärään.

Ensimmäisen ja toisen asteen kertoimet edustavat samaa suuruusluokkaa, mutta e^2 on suuruusluokkaa $b^{-t}e$. Jotta toisen asteen termien vaikutus olisi samaa suuruusluokkaa kuin ensimmäisen asteen termien, olisi siis N:n oltava suuruusluokkaa b^t. Tällöin pyöristysvirheet ovat kuitenkin niin suuria, ettei polynomin arvon laskeminen t numeron tarkkuudella yleensä enää ole mielekästä tuloksen epätarkkuuden johdosta.

Esimerkkinä mainitusta epätarkkuudesta tarkastelemme polynomia $p = x^N$, missä $N = b^t$. Koska $a_0 = 1$ ja $a_i = 0$, $i = 1,\ldots,N$, on $e_{1.1} = \theta$ ja $e_i = e_{i.2} = 0$, $i = 1,\ldots,N$. Kaavojen (9.20), (3.23) ja (4.9) perusteella on $E^{(1)}_{N.2}$:n odotusarvo katkaisevassa aritmetiikassa

$$E(E^{(1)}_{N.2}) = \frac{1}{p}(\gamma_N\mu_A + \sum_{i=2}^{N}\beta_{i-1}\mu_T) = \frac{1}{p}(Nx^N\mu_A + (N-1)x^N\mu_T)$$

$$\approx 2N\frac{u(1-b)}{2\ln b} = \frac{1-b}{\ln b} \quad .$$

Toisistaan riippumattomien satunnaismuuttujien ξ ja η tulon odotusarvolle voimassa olevan yhtälön [2]

$$E(\xi\eta) = E(\xi)E(\eta) \qquad (9.30)$$

avulla saamme $E^{(2)}_{N.2}$:n odotusarvoksi katkaisevassa aritmetiikassa

$$E(E^{(2)}_{N.2}) = \frac{1}{p}(\tfrac{1}{2}x^2 p''\mu_A^2 + \sum_{i=2}^{N}\gamma_{i-1}\mu_A\mu_T + \sum_{i=3}^{N}\sum_{j=2}^{i-1}\beta_{j-1}\mu_T^2)$$

$$= \frac{1}{p}(\tfrac{1}{2}x^2 N(N-1)x^{N-2}\mu_A^2 + (N-1)Nx^N\mu_A\mu_T + \frac{(N-2)(N-1)}{2}x^N\mu_T^2)$$

$$\approx 2N^2\left[\frac{u(1-b)}{2\ln b}\right]^2 = \frac{1}{2}\left[\frac{1-b}{\ln b}\right]^2 \quad .$$

Alla olevaan taulukkoon on laskettu saatujen odotusarvojen arvoja kantaluvun vaihdellessa.

b	$E(E^{(1)}_{N.2})$	$E(E^{(2)}_{N.2})$
2	-1.44	1.04
8	-3.37	5.67
10	-3.91	7.64
16	-5.41	14.6
64	-15.1	115

Kuten odotimme, edustavat $E(E^{(1)}_{N.2})$ ja $E(E^{(2)}_{N.2})$ samaa suuruusluokkaa, ja niiden arvot ovat niin suuria, että virheen odotusarvo varsinkin suurilla kantaluvun arvoilla on moninkertainen itse polynomin arvoon verrattuna.

Polynomille $p = x^N$, missä $N = \frac{1}{2}b^{t-1}$, saamme vastaavaksi taulukoksi

b	$E(E_{N.2}^{(1)})$	$E(E_{N.2}^{(2)})$
2	-0.36	0.07
8	-0.21	0.022
10	-0.19	0.019
16	-0.17	0.014
64	-0.12	0.0045

Toisen asteen termien vaikutus on jo selvästi ensimmäisen asteen termien vaikutusta pienempi, mutta virheen odotusarvo on vielä varsin suuri.

Saatu tulos vahvistaa yhtälössä (5.4) esitettyä käsitystä, jonka mukaan toisen ja korkeamman asteen termeillä ei ole käytännössä merkitystä kumulatiivista pyöristysvirhettä tarkasteltaessa.

Nollakohtien lähekkäisyyden vaikutus kertoimiin

On tunnettua, että polynomin nollakohtien suhteellinen lähekkäisyys vaikuttaa heikentävästi polynomin arvon laskutarkkuuteen. Tutkimme seuraavassa tätä ilmiötä Taylor-kehitelmän (9.19) kertoimien avulla.

Polynomin p (9.1) kerroin a_i voidaan lausua nollakohtien z_j, $j = 1,\ldots,N$, avulla muodossa

$$a_i = a_o(-1)^i \sum_{j=1}^{K}(z_{j_1} z_{j_2} \cdots z_{j_i}), \; i = 1,\ldots,N, \quad (9.31)$$

missä summattavien tulojen tekijöinä ovat kaikki mahdolliset i nollakohdan kombinaatiot, jolloin $K = N!/(i!(N-i)!)$.

Tarkastelemme nyt polynomia $\bar{p}$, jonka nollakohdat ovat $\bar{z}_j = z_j + M$, $j = 1,\ldots,N$, ja jolle

$$\bar{p}(\bar{x}) = \bar{p}(x+M) = p(x) \quad . \qquad (9.32)$$

Tällöin

$$\bar{p}(\bar{x}) = \bar{a}_0[(x+M)-(z_1+M)] \cdots [(x+M)-(z_N+M)] \\ \equiv \bar{a}_0(x-z_1)\cdots(x-z_N) = \frac{\bar{a}_0}{a_0}p(x) , \tag{9.33}$$

joten

$$\bar{a}_0 = a_0 \tag{9.34}$$

ja kaavan (9.31) mukaan

$$\bar{a}_i = a_0(-1)^i \sum_{j=1}^{K}[(z_{j_1}+M)\cdots(z_{j_i}+M)] \ . \tag{9.35}$$

Kun M kasvaa itseisarvoltaan riittävän isoksi, ovat $\bar{z}_j$:t, $\bar{x}$ ja M samanmerkkisiä. $|M|$:n kasvaessa edelleen kasvaa myös

$$\left|\frac{\bar{\alpha}_i}{\bar{p}}\right| = \left|\frac{\bar{a}_i\bar{x}^{N-i}}{\bar{p}}\right| = \frac{|a_0|\sum(|z_{j_1}+M|\cdots|z_{j_i}+M|)|x+M|^{N-i}}{|p|} , \tag{9.36}$$

$i = 1,\ldots,N$, sekä $|\bar{\alpha}_0/\bar{p}| = |a_0||x+M|^N/|p|$.

Summan ja sen suurimman jäsenen itseisarvot edustavat yleensä samaa suuruusluokkaa, joten voimme odottaa myös $|\bar{\beta}_i/\bar{p}|$:n ja $|\bar{\gamma}_i/\bar{p}|$:n ja siis kaikkien $\bar{E}^{(1)}_{N2}$:n lausekkeessa (9.19) esiintyvien kertoimien kasvavan $|M|$:n kasvaessa.

Nollakohtien arvot lähestyvät toisiaan suhteellisesti $|M|$:n suuretessa, mutta myös $|x|$:t kasvavat, joten on syytä tarkastella nollakohtien origosta mitattujen absoluuttisten etäisyyksien vaikutusta kertoimiin.

Ajattelemme kaikkien x-koordinaattien tulevan kerrotuksi kertoimella k y-koordinaattien pysyessä ennallaan. Tällöin polynomia p vastaa polynomi $\hat{p}$, jolle

$$\hat{p}(\hat{x}) = \hat{p}(kx) = p(x) \ . \tag{9.37}$$

Yhtälön (9.31) mukaan

$$\hat{a}_i = \hat{a}_0(-1)^i \sum_{j=1}^{K}(kz_{j_1}\cdots kz_{j_N}) = \frac{\hat{a}_0}{a_0}k^i a_i, \ i = 1,\ldots,N. \tag{9.38}$$

Tällöin

$$\hat{p}(\hat{x}) = \sum_{i=0}^{N} \hat{a}_i \hat{x}^{N-i} = \sum_{i=0}^{N} \frac{\hat{a}_0}{a_0} k^i a_i k^{N-i} x^{N-i} = \frac{\hat{a}_0 k^N}{a_0} p(x), \quad (9.39)$$

joten yhtälöiden (9.37) ja (9.38) perusteella

$$\hat{a}_i = k^{i-N} a_i \ , \quad i = 0,\dots,N \ . \quad (9.40)$$

Edelleen

$$\frac{\hat{\alpha}_i}{\hat{p}} = \frac{\hat{a}_i \hat{x}^{N-i}}{\hat{p}} = \frac{k^{i-N} a_i k^{N-i} x^{N-i}}{p} = \frac{\alpha_i}{p} \quad (9.41)$$

ja yhtälöiden (9.10) perusteella myös $\hat{\beta}_i/\hat{p} = \beta_i/p$, ja $\hat{\gamma}_i/\hat{p} = \gamma_i/p$, $i = 0,\dots,N$. Kaikki E :n kertoimet pysyvät siis ennallaan mielivaltaisella k:n arvolla.

Myöskään y-koordinaattien kertominen vakioarvolla x-koordinaattien pysyessä ennallaan ei vaikuta $E_{N.2}^{(1)}$:n kertoimiin, sillä $kp(x) = ka_0(x-z_1)\cdots(x-z_N)$ eli tämä kertominen voidaan samaistaa a_0:n kertomiseen vakioarvolla. Voimme helposti todeta, että

$$\frac{\alpha_i}{p} = \frac{a_0(-1)^i \sum (z_{j_1}\cdots z_{j_i}) x^{N-i}}{a_0(x-z_1)\cdots(x-z_N)} \quad (9.42)$$

on riippumaton a_0:sta.

Edellä saatujen tulosten perusteella $E_{N.2}^{(1)}$:n kertoimet määräytyvät yksinomaan x:n ja nollakohtien arvojen suhteista toisiinsa. Nollakohtien ollessa samanmerkkisiä ja lähestyessä toisiaan kasvavat kertoimet yleensä itseisarvoltaan.

Tarkastellessamme tietyn polynomin p eri pisteitä ovat erityisen mielenkiintoisia polynomin nollakohtien ohella $|x|$:n erittäin suuret arvot sekä piste $x = 0$. Yleisenä huomiona voidaan todeta, että $e_{N.2}$:n kerroin β_N/p on aina yksi.

Jos a_i ei ole nolla, α_i/p $(i = 0,\dots,N)$ kasvaa rajatta x:n lähestyessä polynomin nollakohtaa. Samalla yleensä myös β_i/p ja γ_i/p kasvavat rajatta.

Kun $|x|$ kasvaa rajatta, $\alpha_0/p \to 1$ ja $\alpha_i/p \to 0$, $i = 1,\ldots,N$, joten

$$\lim_{|x|\to\infty} E^{(1)}_{N.2} = Ne_x + e_0 + \sum_{i=1}^{N} e_{i.1} + \sum_{i=1}^{N} e_{i.2} \tag{9.43}$$

ja kertoimet riippuvat siis ainoastaan polynomin asteluvusta.

Jos $a_k \neq 0$ ja $a_j = 0$, $j > k$, niin x:n lähestyessä nollaa $\alpha_k/p \to 1$ ja $\alpha_j/p \to 0$, $j \neq k$. Kun otamme huomioon, että tällöin $e_{j.2} = 0$, $j > k$, saamme

$$\lim_{x\to 0} E^{(1)}_{N.2} = (N-k)e_x + e_k + \sum_{i=k+1}^{N} e_{i.1} , \tag{9.44}$$

joten kertoimet riippuvat vain asteluvusta ja indeksin k arvosta. Tapaus $k = 0$, jolloin $p = a_0 x^N$, on erityisen huomion arvoinen. Tällöin E :n lauseke on x:n arvosta riippumatta muotoa

$$E^{(1)}_{N.2} = Ne_x + e_0 + \sum_{i=1}^{N} e_{i.1} . \tag{9.45}$$

Nollakohtien lähekkäisyyden vaikutusta tutkittiin kokeellisesti kolmannen asteen polynomin avulla, jonka alkuperäiset nollakohdat olivat väliltä (-5,5) valitut satunnaiset pisteet

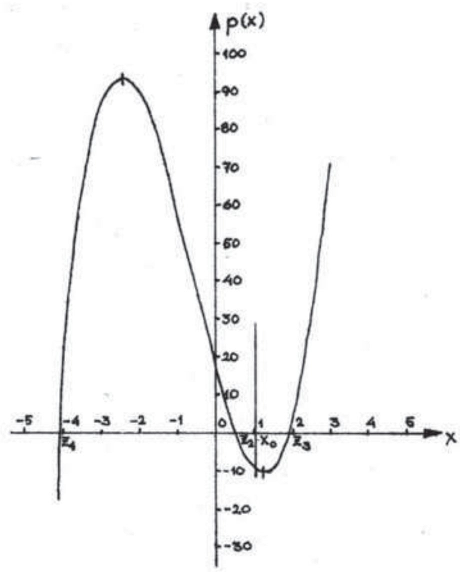
z_1 = -4.0467951

z_2 = 0.54756939

z_3 = 1.9969324

ja alkuperäinen a_0

a_0 = 3.9603187 .



kuva 8

Alkuperäiselle polynomille (kuva 8) suoritettiin yhtälön (9.32) mukainen siirto, kun M sai arvot 2^m, $m = -7,-6,\ldots,6,7$. Kullakin m:n arvolla laskettiin polynomin kertoimet a_1, a_2 ja a_3 käyttäen kaavaan (9.31) perustuvaa kertoimienlaskualgoritmia K.

<u>Algoritmi K</u> (Polynomin kertoimien lasku). Algoritmi laskee N:nnen asteen polynomin $p = a_0 x^N + a_1 x^{N-1} + \ldots + a_{N-1}x + a_N$ kertoimet a_0:n ja nollakohtien $z_1, z_2, \ldots, z_N$ avulla käyttäen apuvektoria $s_1, s_2, \ldots, s_N$.

K1. [Alkuasetukset.] $s_i \leftarrow z_i$, $i = 1,\ldots,N$, $sign \leftarrow -1$, $i \leftarrow 1$.

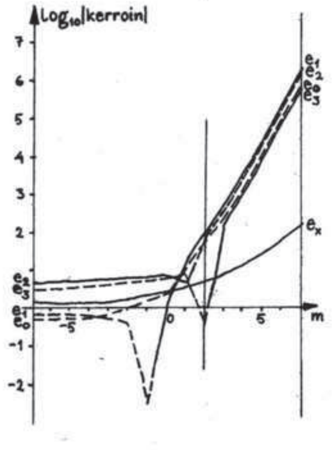
K2. [Kertoimen a_i lasku.] $a_i \leftarrow a_0 * sign * \sum_{j=1}^{N-i+1} s_j$.

K3. [Seuraava kerroin.] Jos $i = N$, algoritmi päättyy. Muuten $sign \leftarrow -sign$, $i \leftarrow i+1$, $j \leftarrow 1$.

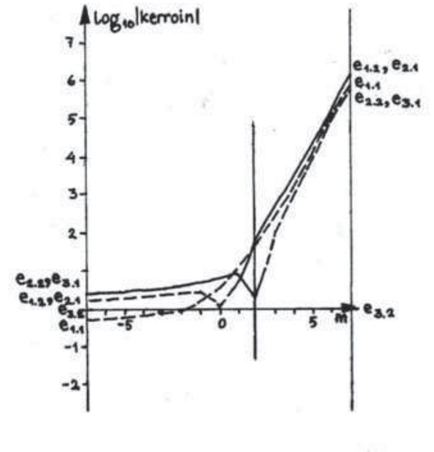
K4. [Apuvektorin täyttö.] $s_j \leftarrow z_j * \sum_{k=j+1}^{N-i+2} s_k$. Jos $j = N-i+1$, $\rightarrow$K2, muuten $j \leftarrow j+1$, $\rightarrow$K4. ▮

Saaduille polynomeille suoritettiin myös yhtälön (9.37) mukainen supistus siten, että satunnainen piste $x_0 = 1.0768227$ pysyi paikoillaan kaikilla m:n arvoilla. Tällöin k sai kullakin M:n arvolla arvon $x_0/(x_0+M)$.

Taylor-sarjan (9.19) kertoimia tutkittiin useilla eri x:n arvoilla, joille polynomin siirtyessä suoritettiin vastaavat siirrot. Suoritettu supistus ei odotusten mukaisesti vaikuttanut kertoimiin.

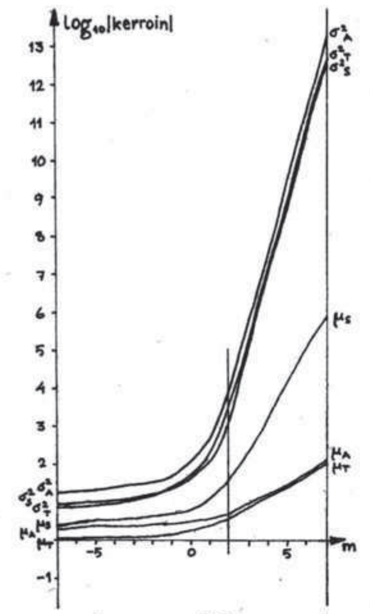


kuva 9a



kuva 9b

Kuvissa 9a ja 9b nähdään kertoimien itseisarvojen logaritmit tekijöittäin pisteessä x_0 m:n vaihdellessa -7:stä 7:ään. Nollakohtien ollessa erimerkkisiä eivät kertoimet muutu oleellisesti. Kaikki nollakohdat tulevat samanmerkkisiksi suunnil-



kuva 10

leen m:n arvolla kaksi, jolloin M = 4, ja alkavat tämän jälkeen lähestyä toisiaan suhteellisesti, jolloin kertoimien itseisarvot alkavat odotusten mukaan kasvaa voimakkaasti. Eräiden kertoimien etumerkki vaihtui tarkastelun kuluessa. Kuvissa 9 merkitsee ehyt viiva positiivista ja katkoviiva negatiivista kerrointa.

Kuvassa 10 on vastaavasti yhtälöiden (9.20) ja (9.21) määrittämien $E^{(1)}_{N,2}$:n odotusarvon ja varianssin lausekkeiden kertoimet, jotka antavat kokonaiskuvan nollakohtien tiivistymisen vaikutuksesta Hermer-Shaman kumulatiiviseen pyöristysvirheeseen.

Myös muissa tarkastelluissa pisteissä olivat tulokset vastaavanlaisia. Tosin kertoimet alkoivat kasvaa pienintä nollakohtaa pienemmillä x:n arvoilla vasta näidenkin muuttuessa lisäyksen M vaikutuksesta positiivisiksi. Tämä olikin yhtälön (9.36) mukaan edellytyksenä kertoimien kasvulle.

Vastaava koe suoritettiin myös viidennentoista asteen polynomilla, mutta laskentatarkkuus ei riittänyt luotettavien arvojen saamiseen kertoimille. Esimerkkinä mainittakoon, että nollakohtien ollessa lähekkäimmillään (m = 7) saatiin polynomin arvoksi sen eräässä nollakohdassa $4.2 \cdot 10^{28}$. Tämäkin tosin omalla tavallaan osoittaa, että pyöristysvirheet kasvavat voimakkaasti nollakohtien lähestyessä toisiaan.

10. MATRIISIN KÄÄNTÖ

Käännettäessä $m \times m$-matriisi Gauss-Jordanin menetelmällä [7] otetaan käyttöön apumatriisina $m \times m$-yksikkömatriisi. Nämä kaksi matriisia yhdessä muodostavat $m \times 2m$-matriisin

$$(A \,\vdots\, I) = \left[\begin{array}{cccc:ccc} a_{11}^{[0]} & a_{12}^{[0]} & \dots & a_{1m}^{[0]} & a_{1.m+1}^{[0]} & \dots & a_{1.2m}^{[0]} \\ a_{21}^{[0]} & a_{22}^{[0]} & \dots & a_{2m}^{[0]} & a_{2.m+1}^{[0]} & \dots & a_{2.2m}^{[0]} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ a_{m1}^{[0]} & a_{m2}^{[0]} & \dots & a_{mm}^{[0]} & a_{m.m+1}^{[0]} & \dots & a_{m.2m}^{[0]} \end{array}\right] \qquad (10.1)$$

missä $a_{ij}^{[0]} = a_{ij}$, kun $j \le m$, $a_{ij}^{[0]} = 1$, kun $i = j-m$ ja $a_{ij}^{[0]} = 0$ muulloin.

Matriisin kääntö tapahtuu algoritmin

$$\begin{cases} a_{ij}^{[k]} = a_{ij}^{[k-1]} - \dfrac{a_{ik}^{[k-1]}}{a_{kk}^{[k-1]}} a_{kj}^{[k-1]}, & i = 1,\dots,m,\ i \ne k, \\ & j = k+1,\dots,k+m, \\ & k = 1,\dots,m \\ a_{ij}^{[m+1]} = \dfrac{a_{ij}^{[m]}}{a_{ii}^{[m]}}, & i = 1,\dots,m \\ & j = m+1,\dots,2m \end{cases} \qquad (10.2)$$

avulla. Mikäli k:nnella iteraatiokierroksella $a_{ij}^{[k]}$:lle ei lasketa uutta arvoa, on $a_{ij}^{[k]} = a_{ij}^{[k-1]}$. Tällöin saamme käänteismatriisiksi

$$A^{-1} = \begin{bmatrix} a_{1.m+1}^{[m+1]} & a_{1.m+2}^{[m+1]} & \dots & a_{1.2m}^{[m+1]} \\ \vdots & \vdots & & \vdots \\ a_{m.m+1}^{[m+1]} & a_{m.m+2}^{[m+1]} & \dots & a_{m.2m}^{[m+1]} \end{bmatrix}. \qquad (10.3)$$

Alkioiden indeksointijärjestystä sopivasti vaihtamalla on mahdollista suorittaa kääntäminen tehokkaammin, so. saada tulos tarkemmaksi. Seuraavassa tarkastelussa tätä ei ole otettu huomioon.

Kun hajoitamme algoritmin (10.2) yksittäisiin laskutoimituksiin, saamme sen muotoon

$$\left\{\begin{array}{lll} q_{i.j}^{[0]} = a_{ij}^{[0]}\,, & & i=1,\dots,m,\ j=1,\dots,2m \\ q_{i.0.1}^{[k]} = a_{i.k}^{[k-1]}/a_{k.k}^{[k-1]} & & \multirow{3}{*}{$\left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \begin{array}{l} i=1,\dots,m,\ i\neq k, \\ k=1,\dots,m \end{array}$} \\ \left.\begin{array}{l} q_{i.j.2}^{[k]} = q_{i.0.1}^{[k]}\cdot a_{k.j}^{[k-1]} \\ q_{i.j.3}^{[k]} = a_{ij}^{[k]} = a_{ij}^{[k-1]}-q_{i.j.2}^{[k]} \end{array}\right\} & j=k+1,\dots,k+m & \\ q_{i.j}^{[m+1]} = a_{ij}^{[m+1]} = a_{ij}^{[m]}/a_{ii}^{[m]} & & i=1,\dots,m,\ j=m+1,\dots,2m \end{array}\right. \qquad (10.4)$$

jossa q:n indekseistä ensimmäinen osoittaa riviä, toinen saraketta ja kolmas laskutoimituksen vaihetta. Kuhunkin q:hun liittyy täsmälleen yksi yksittäinen pyöristysvirhe. Pyöristysvirheiden lukumääräksi saamme tyypeittäin

tyyppi	lukumäärä
$e_{i.j}^{[0]}$	$m\cdot 2m$
$e_{i.0.1}^{[k]}$	$(m-1)\cdot m$
$e_{i.j.2}^{[k]}$	$(m-1)\cdot m\cdot m$
$e_{i.j.3}^{[k]}$	$(m-1)\cdot m\cdot m$
$e_{i.j}^{[m+1]}$	$m\cdot m$.

Kaikkiaan pyöristysvirheitä ja samalla ensimmäisen asteen termejä kumulatiivisen pyöristysvirheen Taylor-kehitelmässä on $2m^3+2m^2-m$ kappaletta.

Taylor-kehitelmän kertoimien analyyttinen määrittäminen kullekin käänteismatriisin alkiolle on algoritmin (10.4) laajuuden johdosta varsin hankalaa, mutta kokeellisissa tarkasteluissa voidaan käyttää esimerkiksi kertoimienlaskualgoritmia L. Esimerkiksi 10×10-matriisin arvojen ollessa valmiina VA-taulukossa saamme alkion 2.7 (E,e)-sarjan taulukkoon C1 sekä alkion 7.2 (E,e)-sarjan taulukkoon C2 ja (R,e)-sarjan taulukkoon C3 ohjelmalla

```
C MÄÄRITTELYT
      REAL VA(10,10),V(2200),C1(2200)
      REAL C2(2200),C3(2200)
      INTEGER A(10,10),T(2200),O1(2200),O2(2200),Q
```

```
C ALKUARVOT
      I=LBEGIN(V,T,01,02,C3,2200)
      DO 30 I=1,10
      DO 10 J=1,10
   10 A(I,J)=LNAME(VA(I,J))
      DO 30 J=11,20
      VAL=0.
      IF(J-I-10) 30,20,30
   20 VAL=1.
   30 A(I,J)=LNAMEX(VAL)
C ITERAATIOKIERROKSET 1...10
      DO 50 K=1,10
      DO 50 I=1,10
      IF(I-K) 40,50,40
   40 Q=LDIV(A(I,K),A(K,K))
      DO 50 L=1,10
      J=K+L
      A(I,J)=LSUB(A(I,J),LMUL(Q,A(K,J)))
   50 CONTINUE
C ITERAATIOKIERROS 11
      DO 60 I=1,10
      DO 60 J=11,20
   60 A(I,J)=LDIV(A(I,J),A(I,I))
C KERTOIMIEN LASKU
      I=LEND(A(2,17))+LREL(C1,2200)
      I=LEND(A(7,12))+LREL(C2,2200)+LABS(C3,2200)
```

Yksittäisten pyöristysvirheiden lukumäärä kasvaa niin voimakkaasti matriisin ulottuvuuden kasvaessa, että algoritmi L on sovellettava käyttämään tukimuisteja, mikäli sitä aiotaan käyttää suurialla matriiseja käännettäessä.

Algoritmia L hyväksi käyttäen laskettiin (E,e)-sarjat symmetrisen 5×5-matriisin A = (a_{ij}) käänteismatriisin alkioille. Matriisin A satunnaislukualkiot olivat neljän desimaalinumeron tarkkuudella

$$A = \begin{bmatrix} 0.5758 & -0.1035 & -0.0824 & 0.1051 & -0.2077 \\ -0.1035 & 0.2601 & -0.0850 & -0.1125 & 0.2108 \\ -0.0824 & -0.0850 & 0.3935 & 0.0737 & -0.1073 \\ 0.1051 & -0.1125 & 0.0737 & 0.3049 & -0.1401 \\ -0.2077 & 0.2108 & -0.1073 & -0.1401 & 0.3595 \end{bmatrix}.$$

Käänteismatriisiksi $A^{-1} = (c_{ij})$ saatiin

$$A^{-1} = \begin{bmatrix} 2.5555 & -0.2655 & 1.0407 & -0.4111 & 1.7821 \\ -0.2655 & 7.6127 & 0.2779 & 0.9140 & -4.1777 \\ 1.0407 & 0.2779 & 3.2415 & -0.4788 & 1.2193 \\ -0.4111 & 0.9140 & -0.4788 & 4.2070 & 0.7235 \\ 1.7821 & -4.1777 & 1.2193 & 0.7235 & 6.9063 \end{bmatrix}.$$

Saatujen Taylor-kehitelmien suurimmat kertoimet olivat alkion $c_{1.2}$ sarjassa virheen $e^{[0]}_{5.2}$ (eli alkuarvon $a_{5.2}$ pyöristysvirheen) sekä alkion $c_{2.1}$ sarjassa virheen $e^{[0]}_{2.5}$ (eli alkuarvon $a_{2.5}$ pyöristysvirheen) kertoimet. Näiden molempien arvo oli 10.7718. Noin kaksi kolmasosaa kaikista kertoimista oli nollia.

Kokonaiskuvan saaminen yksittäisten kertoimien perusteella on niiden lukuisuuden vuoksi varsin vaikeaa. Tässä suhteessa saamme paremman käsityksen yhtälöiden (5.8) ja (5.9) antamien keskiarvon ja varianssin kertoimien avulla. Tällöin on syytä huomioida, että mahdollisista virhelähteistä on m^2 kappaletta tarkkoja alkuarvoja (yksikkömatriisin alkiot), $m \cdot (m-1)$ kappaletta muotoa $0-q^{[k]}_{i.j.2}$ olevia ja siten tarkkoja vähennyslaskuja sekä $m \cdot (m-1)$ kappaletta muotoa $q^{[k]}_{i,0,1} \cdot 1$ olevia tarkkoja kertolaskuja. Näitä alkuarvoja ja laskutoimituksia vastaavat yksittäiset pyöristysvirheet ovat nollia, joten niiden kertoimet, yhteensä $3m^2-2m$ kappaletta, on syytä poistaa keskiarvon ja varianssin lausekkeita laskettaessa. Tähän on varauduttu aliohjelmaryhmässä L (vastaavat TYPE-kentät ovat negatiivisia).

Esimerkkimatriisille A^{-1} saatiin odotusarvojen lausekkeiden kertoimiksi kahden desimaalin tarkkuudella, kun mainitut poistot oli suoritettu (kertoimet tekijöittäin järjestyksessä $\mu_A/\mu_S/\mu_T$)

$$\begin{bmatrix} -1.00/3.68/2.32 & -1.00/2.00/12.68 & -1.00/1.00/3.81 & -1.00/0.00/3.18 & -1.00/-1.00/3.65 \\ -1.00/2.00/12.68 & -1.00/1.54/2.95 & -1.00/0.00/-3.33 & -1.00/-1.00/2.79 & -1.00/-2.00/3.67 \\ -1.00/1.00/3.81 & -1.00/0.00/-3.33 & -1.00/-0.09/1.68 & -1.00/-2.00/2.48 & -1.00/-3.00/3.36 \\ -1.00/0.00/3.18 & -1.00/-1.00/2.79 & -1.00/-2.00/2.48 & -1.00/-2.02/1.23 & -1.00/-4.00/-2.23 \\ -1.00/-1.00/3.65 & -1.00/-2.00/3.67 & -1.00/-3.00/3.36 & -1.00/-4.00/-2.23 & -1.00/-4.00/2.93 \end{bmatrix}.$$

Vastaavat varianssien lausekkeiden kertoimet olivat tekijöittäin järjestyksessä $\sigma_A^2/\sigma_S^2/\sigma_T^2$

$$\begin{bmatrix} 2.71/3.34/1.29 & 366.14/326.72/179.08 & 5.76/6.51/2.34 & 22.41/16.71/7.95 & 14.37/11.31/4.40 \\ 366.14/326.72/196.11 & 6.19/7.53/1.95 & 207.84/264.30/95.22 & 32.69/29.60/8.00 & 18.31/18.07/4.14 \\ 5.76/6.51/2.44 & 207.84/264.30/155.25 & 1.76/4.74/1.08 & 12.75/18.96/5.93 & 16.46/20.74/7.28 \\ 22.41/16.71/13.82 & 32.69/29.60/16.66 & 12.75/18.96/6.80 & 17.02/4.62/1.05 & 56.18/41.51/20.69 \\ 14.37/11.31/7.20 & 18.31/18.07/7.17 & 16.46/20.74/7.37 & 56.18/41.51/16.57 & 8.59/8.65/2.78 \end{bmatrix}.$$

Edellä esitetyistä matriiseista voidaan tehdä mielenkiintoisia havaintoja:

- **Kaikki kertoimet σ_T^2:n kertoimia lukuunottamatta ovat symmetrisiä (so. c_{ij}:n tietty kerroin on sama kuin c_{ji}:n vastaava kerroin).**
- **σ_T^2:n kertoimet yläkolmiossa ovat pienempiä kuin symmetriset kertoimet alakolmiossa (paitsi c_{45}:n σ_T^2:n kerroin), joten muiden kertoimien yhtäsuuruudesta johtuen yläkolmioon lasketut käänteismatriisin alkiot ovat tarkempia kuin alakolmioon lasketut.**
- **Varianssin kerroinmatriisin kerrointen suuruusluokka on päälävistäjällä pienempi kuin muualla, so. päälävistäjän alkioiden arvot ovat tarkempia kuin muiden alkioiden.**
- **μ_S:n kertoimien arvo alkioissa c_{ij}, $i \neq j$, on m-i-j (m = 5) sekä alkiossa c_{mm} m-1.**
- **Kaikkien μ_A:n kertoimien arvo on -1.**

Jätän tässä esityksessä avoimeksi, ovatko nämä havainnot sattumia, vain käsiteltyyn matriisiin liittyviä, vai voidaanko vastaavia havaintoja tehdä yleisesti. Odotusarvomatriisia koskevat havainnot voidaan yleistää ainakin mielivaltaiselle 2×2-matriisille, sillä tällöin saadaan odotusarvomatriisiksi

$$\begin{bmatrix} -1/\frac{D}{a_{11}a_{22}} \ /1+2\frac{a_{12}a_{21}}{D}+\frac{a_{12}a_{21}}{a_{11}a_{22}} & -1/\ -1\ /\ 2\frac{a_{11}a_{22}}{D} \\ -1/\ -1\ /\ 2\frac{a_{11}a_{22}}{D} & -1/\ -1\ /1+2\frac{a_{12}a_{21}}{D} \end{bmatrix}$$

missä $D = a_{11}a_{22}-a_{12}a_{21}$ on ko. matriisin determinantti.

Liite: ALGORITMI L FORTRAN IV-OHJELMANA

algoritmin
askel

```
      C
L1    C     FUNCTIØN SUBPRØGRAM GRØUP L
      C
            INTEGER FUNCTIØN QI
            ENTRY LBEGIN (VALUE,TYPE,ØPER1,ØPER2,CØEFF,M)
            INTEGER TYPE,ØPER1,ØPER2,QJ,QK,QN,EX
            DIMENSIØN VALUE(M),TYPE(M),ØPER1(M),ØPER2(M)
            DIMENSIØN CØEFF(M)
            EX=0
            QI=0
            I=1
            RETURN
      C
L3    C     INITIAL VALUES
      C
            ENTRY LNAME(VAL)
            TYPE(I)=1
         10 VALUE(I)=VAL
            GØ TØ 66
            ENTRY LNAMEX(VAL)
            TYPE(I)=-1
            GØ TØ 10
      C
L4    C     NEGATIØN
      C
            ENTRY LNEG(QJ)
            TYPE(I)=-2
            VALUE(I)=-VALUE(QJ)
            GØ TØ 65
      C
L5    C     ADDITIØN
      C
            ENTRY LADD(QJ,QK)
            IF(VALUE(QJ)) 20,21,20
         20 IF(VALUE(QK)) 22,21,22
         21 EX=1
         22 TYPE(I)=3
            VALUE(I)=VALUE(QJ)+VALUE(QK)
            GØ TØ 60
            ENTRY LADDX(QJ,QK)
            GØ TØ 21
            ENTRY LADDN(QJ,QK)
            EX=-1
            GØ TØ 22
```

```
      C
L6    C       SUBTRACTION
      C
              ENTRY LSUB(QJ,QK)
              IF(VALUE(QJ)) 30,31,30
           30 IF(VALUE(QK)) 32,31,32
           31 EX=1
           32 TYPE(I)=4
              VALUE(I)=VALUE(QJ)-VALUE(QK)
              GO TO 60
              ENTRY LSUBX(QJ,QK)
              GO TO 31
              ENTRY LSUBN(QJ,QK)
              EX=-1
              GO TO 32
      C
L7    C       MULTIPLICATION
      C
              ENTRY LMUL(QJ,QK)
              IF(ABS(VALUE(QJ))-1.) 40,41,40
           40 IF(ABS(VALUE(QK))-1.) 42,41,42
           41 EX=1
           42 TYPE(I)=5
              VALUE(I)=VALUE(QJ)*VALUE(QK)
              GO TO 60
              ENTRY LMULX(QJ,QK)
              GO TO 41
              ENTRY LMULN(QJ,QK)
              EX=-1
              GO TO 42
      C
L8    C       DIVISION
      C
              ENTRY LDIV(QJ,QK)
              IF(ABS(VALUE(QK))-1.) 51,50,51
           50 EX=1
           51 TYPE(I)=6
              VALUE(I)=VALUE(QJ)/VALUE(QK)
              GO TO 60
              ENTRY LDIVX(QJ,QK)
              GO TO 50
              ENTRY LDIVN(QJ,QK)
              EX=-1
              GO TO 51
      C
L9         60 OPER2(I)=QK
              IF(VALUE(I)) 62,61,62
           61 EX=EX+1
           62 IF(EX) 64,64,63
           63 TYPE(I)=-TYPE(I)
           64 EX=0
L10        65 OPER1(I)=QJ
L11        66 QI=I
              I=I+1
              RETURN
```

```
      C
      C         COEFFICIENTS
      C
                ENTRY LEND(QN)
L12             DO 80 K=1,M
             80 COEFF(K)=0.
                COEFF(QN)=1.
                K=QN
L13          81 ITYP=IABS(TYPE(K))
                K1=OPER1(K)
                K2=OPER2(K)
L14             GO TO (83,102,103,104,105,106),ITYP
L15         102 COEFF(K1)=COEFF(K1)-COEFF(K)
                COEFF(K)=0.
                GO TO 83
L16         103 D1=1.
                D2=1.
                GO TO 82
L17         104 D1=1.
                D2=-1.
                GO TO 82
L18         105 D1=VALUE(K2)
                D2=VALUE(K1)
                GO TO 82
L19         106 D1=1./VALUE(K2)
                D2=-VALUE(K1)/VALUE(K2)**2
L20          82 COEFF(K1)=COEFF(K1)+D1*COEFF(K)
                COEFF(K2)=COEFF(K2)+D2*COEFF(K)
L21          83 K=K-1
                IF(K) 84,84,81
             84 I=1
                QI=0
                RETURN
      C
L22   C         RELATIVE OR ABSOLUTE COEFFICIENTS
      C
                ENTRY LREL(COEFIC,M)
                RES=VALUE(QN)
                ASSIGN 91 TO IG
                GO TO 90
                ENTRY LABS(COEFIC,M)
                ASSIGN 92 TO IG
             90 CONTINUE
                DIMENSION COEFIC(M)
                DO 92 K=1,QN
                COEFIC(K)=COEFF(K)*VALUE(K)
                GO TO IG,(91,92)
             91 COEFIC(K)=COEFIC(K)/RES
             92 CONTINUE
                QI=0
                RETURN
                END
```

KÄYTETTYJÄ MERKINTÖJÄ

merkintä	selitys	sivu
*	pyöristyssymboli, esim. $z^* = z(1+e)$	3
$\langle\ \rangle$	sarjan jäännöstermi, sulkeissa sarjan muuttujatyyppi ja pienin jäännöstermissä esiintyvä asteluku, esim. $\langle e^3 \rangle$	15
$\bar{\ }$	x-koordinaattien siirto: $\bar{x} = x+M$	46
$\hat{\ }$	x-koordinaattien lavennus/supistus: $\hat{x} = kx$	47
α	polynomille p $\alpha_i = x^{N-i} a_i$	40
β	polynomille p $\beta_i = \sum_{j=0}^{i} x^{N-i} a_i$	40
γ	polynomille p $\gamma_i = \sum_{j=0}^{i} (N-j) x^{N-i} a_i$	40
ε	mantissan absoluuttinen pyöristysvirhe pyöristävässä aritmetiikassa	4
ε'	mantissan absoluuttinen pyöristysvirhe katkaisevassa aritmetiikassa	5
μ	odotusarvo; μ_A:alkuarvon, μ_S:summan ja erotuksen, μ_T:tulon ja osamäärän pyöristysvirheen odotusarvo	8,9,11, 12,16,18
σ	keskihajonta; σ^2:varianssi; σ_A^2:alkuarvon, σ_S^2:summan ja erotuksen, σ_T^2:tulon ja osamäärän pyör.virheen varianssi	8,9,11, 12,16,18
ξ,ζ,η	satunnaismuuttujia	7,8,16,45
a	polynomin p kerroin;matriisin A alkio;	39,52
	(E,e)-sarjan kerroin	15
b	käytetyn aritmetiikan kantaluku	2
c	(R,e)-sarjan kerroin;matriisin A^{-1}alkio	20,55
D	Q:n derivaatta operandinsa suhteen, esim. $D = \partial Q_n(q_i, q_j)/\partial q_i$;	20
	keskihajonta; D^2:varianssi, esim. $D^2(e)$	8
d	(R,r)-sarjan kerroin	22
E	kumulatiivinen suhteellinen pyör.virhe;	15
	odotusarvo, esim. E(e)	8
$E_n^{(i)}$	(E,e)-sarjan i:nnen asteen termien summa	15

(E,e)	(E,e)-sarja: muotoa $E_n = \sum_i a_i e_i + \ldots$ oleva Taylor-kehitelmä	15
e	yksittäinen suhteellinen pyör.virhe, esim. $z^* = z(1+e)$	3
f	tiheysfunktio	5
i,j,k,l,m,n,N	indeksejä	
k	x-koordinaattien lavennus/supistuskerroin, vrt. "^"	47
M	x-koordinaattien siirron määrä, vrt. "−"	46
m	liukuluvun mantissa	2
p	liukuluvun eksponentti;	2
	polynomi $p = p(x) = a_0 x^N + a_1 x^{N-1} + \ldots + a_N$	39
Q	laskutoimitus, esim. voi olla $Q(a,b) = a-b$	19
q	algoritmin alkuarvo tai tulos	19
R	kumulatiivinen absoluuttinen pyör.virhe	19
$R_n^{(i)}$	(R,e)-sarjan i:nnen asteen termien summa	20
(R,e)	(R,e)-sarja: muotoa $R_n = \sum_i c_i e_i + \ldots$ oleva Taylor-kehitelmä	20
(R,r)	(R,r)-sarja: muotoa $R = \sum_i d_i r_i + \ldots$ oleva Taylor-kehitelmä	22
r	yksittäinen absoluuttinen pyör.virhe	3
t	liukuluvun mantissan numeroiden lkm.	2
u	b^{-t}	4
x	piste, jossa polynomin p arvo lasketaan	39
z	liukuluku, $z = m \cdot b^p$;	2
	polynomin nollakohta	46

VIITELUETTELO

[1] Babuška, I., Numerical Stability in Mathematical Analysis, Information Processing 68, 11-23, Amsterdam, 1969

[2] Elfving, G., Todennäköisyyslaskenta, II luku, Otava, Helsinki, 1966

[3] Hamming, Numerical Methods for Scientifists and Engineers, luku 2, McGraw-Hill, New York, 1962

[4] Henrici, P., Elements of Numerical Analysis, luvut 15 ja 16, Wiley, New York, 1964

[5] Henrici, P., Discrete Variable Methods in Ordinary Differential Equations, Wiley, New York, 1962

[6] Hull, T. E. ja Swenson, J. R., Tests of Probabilistic Models for Propagation of Roundoff Errors, Comm. ACM, vol.9, 108-113, 1966

[7] Isaacsson, E. ja Keller, H. B., Analysis of Numerical Methods, luvut 1 ja 2, Wiley, New York, 1966

[8] Knuth, D. E., The Art of Computer Programming, vol.2, luku 4, Addison-Wesley, New York, 1969

[9] Tienari, M., A Statistical Model of Roundoff Errors in Varying Lenght Floating-point Arithmetic, Helsinki, 1970

[10] Wilkinson, J., Rounding Errors in Algebraic Processes, Prentice Hall, Englewood Cliffs, N.J., 1963