Computerstøttet beregning

Lektion 2. Repetition

Martin Qvist

qvist@math.aau.dk

Det Ingeniør-, Natur-, og Sundhedsvidenskabelige Basisår

Aalborg Universitet

10. februar 2009

people.math.aau.dk/~qvist/teaching/csb-09

Flydende tal

Normaliseret flydende tal $x = \pm f \times \beta^E = \sum_{k=0}^{N} d_k \beta^{E-k}$, hvor

$\pm$: Fortegn

f: Mantissa

$f = d_0, d_1 d_2 \cdots d_N;\ 1 \leq d_0 < \beta,\ 0 \leq d_k < \beta$

β: Grundtal/base

E: Eksponent.

Bemærk:

- Typisk er $\beta = 2$ eller $\beta = 10$.
- Antal betydende cifre er $N + 1$.
- Væsentligt for flydende tals aritmetik er afrundingsmetoden. Enten symmetrisk, afskæring (chopping), op mod ∞ eller ned mod $-\infty$.

Mål for fejl

Hvis tallet x approksimeres med tallet $\hat{x}$:

Absolut fejl: $|x - \hat{x}|$

Relativ fejl: $\dfrac{|x - \hat{x}|}{|x|}$ eller $\dfrac{|x - \hat{x}|}{|\hat{x}|}$

Mål for fejl

Hvis tallet x approksimeres med tallet $\hat{x}$:

Absolut fejl: $|x - \hat{x}|$

Relativ fejl: $\frac{|x - \hat{x}|}{|x|}$ eller $\frac{|x - \hat{x}|}{|\hat{x}|}$

Fortolkning:

- Absolut fejl angiver på hvilken decimalplads $\hat{x}$ afviger fra x
- Relativ fejl angiver (omtrent) på hvormange cifre $\hat{x}$ og x stemmer overens

Mål for fejl

Hvis tallet x approksimeres med tallet $\hat{x}$:

Absolut fejl: $|x - \hat{x}|$

Relativ fejl: $\frac{|x - \hat{x}|}{|x|}$ eller $\frac{|x - \hat{x}|}{|\hat{x}|}$

Fortolkning:

- Absolut fejl angiver på hvilken decimalplads $\hat{x}$ afviger fra x
- Relativ fejl angiver (omtrent) på hvormange cifre $\hat{x}$ og x stemmer overens

Eksempel:

$$x = 1000 \qquad \hat{x} = 1000.1$$

Afrundingsfejl

Vurdering af fejl ved repræsentation af reelt tal x som binært tal (fremkommet ved afskæring):

$$\text{Eksakt: } x = +f \times 2^E = \sum_{k=0}^{\infty} d_k 2^{E-k}, \;\; f = 1, d_1 d_2 d_3 \ldots$$

$$\text{Approksimation: } \hat{x} = +\hat{f} \times 2^E = \sum_{k=0}^{N} d_k 2^{E-k}, \;\; \hat{f} = 1, d_1 d_2 d_3 \ldots d_N.$$

Afrundingsfejl

Vurdering af fejl ved repræsentation af reelt tal x som binært tal (fremkommet ved afskæring):

$$\text{Eksakt: } x = +f \times 2^E = \sum_{k=0}^{\infty} d_k 2^{E-k}, \;\; f = 1, d_1 d_2 d_3 \ldots$$

$$\text{Approksimation: } \hat{x} = +\hat{f} \times 2^E = \sum_{k=0}^{N} d_k 2^{E-k}, \;\; \hat{f} = 1, d_1 d_2 d_3 \ldots d_N.$$

$$\text{Absolut fejl: } |x - \hat{x}| = \sum_{k=N+1}^{\infty} d_k 2^{E-k} \leq 2^{E-N},$$

$$\text{Relativ fejl: } \frac{|x - \hat{x}|}{|x|} \leq 2^{-N}.$$

Afrundingsfejl

Vurdering af fejl ved repræsentation af reelt tal x som binært tal (fremkommet ved afskæring):

$$\text{Eksakt: } x = +f \times 2^E = \sum_{k=0}^{\infty} d_k 2^{E-k}, \;\; f = 1, d_1 d_2 d_3 \ldots$$

$$\text{Approksimation: } \hat{x} = +\hat{f} \times 2^E = \sum_{k=0}^{N} d_k 2^{E-k}, \;\; \hat{f} = 1, d_1 d_2 d_3 \ldots d_N.$$

$$\text{Absolut fejl: } |x - \hat{x}| = \sum_{k=N+1}^{\infty} d_k 2^{E-k} \leq 2^{E-N},$$

$$\text{Relativ fejl: } \frac{|x - \hat{x}|}{|x|} \leq 2^{-N}.$$

Tilsvarende beregning kan laves med ethvert andet grundtal

Fixed point

- I modsætning til flydende tal (floating point), hvor kommaet flyttes ved at gange med grundtal opløftet i eksponent opererer nogle arkitekturer med fixed point.
- Fixed point svarer til floating point, hvor eksponenten er fastsat på forhånd. Antal cifre før og efter kommaet ligger således fast.
- Fixed point er mindre fleksibel end floating point; til gengæld er fixed point mere effektiv.

Flydende tal i Maple

Maple regner som udgangspunkt eksakt (heltal, heltalsbrøker)
Software floats fås med kommandoen evalf

- Antal betydende cifre (svarende til $N+1$) styres med systemvariablen Digits (default er 10)

Afrunding styres med Rounding, som kan antage værdierne 0, infinity, -infinity, nearest.
Hardware floats fås med kommandoen evalhf

- Basen er typisk 2 (afhængig af platformen) og $N = 52$
- Antal betydende cifre er fast. Et estimat kan fås med kommandoen evalhf(Digits)