Numeriikan perusteita

Lukujen esitystavoista

Yksinkertaiset muuttujat

Pienin suoraan osoitettavissa oleva yksikkö yleensä 8 bitin tavu (poikkeuksena sanakoneet, kuten Cray).

Kokonaisluku (integer) tavallisesti 2 tavua. Kokonaisluvun itseisarvo voi olla korkeintaan $2^{15} - 1 = 32\ 767$.

Usein käytettävissä myös kaksoistarkkuuden kokonaisluku, 4 tavua, jolloin suurin esitettävissä oleva luku on $2^{31} - 1 = 2\ 147\ 483\ 647$.

Reaaliluku (real) tavallisesti 4 tavua

- mantissan etumerkki
- mantissa normitettu välille 0.1 – 1
- eksponentti: etumerkki+itseisarvo tai 2:n komplementti

Esim. PC:n yksinkertaisen tarkkuuden reaaliluvun mantissa on 23 bittiä. Desimaalipisteen jälkeinen bitti on aina ykkönen, joten sitä ei tarvitse tallettaa. Tarkkuus on 24 bittiä eli $\log_{10} 2^{24} \approx 7$ desimaalia. Eksponenttiosan pituus 8 bittiä. Eksponenttiosa e = 127 (bias) + todellinen eksponentti; $1 \leq e \leq 254$. Lukualue $2^{-126} \approx 10^{-38}$ – $2^{128} \approx 3 \times 10^{38}$.

Kaksoistarkkuus: merkitsevien numeroiden määrä noin kaksinkertainen ja mahdollisesti myös lukualue laajempi. Esimerkiksi PC:n kaksoistarkkuuden muuttujien tarkkuus on noin 15 desimaalia ja lukualue noin $10^{-308} \ldots 10^{308}$.

Äärellinen laskentatarkkuus

Konevakiot ovat laskentatarkkuutta ja liukulukujen esitystapaa kuvaavia lukuja.

"Kone-epsilon" ϵ on pienin luku, joka ykköseen lisättynä antaa ykköstä suuremman tuloksen:

$$\epsilon = \min\{x | 1 + x > 1\}.$$

PC:n kaksoistarkkuuden lukuja vastaava konevakio on $\epsilon = 2.2 \times 10^{-16}$. Huom! tämä on hyvin paljon suurempi kuin pienin esitettävissä oleva positiivinen luku.

Virheet

Jos luvun tarkka arvo on a, sen likimääräisen esityksen $\tilde{a}$ absoluuttinen virhe on

$$\Delta a = \tilde{a} - a.$$

Suhteellinen virhe on

$$e = \frac{\Delta a}{a} = \frac{\tilde{a} - a}{a}.$$

Usein a ei ole tiedossa, mutta Δa tunnetaan esimerkiksi tilastollisten ominaisuuksien avulla. Arvio suhteelliselle virheelle on silloin

$$e \approx \frac{\Delta a}{\tilde{a}}.$$

Pyöristys:

```
1.490      1, 1.5
1.551      2, 1.6
1.500      2
2.500      2
```

Katkaisussa luvun loppuosa heitetään menemään.

Fortranissa muunnos reaaliluvusta kokonaisluvuksi tehdään katkaisemalla.

Virheiden käyttäytyminen laskutoimituksissa

Yhteenlasku

Yhteenlaskussa lasketaan yhteen myös lukujen virheet. Jos virheiden merkit ovat satunnaisia, virheet osittain kumoavat toisensa.

$$1.57 + 0.76 = 2.33.$$

Lasketaan summa pyöristämällä luvut yhteen desimaaliin:

$$1.6 + 0.8 = 2.4$$

Termien suhteelliset virheet ovat

$$\frac{1.6 - 1.57}{1.57} = 0.019, \quad \frac{0.8 - 0.76}{0.76} = 0.053.$$

Summan suhteellinen virhe on

$$\frac{2.4 - 2.33}{2.33} = 0.030,$$

Summan suhteellinen virhe ei koskaan voi olla suurempi kuin suurin yksittäisten positiivisten termien suhteellisista virheistä.

Muutetaan esimerkkiä hieman:

$$1.57 + 0.74 = 2.31.$$

$$1.6 + 0.7 = 2.3$$

Termien suhteelliset virheet ovat

$$\frac{1.6 - 1.57}{1.57} = 0.019, \quad \frac{0.7 - 0.74}{0.74} = -0.054.$$

Summan suhteellinen virhe on

$$\frac{2.3 - 2.31}{2.31} = -0.004.$$

Yhteenlasku on virheitä tasoittava operaatio, mikäli yksittäisten termien virheiden merkit ovat satunnaisia.

Ongelmia samanmerkkisten lukujen yhteenlasku voi aiheuttaa, kun lasketaan yhteen hyvin eri suuruisia lukuja. Jos esitystarkkuus on 7 desimaalia, on $1.0 + 3 \times 10^{-8} = 1.0$.

Vähennyslasku

$\pi_1 = 3.160494$ ja $\pi_2 = 3.142857$. Esitetään luvut kolmella desimaalilla, ja vähennetään toisistaan:

	tarkka arvo	likiarvo	abs. virhe	suht. virhe
π_1	3.160494	3.160	-0.00049	-0.00016
π_2	3.142857	3.143	0.00014	0.00005
$\pi_1 - \pi_2$	0.017637	0.017	-0.00064	-0.03612

Erotuksen suhteellinen virhe on paljon suurempi kuin alkuperäiset virheet, sillä mantissojen merkitsevimmät numerot kumoavat toisensa ja jäljelle jää alkuperäistä vähemmän merkitseviä numeroita (catastrophic cancellation).

Varo lähes yhtäsuurten lukujen vähennyslaskua!

Ongelmia aiheuttaa esimerkiksi

$$\sqrt{1+x} - 1,$$

kun x on pieni. Mikäli $x << 1$, voidaan neliöjuuri korvata Taylor-sarjan alkupäällä:

$$\sqrt{1+x} - 1 \approx 1 + \frac{1}{2}x - 1 = \frac{1}{2}x.$$

Lauseke voidaan myös muuntaa toiseen muotoon

$$\sqrt{1+x} - 1 = \frac{x}{1+\sqrt{1+x}}.$$

Esimerkki: lasketaan integraali

$$I_n = \int_0^1 \frac{x^n}{x+10} dx$$

$$\begin{aligned} I_{n+1} + 10I_n &= \int_0^1 \left(\frac{x^{n+1}}{x+10} + \frac{10x^n}{x+10} \right) dx \\ &= \int_0^1 \frac{x^n(x+10)}{x+10}\, dx \\ &= \int_0^1 x^n\, dx = \frac{1}{n+1}, \end{aligned}$$

josta saadaan palautuskaava

$$I_{n+1} = \frac{1}{n+1} - 10I_n.$$

$$I_0 = \int_0^1 \frac{dx}{x+10} = \Big|_0^1 \ln(x+10) = \ln 11 - \ln 10 \approx 0.0953.$$

$$\begin{aligned} I_1 =& \frac{1}{1} - 10 \times 0.0953 \approx 0.0470, \\ I_2 =& \frac{1}{2} - 10 \times 0.0470 \approx 0.0300, \\ I_3 =& \frac{1}{3} - 10 \times 0.0300 \approx 0.0333, \\ I_4 =& \frac{1}{4} - 10 \times 0.0333 \approx -0.0833.???? \end{aligned}$$

Integroitava on positiivinen, joten intgegraalinkin pitäisi olla positiivinen. Mistä negatiivinen tulos on peräisin?

Nimittäjä on likimain vakio 10, joten

$$I_n \approx \frac{1}{10} \int_0^1 x^n\, dx = \frac{1}{10} \frac{1}{n+1}.$$

Tässäkin on kyseessä lähes yhtä suurten lukujen vähennyslasku.