Fra sist:

Observasjon: Hvis $\tilde{a}$ er nærmeste (64 bits) flyttall til a, så er $\left|\frac{\tilde{a}-a}{a}\right| \lesssim 10^{-16}$.

Hvorfor? La

$$a = \alpha \cdot 10^n$$
$$\tilde{a} = \tilde{\alpha} \cdot 10^n$$

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

normalformer

Vi vet at vi har $\lesssim 16$ signifikante 10-talls sifre som stemmer overens i a og $\tilde{a}$, slik at:

$$|\tilde{\alpha} - \alpha| \lesssim 0.5 \cdot 10^{-16}$$



$$\Rightarrow \left|\frac{\tilde{a}-a}{a}\right| = \left|\frac{\tilde{\alpha}10^n - \alpha 10^n}{\alpha 10^n}\right| = \left|\frac{\tilde{\alpha}-\alpha}{\alpha}\right| \leq \frac{0.5 \cdot 10^{-16}}{\frac{1}{10}} = 5 \cdot 10^{-16}$$

$\alpha \geq \frac{1}{10}$

Dette resultatet forbinder relativ feil med antall sifre som stemmer overens. På samme måte:

Observasjon 5.20

$$\left|\frac{\tilde{a}-a}{a}\right| \lesssim 10^{-m}$$

$\Updownarrow$

"omtrent" de første m signifikante sifrene i a og $\tilde{a}$ stemmer overens.

(Viste dette over for $m = 16$).

Altså: sto relativ feil $\Leftrightarrow$ liten m $\Leftrightarrow$ få sifre stemmer overens.

Feil ved regneoperasjoner

La $a \in \mathbb{R}$, $\tilde{a}$ en tilnærming til a med 64 bits flyttall.

relativ feil: $\delta = \frac{\tilde{a}-a}{a}$ ($\lesssim 10^{-16}$ hvis $\tilde{a}$ er nærmeste flyttall)

$\Downarrow$

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

Multiplikasjon

$\tilde{a}, \tilde{b}$ nærmeste flyttall.

$\tilde{a} = a(1+\delta_1)$ $\quad \tilde{b} = b(1+\delta_2)$ $\quad \delta_1, \delta_2 \approx 10^{-16}$

Hva er relativ feil mellom ab og $\tilde{a}\tilde{b}$

$$\tilde{a}\tilde{b} = a(1+\delta_1)\,b(1+\delta_2) = ab(1+\delta_1+\delta_2+\underbrace{\delta_1\delta_2}_{\approx 10^{-32},\ \text{lite i forhold til } \delta_1,\ \delta_2}) \approx ab(1+\delta_1+\delta_2)$$

$$\left|\frac{\tilde{a}\tilde{b}-ab}{ab}\right| \approx \left|\frac{ab(1+\delta_1+\delta_2)-ab}{ab}\right| = |(1+\delta_1+\delta_2)-1| = |\delta_1+\delta_2| \le |\delta_1|+|\delta_2|$$

Relativ feil blir med andre ord ikke mer enn fordoblet.

Divisjon

$$\frac{\tilde{a}}{\tilde{b}} = \frac{a(1+\delta_1)}{b(1+\delta_2)} = \frac{a}{b}\frac{(1+\delta_1)(1-\delta_2)}{(1+\delta_2)(1-\delta_2)} = \frac{a}{b}\frac{1+\delta_1-\delta_2-\delta_1\delta_2}{1-\delta_2^2}$$

liten, liten $\approx 10^{-32}$

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

Relativ feil:

$$\left|\frac{\frac{\tilde{a}}{\tilde{b}}-\frac{a}{b}}{\frac{a}{b}}\right| \approx \left|\frac{\frac{a}{b}(1+\delta_1-\delta_2)-\frac{a}{b}}{\frac{a}{b}}\right| = |(1+\delta_1-\delta_2)-1| = |\delta_1-\delta_2| \le |\delta_1|+|\delta_2|$$

Igjen: Feilen blir ikke mer enn fordoblet.

Addisjon:

$\tilde{a} = a(1+\delta_1)$ $\quad \tilde{b} = b(1+\delta_2)$.

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

$$(\tilde{a}+\tilde{b})-(a+b) = a\delta_1+b\delta_2$$

relativ feil:

$$\left|\frac{(\tilde{a}+\tilde{b})-a+b}{a+b}\right| = \left|\frac{a\delta_1+b\delta_2}{a+b}\right| \le \frac{|a\delta_1|+|b\delta_2|}{|a+b|}$$

Kan bli stor hvis $a+b$ er liten sammenlignet med a og b.

Det samme gjelder for subtraksjon.

Problem subtraksjon/addisjon:

Når vi regner ut $a-b$, så kan vi miste mange signifante sifre hvis $a-b \approx 0$. Dette kalles kansellering.

Omskriving av formler for å unngå kansellering:

Eksempel 1: Vi skal regne ut $\dfrac{1}{\sqrt{x^2+1}-x}$, med x stor, $x>0$.

Når x er stor, og >0, så vil $\sqrt{x^2+1} \approx x$, og vi får problemer med kansellering. Vi skriver om:

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

I dette uttrykket får vi ikke problemer med kansellering.

Eksempel 2 Røttene til $ax^2+bx+c=0$ er $x = \dfrac{-b \pm \sqrt{b^2-4ac}}{2a}$

Hvis $b>0$, og stor sammenlignet med a og c, så får vi problemer med kansellering i $x = \dfrac{-b+\sqrt{b^2-4ac}}{2a}$

Vi kan her skrive om:

$$\frac{-b+\sqrt{b^2-4ac}}{2a} = \frac{(-b+\sqrt{b^2-4ac})(-b-\sqrt{b^2-4ac})}{2a(-b-\sqrt{b^2-4ac})}$$

$$= \frac{b^2-(b^2-4ac)}{2a(-b-\sqrt{b^2-4ac})}$$

$$= \frac{4ac}{2a(-b-\sqrt{b^2-4ac})}$$

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

Denne har ikke problemet med kansellering for $b>0$, b stor.

Oppgaver kap. 5

5.2: 1, 2, 4, 6 a de, 9
5.3: 1, 2, 3, 4, 6
5.4: 1 b, 2

Oppgave 5.2.3

d) Runde av til 3 sifre: $10.473 \longrightarrow 10.5$

e) Trunkere til 3 sifre: $10.473 \longrightarrow 10.4$

Oppgave 5.2.6 Regn ut $9.834 + 2.45$ når vi bruker 4 sifre for signifikanden.

Løsning:

$$9.834 = 0.9834 \cdot 10^{1}$$
$$2.45 = 0.2450 \cdot 10^{1} \leftarrow \text{ingen avrunding enda.}$$

$$\begin{array}{r} {\scriptstyle 1\ \ 1\ \ \ \ } \\ 0.9834 \\ +\ 0.2450 \\ \hline = 1.2284 \end{array}$$

$$1.2284 \cdot 10^{1} = 0.12284 \cdot 10^{2} \approx 0.1228 \cdot 10^{2} = 12.28$$

egentlig svar: 12.284.

Differensligninger, kap. 4 i kalkulus

En differensligning er en ligning på formen

$$x_{n+2} + b x_{n+1} + c x_n = 0,$$

der $x_1, x_2, x_3, \ldots$ er de ukjente (alternativt $x_0, x_1, x_2, \ldots$)

Hvis vi vet x_1, x_2, så kan vi finne $x_3, x_4, x_5, \ldots$ ved

$$x_{n+2} = -b x_{n+1} - c x_n \qquad n \geq 1$$

Vi sier at $x_{n+2} + b x_{n+1} + c x_n = 0$ er

- homogen (står 0 på høyre side)
- av andre orden (tre ledd involvert)
- lineær (ingen kryssledd av x_i'ene)
- konstante koeffisienter.

Eksempel 1: Vi antar harer formerer seg slik:

1. starter å formere seg når 2 måneder gamle
2. får et nytt par hver måned
3. Aldri dør.

Hvordan vil hamrebestanden vokse?

Løsning: La x_n være antall par etter n måneder. Da er

$$x_{n+2} = \#\underbrace{\text{mer enn 1 mnd gamle}}_{x_{n+1}} + \#\underbrace{\text{nyfødte}}_{x_n} \quad \Rightarrow x_{n+2} = x_{n+1} + x_n$$

($\geq$ 2 mnd. gamle.

I boka: $x_1 = x_2 = 1$. Disse kalles for startbetingelser.