Representasjon av tall på datamaskin.

64 bits flyttall

Representere et reelt tall a i formen

$$a = b \times 2^n, \qquad \frac{1}{2} \leq |b| < 1$$

b er signifikanten

n er eksponenten.

Deler bits: 53 på b, 11 på n.

Begge har fortegn, 52 til $|b|$, 10 til $|n|$

$|b|$ har opptil 52 signifikante binære siffre

Konverter: $2^{52} = 10^{?}$ Svaret $\approx 10^{16}$

Hvorfor: $2^{10} = 1024 \approx 1000 = 10^3$

$\Rightarrow 2^{52} \approx 10^{16}$

Da har a cirka 16 signifikante desimale siffre

Eksponenten: 10 bits gir 1024

$\Rightarrow$ største tallet er $\approx 2^{1023} \approx 10^{308}$

Måling av feil

Anta at $\tilde{a}$ er en tilnærming til a

For eksempel, $\tilde{a}$ kan være flyttallsrepresentasjon av a.

Vi har to måter å måle feilen på:

Absolutt feil $\tilde{a} - a$

Relativ feil $\frac{\tilde{a}-a}{a}$, så lenge $a \neq 0$

Noen ganger er vi ikke interessert i fortegnet, så kan sier absolutt feilen er $|\tilde{a} - a|$

rel. " er $\left|\frac{\tilde{a}-a}{a}\right|$.

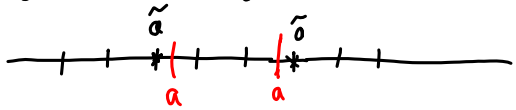
Analogi: Du har penger i en bankkonto.
Og du får 100 kr. i rente. Er det bra?
Det avhenger av hvor mye du startet med.

Eks. Du startet med 1000 kr.
Da er avkastningen $\frac{100}{1000} = 0{,}1 = 10\%$

Eks. Du startet med 200 kr.
Da er avkastningen $\frac{100}{200} = 0{,}5 = 50\%$

Moralen: mål rente i %, ikke i kroner.

Anta nå at $\tilde{a}$ er flyttalls representasjonen av a i 64 bits



Da kan vise $\left|\frac{\tilde{a}-a}{a}\right| \approx 10^{-16}$

Hvorfor? $a = \alpha \times 10^n$, $\tilde{a} = \tilde{\alpha} \times 10^n$,

normal form. $\frac{1}{10} \leq \alpha, \tilde{\alpha} < 1$.

$$\Rightarrow \left|\frac{\tilde{a}-a}{a}\right| = \left|\frac{(\tilde{\alpha}-\alpha)\times 10^n}{\alpha \times 10^n}\right| = \left|\frac{\tilde{\alpha}-\alpha}{\alpha}\right|$$

$$|\tilde{\alpha} - \alpha| \leq 0.5 \times 10^{-16}, \quad |\alpha| \geq 0.1$$

$$\Rightarrow \frac{|\tilde{\alpha} - \alpha|}{|\alpha|} \leq 5 \times 10^{-16}$$

Tap av presisjon i operasjoner (flops)

$a \in \mathbb{R}$, $\tilde{a}$ er tilnærmingen i 64 bits flyttall.

La $\delta = \frac{\tilde{a} - a}{a}$ være relativ feil.

Omskrive: $a\delta = \tilde{a} - a \Rightarrow \tilde{a} = a(1+\delta)$,

hvor $|\delta| \leq 10^{-16}$

Multiplikasjon Gitt $\tilde{a} = a(1+\delta_1)$, $\tilde{b} = b(1+\delta_2)$

hva er den relative feilen i ab?

Vi regner $\tilde{a}\tilde{b} = ab(1+\delta_1)(1+\delta_2)$

$\approx ab(1+\delta_1+\delta_2)$

fordi $\delta_1\delta_2$ veldig liten $(\leq 10^{-32})$

Rel. feilen er da $\frac{\tilde{a}\tilde{b} - ab}{ab} \approx \delta_1 + \delta_2$

Da er $\left|\frac{\tilde{a}\tilde{b} - ab}{ab}\right| \leq |\delta_1| + |\delta_2|$

Feilen er ikke verre enn 2 ganger større.

Divisjon $\frac{\tilde{a}}{\tilde{b}} \approx \frac{a}{b}(1+\delta_1-\delta_2)$

$\Rightarrow$ rel. feil. $\left|\frac{\frac{\tilde{a}}{\tilde{b}} - \frac{a}{b}}{\left(\frac{a}{b}\right)}\right| \leq |\delta_1| + |\delta_2|$

Igjen, lite feil.

Potenser? a^n?

$(\tilde{a})^n = a^n(1+\delta)^n = a^n(1+n\delta+\cdots)$

$\approx a^n(1+n\delta)$

$\Rightarrow \left|\frac{(\tilde{a})^n - a^n}{a^n}\right| \approx n|\delta|$

Igjen, relativt bra!

Addisjon : $\tilde{a} = a(1+\delta_1)$, $\tilde{b} = b(1+\delta_2)$

$$\tilde{a}+\tilde{b} = a+b+a\delta_1+b\delta_2$$
$$= a+b+\varepsilon_1+\varepsilon_2,$$

$(\varepsilon_1 = \tilde{a}-a,\ \varepsilon_2 = \tilde{b}-b)$

Rel. feil: $$\frac{(\tilde{a}+\tilde{b})-(a+b)}{a+b} = \frac{\varepsilon_1+\varepsilon_2}{a+b}$$

$$\Rightarrow \left|\frac{(\tilde{a}+\tilde{b})-(a+b)}{a+b}\right| \leq \frac{|\varepsilon_1|+|\varepsilon_2|}{|a+b|}$$

Denne feilen kan være stor hvis $a+b \approx 0$

Eks : $a = 1.429$, $b = -1.420$

Samme problem med subtraksjon : $a-b$

rel. feil stor hvis $a-b \approx 0$

Omskrivning av formler

Noen ganger kan vi unngå store feil.

Eksempel: evaluer $\dfrac{1}{\sqrt{x^2+1} - x}$

for stor x, for eksempel, $x = 10^8$

Omskrive: et triks:

$$\frac{1}{\sqrt{x^2+1} - x}\left(\frac{\sqrt{x^2+1} + x}{\sqrt{x^2+1} + x}\right)$$

$$= \frac{\sqrt{x^2+1} + x}{1} = \sqrt{x^2+1} + x$$

Eks 2. Finne røttene:

$$ax^2 + bx + c = 0$$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Anta at $b > 0$, og stor, for eks. $b = 10^8$

og at $a, c \simeq 1$.

Feil når $x = \dfrac{-b + \sqrt{b^2-4ac}}{2a}$

Kan fikse:

$$\frac{-b + \sqrt{b^2-4ac}}{2a} \quad \frac{b + \sqrt{b^2-4ac}}{b + \sqrt{b^2-4ac}} = \frac{-2c}{b + \sqrt{b^2-4ac}}.$$

Neste tema: differensligninger.

Følger : $x_1, x_2, x_3, \ldots$

Fibonacci tall : $1, 1, 2, 3, 5, 8, 13, 21, \ldots$

$x_1 = 1, \quad x_2 = 1$: initialbetingelser

$$\boxed{x_{n+2} = x_{n+1} + x_n}, \qquad n = 1, 2, 3, \ldots$$

Eksempel av en differensligning.

Skriver ligninger i formen

$$x_{n+2} - x_{n+1} - x_n = 0$$

Generelt: $x_{n+2} - a x_{n+1} - b x_n = 0$