N:1 Felanalys och datoraritmetik

Håkon Hoel

11 april 2018

Översikt

Approximationsfel och felfortplantning

Approximationsfel

Av olika skäl behöver vi studera approximationsfel.

- Mätfel och avrundingsfel (datorer representerar närapå alla tal approximativt) ...
- T ex alla tal i intervallet $(1-\epsilon/4, 1+\epsilon/2)$ var $\epsilon = 2.2 * 10^{-16}$ sparas som (64-bit flyttalet) 1 i Matlab.
- **Notation:** För en skalär $x \in \mathbb{R}$ låter vi $\hat{x}$ beteckna en approximation av x. T ex:
$$\hat{x} = 3.14 \approx 3.14159265\ldots = x.$$
- Ofta måste man använda en approximativ algoritm för att lösa ett problem (även i fall där det finns en exakt algoritm).
- Exempel: Exakt algoritm $f(x) = e^x$ och approximativ alg.
$$\hat{f}(x) := 1 + x + x^2/2 \approx e^x.$$

Absoluta och relativa felet

- Två sätt att beskriva approximationsfelet:
  - **Absoluta felet:** $\delta x := \hat{x} - x$.
  - **Relativa felet** är definierad som
$$\frac{\delta x}{x} = \frac{\hat{x} - x}{x} = \frac{\textbf{absoluta felet}}{\textbf{exakta värdet}}.$$
- Exempel: Approximationen $\hat{x} = 3.14$ av $x = 3.14159265\ldots$ ger
$$\delta x := -0.00159265\ldots \approx -0.15927 \cdot 10^{-2} \quad \text{och} \quad \frac{\delta x}{x} \approx -0.508 \cdot 10^{-3}. \tag{1}$$
- Om man inte känner exakta värdet till x (och $x, \hat{x} \neq 0$) kan relativa felet approximeras
$$\frac{\delta x}{x} \approx \frac{\delta x}{\hat{x}}.$$

Felgränser och noggrannhet

- Ofta behövs endast konservativa skattningar av approximationsfelets storlek. Konstanter $c_1, c_2 > 0$ är **felgränser** för resp fel om följande gäller
$$|\delta x| \leq c_1, \qquad \left|\frac{\delta x}{x}\right| \leq c_2.$$
- I exemplet ekv. (1) är $c_1 = 0.2 \cdot 10^{-2}$ $c_2 = 0.6 \cdot 10^{-3}$ möjliga felgränser sedan
$$|\delta x| \approx 0.15927 \cdot 10^{-2} \leq 0.2 \cdot 10^{-2} \quad \text{och} \quad \left|\frac{\delta x}{x}\right| \approx 0.508 \cdot 10^{-3} \leq 0.6 \cdot 10^{-3}.$$
- Om det absoluta felet uppfyller $|\delta x| \leq 0.5 \cdot 10^{-n}$ säger vi att approximationen $\hat{x}$ har (minst) n **korrekta decimaler** (Obs! n kan vara negativ.)
- (Obs! inte helt samma definition som Num Linj Alg. boken s. 9) Om relativa felet uppfyller $|\delta x|/|x| < 0.5 \cdot 10^{-n}$ säger vi att approximationen $\hat{x}$ har (minst) n **signifikanta siffror**.

Se: `https://ece.uwaterloo.ca/~dwharder/NumericalAnalysis/01Error/SignificantDigits/`

- Approximationen $\hat{x} = 3.14$ av π har $n = 2$ korrekta decimaler och signifikanta siffror.

Felfortplantning

- Antag vi har en funktion/algoritm $f : I \to \mathbb{R}$, var $I \subset \mathbb{R}$ och att vi onskar att beräkna $f(x)$ för indatan $x \in I$.
- Felfortplanting studerar hur fel i indata $\hat{x} - x$ fortplantar sig till fel i utdata $f(\hat{x}) - f(x) =: \delta f(x)$.
- Om vi antar $f \in C^1(I, \mathbb{R})$ ger medelvärdessatsen

$$\delta f(x) = f'(x + \theta\delta x)(\hat{x} - x) = f'(x + \theta\delta x)\delta x, \quad \theta \in [0, 1].$$

- Observation: $|utdatafelet| > |indatafelet|$ om $|f'| > 1$, och annars är $|indatafelet| \geq |utdatafelet|$.
- Om inte θ och/eller x är bekanta, används ofta första ordningens approximativa felfortplaningsformel

$$\delta f(x) \approx f'(\hat{x})\delta x \tag{2}$$

med motsvarande absoluta och relativa felgränser

$$|\delta f(x)| \lesssim |f'(\hat{x})||\delta x| \quad \text{och} \quad \left|\frac{\delta f(x)}{f(\hat{x})}\right| \lesssim \left|\frac{f'(\hat{x})}{f(\hat{x})}\right| |\delta x|$$

Exempel 1.2

Vi önskar att bestämma volymen till en boll med radien r.
Antag vår mätning av radien $\hat{r}$ är bestamd med 1% relativfel (indatafel).
Ange felgräns för relativfelet till volymbestämmningen.

Lösning:

1. Formeln för volymen: $V(r) = (4/3)\pi r^3$ med $V'(r) = 4\pi r^2$.
2. Relativfelet högst 1% implicerar att $|\delta r| \leq 0.01|r| \lesssim 0.01|\hat{r}|$ (var vi använder approximativa felgränsen sedan vi inte känner r).
3. Approximativa felfortplantningsformeln ger

$$\begin{aligned} \left|\frac{\delta V(r)}{V(\hat{r})}\right| &\lesssim \left|\frac{V'(\hat{r})}{V(\hat{r})}\right| |\delta r| \\ &\lesssim \left|\frac{4\pi\hat{r}^2}{(4/3)\pi\hat{r}^3}\right| \cdot 0.01 \cdot |\hat{r}| = 0.03. \end{aligned}$$

Konklusion: 1% relativfel i indata ger att volymen kan bestämmas med 3% relativfel.

Högre ordnings felapproximation

- Medelvärdssatsen (2) leder till första ordningens approximation av felfortplantningen

$$\delta f(x) \approx f'(\hat{x})\delta x \tag{3}$$

- Om $f'(\hat{x}) \approx 0$, så är sannolikt approximationen (3) för optimistisk.
- Om t ex $f \in C^2(I, \mathbb{R})$ kan man i stället härleda en andra ordningens utveckling
  1. $$\delta f(x) = f'(x)\delta x + f''(x + \theta\delta x)\delta x^2/2, \quad \theta \in [0, 1],$$
  2. och, om $f'(\hat{x}) \approx 0$, approximativa felfortplantningsformeln

$$\delta f(x) \approx f''(\hat{x})\delta x^2/2.$$

Exempel 1.4 (ändrat)

Betrakta $f(x) = x^2 + 2x - 1$ och antag vi är givna mätningen $\hat{x} = -1 \approx x$ och $|\hat{x} - x|/|\hat{x}| \le 0.01$. Hur noggrant kan $f(x)$ bestämmas?

Lösning:

1. Sedan $f'(x) = 2x + 2$ är $f'(\hat{x}) = 0$, och vi använder approximationen
$$\delta f(x) \approx f''(\hat{x})\delta x^2/2.$$
2. $f(\hat{x}) = -2$ och $f''(\hat{x}) = 2$ ger
$$\left|\frac{\delta f(x)}{f(\hat{x})}\right| \lesssim \left|\frac{f''(\hat{x})}{f(\hat{x})}\right| \delta x^2/2 \le 0.5 \cdot 10^{-4}.$$

Konklusion: $f(x)$ kan bestämmas med fyra signifikanta siffror.

Felfortplantning i flera variabler

- Generalisering av begreppen ovan från endimensionella fall till n-dimensionella fall:
- För en approximation $\hat{x} \approx x$ var $\hat{x}, x \in \mathbb{R}^n$, och $\delta x = \hat{x} - x$ säger vi att $c_1, c_2 > 0$ är resp. felgränser för absoluta och relativfelen om
$$\|\delta x\| \le c_1, \quad \& \quad \frac{\|\delta x\|}{\|x\|} \le c_2.$$
- Om $f \in C^1(I, \mathbb{R})$ var $I \subset \mathbb{R}^n$, dvs $f(x) = f(x_1, x_2, \dots, x_n)$ kan felfortplaningen från fel i indata
$$\hat{x} = x + \delta x = (x_1 + \delta x_1, \dots, x_n + \delta x_n)$$
till fel i utdata $f(\hat{x}) = f(x) + \delta f(x)$ skattas via medelvärdssatsen
$$\delta f(x) = f(\hat{x}) - f(x) = \int_0^1 \frac{d}{ds} f(x + s\delta x)\, ds = \sum_{k=1}^{m} \frac{\partial f}{\partial x_k}(x + \theta\delta x)\delta x_k, \quad \theta \in [0, 1].$$
- Approximation, om x är obekant: $\delta f(x) \approx \nabla f(\hat{x}) \cdot \delta x$.

Felfortplantning i flera variabler forts.

- Cauchy-Schwartz, $|x \cdot y| \le \|x\|\|y\|$, $x, y \in \mathbb{R}^n$, är nyttig vid skattning av felgräns för felfortplantningen:
$$|\delta f(x)| \lesssim |\nabla f(\hat{x}) \cdot \delta x| \le \|\nabla f(\hat{x})\|\|\delta x\| \quad \& \quad \left|\frac{\delta f(x)}{f(x)}\right| \lesssim \frac{\|\nabla f(\hat{x})\|\|\delta x\|}{|f(x)|}. \tag{4}$$

Exempel 1.4 (ändrat): Härleda formel för relativa felet för multiplikation $f(x) = x_1 x_2$

Lösning: Ab initio:
$$|f(\hat{x}) - f(x)| = |(x_1 + \delta x_1)(x_2 + \delta x_2) - x_1 x_2| \lesssim |x_1 \delta x_2 + x_2 \delta x_1|.$$
$$\implies \left|\frac{\delta f(x)}{f(x)}\right| \lesssim \left|\frac{x_1 \delta x_2 + x_2 \delta x_1}{x_1 x_2}\right| \le \left|\frac{\delta x_2}{x_2}\right| + \left|\frac{\delta x_1}{x_1}\right|$$
Sedan $\nabla f = (x_2, x_1)$ ger (4) (den typisk mer konservativa?) felgränsen
$$\implies \left|\frac{\delta f(x)}{f(x)}\right| \lesssim \frac{\|x\|\|\delta x\|}{|x_1 x_2|} = \sqrt{x_1^{-2} + x_2^{-2}}\|\delta x\|.$$

Konditionstal och stabilitet för matematiska problem

Kondition och konditionstal

- (Relativa) koniditionstalet är definierad som

$$\kappa(x) = \lim_{\delta \downarrow 0} \max_{\|\delta x\| \le \delta} \frac{(|f(x+\delta x) - f(x)|/|f(x)|)}{(\|\delta x\|/\|x\|)} \approx \frac{\|\textbf{Relativfel utdata}\|}{\|\textbf{Relativfel indata}\|}$$

- Tolkning: Ett problem med indata relativfel $\alpha\%$ kan ha utdata relativfel $\kappa \cdot \alpha\%$.
- Ett problem med litet konditionstal, t ex $\kappa \in (0, 10)$, sägs vara **stabilt** eller **välkonditionerad**.
- Och omvand, ett problem med stort konditionstal, t ex $\kappa = 10^8$, sägs vara **instabilt** eller **illakonditionerad**.
- Tillämpning av approximationen $|f(x+\delta x) - f(x)| \lesssim \|\nabla f(x)\| \|\delta x\|$ ger följande estimat för κ:

$$\kappa(x) \lesssim \frac{(\|\nabla f(x)\| \|\delta x\|/|f(x)|)}{(\|\delta x\|/\|x\|)} = \|x\| \frac{\|\nabla f(x)\|}{|f(x)|}.$$

Exempel

1. Den endimensionella funktionen $f(x) = cx$ och $x \neq 0$ ger

$$\kappa(x) \lesssim |x| \frac{|f'(x)|}{|f(x)|} = |x| \frac{|c|}{|cx|} = 1.$$

2. Funktionen $f(x_1, x_2) = x_1 - x_2$ och $x = (1, 1 + 10^{-100})$ ger

$$\kappa(x) \approx \|x\| \frac{\|\nabla f(x)\|}{|f(x)|} \approx \sqrt{2} \frac{\|(1, -1)\|}{|1 - (1 + 10^{-100})|} \approx 2 \cdot 10^{100}.$$

   Observation: Subtraktionen $x_1 - x_2$ är en instabil operation när $x_1 \approx x_2$.

3. Funktionen $f(x) = Ax$ var A är en inverterbar $n \times n$ matris ger

$$\kappa(x) \lesssim \frac{(\|A\| \|\delta x\|/\|Ax\|)}{(\|\delta x\|/\|x\|)} = \|A\| \frac{\|x\|}{\|Ax\|}. \le \|A\| \|A^{-1}\|$$

   (Vi återkommer till matrisnorm $\|A\| := \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$ i kapitel 5.)

Felanalys för approximativa algoritmer

Approximativa algoritmer, framåt- och bakåtfel

- I tillägg till indatafel (t ex mätfel, eller trunkerad representation) finns en annan potensiell felkälla: exakta metoden/algoritmen f måste i många fall ersättas av en approximativ algoritm $\hat{f}$.
- **Framåtfelet** för en trippel $(f, \hat{f}, x)$ ges utav $\hat{f}(x) - f(x)$, som beskriver approximativa algoritmens fel i utdata.
- Antag i fortsättningen att f är inverterbar.
- Observera att för värje värde $\hat{f}(x)$ finns då ett indatavärde $\hat{x} = f^{-1}(\hat{f}(x))$ så att
$$f(\hat{x}) = f(f^{-1}(\hat{f}(x))) = \hat{f}(x).$$
- **Bakåtfelet** för en trippel $(f, \hat{f}, x)$ är definierad som
$$\hat{x} - x = f^{-1}(\hat{f}(x)) - f^{-1}(f(x)).$$
Bakåtfelet $\hat{x} - x$ motsvarar det absoluta fel δx som ger felfortplantningen $f(\hat{x}) - f(x) = \hat{f}(x) - f(x)$.

Exempel 1.8

Låt $f(x) = e^x$ och
$$\hat{f}(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!}$$
vara resp. exakta och approximativa algoritmen. Berakna framåt- och bakåtfelet för $(f, \hat{f}, x)$ med $x = 1$.

Lösning:

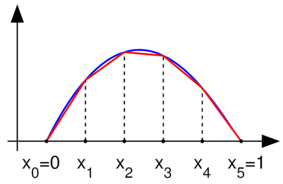
1. Sedan
$$\hat{f}(1) = (1 + 1 + 1/2 + 1/6) = 8/3$$
bliver framåtfelet
$$\hat{f}(1) - f(1) = 8/3 - e^1 = -0.051615\ldots$$
2. För bakåtfelet, observera att $f^{-1}(x) = \ln(x)$. Det ger
$$\hat{x} = f^{-1}(\hat{f}(1)) = \ln(8/3)$$
och
$$\hat{x} - x = \ln(8/3) - 1 = -0.019171\ldots$$

Numerisk algoritm

Exempel, matematisk problem: Beräkna integralen

$$I = \int_0^1 g(x)\, dx.$$

Om vi inte får till att lösa problemet exakt, kan vi prova att lösa det numeriskt. Typiska steg från matematisk problem till numerisk algoritm:



Figur: Wikimedia com.

1. Reduktion av informationsmängd: approximera funktionen $g : [0,1] \to \mathbb{R}$ med styckvis-linjär funktion $\bar{g}$ sådan att $\bar{g}(x_i) = g(x_i)$ för $0 = x_0 < x_1 < \ldots < x_n = 1$.
2. Förenklad (numeriskt) problem: Beräkna integralen

$$\hat{I} = \int_0^1 \bar{g}(x)\, dx.$$

3. Numerisk algoritm

$$\hat{I} = \sum_{i=0}^{n-1} (g(x_i) + g(x_{i+1})) \frac{(x_{i+1} - x_i)}{2}.$$

Stabilitet för numeriska algoritmer

- Kom ihåg att ett problem sägs vara stabilt/välkonditionerad om konditionstalet är litet

$$\kappa(x) = \frac{\|\textbf{relativfel utdata}\|}{\|\textbf{relativfel indata}\|}.$$

- Med andra ord, ett problem är stabilt om relativt små förändringar i indata ger relativt små förändringar i utdata.
- Definitionen: En numerisk algoritm $\hat{f} \approx f$ sägs vara stabil vid $x \neq 0$ om relativa bakåtfelets storlek

$$\frac{|f^{-1}(\hat{f}(x)) - x|}{\|x\|}$$

  är litet.
- Med andra ord, $\hat{f}$ är stabil om $\hat{f}(x)$ är exakta lösningen till nästan rätt problem ($f(\hat{x})$).
- Exempel 1.8: $|f^{-1}(\hat{f}(1)) - 1|/1 \approx 0.02$. Algoritmen är stabil för $x = 1$.

Framåt- och bakåtanalys

- Vi använder bakåtanalys i definitionen av stabilitet bland annat för att det i många fall är lättare att studera än framåtanalys.
- Exempel: Betrakta problemet $Ax = b$, A stor men inverterbar matris, och låt

$$f(b) = A^{-1}b, \quad \text{och} \quad \hat{f}(b) = (A^{-1} + \varepsilon I)b.$$

- Framåtanalys estimerar (framåtstabiliteten)

$$\frac{\|\hat{f}(b) - f(b)\|}{\|f(b)\|},$$

  vilket är svårt att komma åt i fall där A är stor och vi inte kan lösa problemet exakt.
- I bakåtanalysen anvander vi att $f^{-1}(y) = Ay$, och estimerar enkelt (bakåt) stabiliteten

$$\frac{\|f^{-1}(\hat{f}(b)) - b\|}{\|b\|} = \frac{\|A(A^{-1} + \varepsilon I)b - b\|}{\|b\|} = \frac{\|\varepsilon Ab\|}{\|b\|} \leq |\varepsilon| \|A\|.$$

Flyttalsaritmetik

Datoraritmetik – datatyper

- I programmeringsspråk kan tal sparas i en del olika format (datatyper):
  - signed int32 eller 64(32 eller 64-bits heltal), dvs alla heltal i resp. $[-2^{-31}, 2^{31}-1]$ och $[-2^{-63}, 2^{63}-1]$
  - single precission floating point (32-bit flyttal) $\hat{\pi} = 3.14159274101257$
  - double precission floating point (64-bit flyttal) $\hat{\pi} = 3.14159265358979$
  - ...
- I Matlab sparas alla tal som flyttal i datatypen **double 64** per default. Om du önskar att spara i annan datatyp måste du ange det när tilldelar en variabel ett värde.

```
>> x = pi
x =   3.14159265358979
>> y = int64(pi)
y = 3
```

Flyttal

- Ett flyttalssystem definieras av kvadrupeln $\mathcal{F} = (\beta, t, L, U)$.
  - $\beta \in \mathbb{N} \setminus \{1\}$ är basen till systemet. $\beta = 2, 10$ ger resp. det binära- och titalssystemet.
  - t är antalet siffror i systemet (relaterad till antal signifikanta siffror).
  - L och U är resp. lägsta och högsta värde på exponenten (se nedan).
- Varje flyttal x i $\mathcal{F}$ representeras på formen

$$x = \pm(d_0 + d_1\beta^{-1} + d_2\beta^{-2} + \ldots + d_{t-1}\beta^{1-t})\beta^e$$

  var $1 \leq d_0 \leq \beta - 1$, $0 \leq d_i \leq \beta - 1$ för alla $1 \leq i \leq t-1$ och $L \leq e \leq U$.
- Med mantissaen $m = d_0.d_1 \ldots d_{t-1}$ kan talet ovan skrivas $x = \pm m \cdot \beta^e$.
- Exempel:Flyttalssystemet $(10, 3, -1, 1)$ består av tal på formen

$$x = \pm d_0.d_1d_2 \times 10^e, \quad 1 \leq d_0 \leq 9, \quad 0 \leq d_1, d_2 \leq 9, \quad -1 \leq e \leq 1.$$

- T ex flyttalet $x = -0.94 = -9.40 \cdot 10^{-1}$ fås vid $d_0 = 9, d_1 = 4, d_2 = 0$ och $e = -1$.

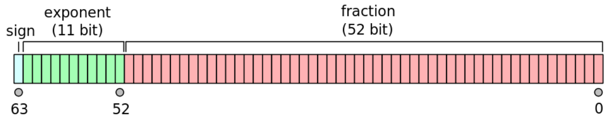
Binära flyttalssystemet double

- Exempel: Flyttalsstemet (IEEE-754) double beskrivs av kvadrupeln $(\beta = 2, t = 53, L = -1022, U = 1023)$ och består av alla tal

$$x = \pm \underbrace{d_0.d_1 \dots d_{52}}_{\text{mantissa}} \times 2^e, \ \text{var} \quad d_0 = 1 \quad 0 \le d_i \le 1, \quad i \ge 1,$$

  och $\quad -1022 \le e \le 1023.$
- T ex talet

$$x = 2.75 = (1 + 0 \cdot 2^{-1} + 2^{-2} + 2^{-3}) \cdot 2^1 \quad \text{skrivas} \quad 1.0110 \dots 0_2 \cdot 2^1.$$



Figur: Hur ett 64-bit flyttal sparas i ett datorregister (Wikimedia commons).

Underflow, overflow och flyttalsavrunding

- Då vi har normaliseringen $1 \le d_0 \le \beta - 1$, uppfyller systemet $\mathcal{F} = (\beta, t, L, U)$ att

$$\min(\mathcal{F} \cap \mathbb{R}_+) = 1.0 \dots 0 \cdot \beta^L = \beta^L =: \text{UFL(underflow level)}$$

  och

$$\max(\mathcal{F} \cap \mathbb{R}_+) = (\beta - 1).\underbrace{(\beta - 1) \dots (\beta - 1)}_{t-1} \cdot \beta^U = \beta^{U+1}(1 - \beta^{-t}) =: \text{OFL},$$

  var OFL star för overflow level.
- För att konvertera/avrunda rella tal till flyttal kan vi tänka oss, något förenklad, att följande avbildning $fl : \mathbb{R} \to \mathcal{F}$ benyttas:

$$fl(x) = \begin{cases} \arg\min_{y \in \mathcal{F}} |y - x| & \text{om } x \in [-OFL, OFL] \\ \text{Inf} & \text{om } |x| > OFL. \end{cases}$$

- Observera, i denna förenklade beskrivningen är $fl(0) = UFL \neq 0$.

Maskinprecision och flyttalsaritmetik

- Följande egenskap gäller för alla $|x| < UFL$, (övning 1.5)

$$\left| \frac{fl(x) - x}{x} \right| \le 0.5 \cdot \beta^{1-t}.$$

- Dvs $fl(x)$ approximerar x med $t - 1$ signifikanta siffror (i basen β) och talet $\mu = 0.5 \cdot \beta^{1-t}$ kallas **maskintalet**.
- Alternativt skriver man

$$fl(x) = x(1 + \delta), \quad \text{för någon} \quad |\delta| \le \mu.$$

- Flyttalssystem är implementerad så att flyttalsaritmetik göras i ett utvidgat register och

$$fl(x \diamondsuit y) \le x \diamondsuit y(1 + \delta) \quad \text{för någon } |\delta| \le \mu,$$

  om $\diamondsuit = +, *, -$ eller $/$, $x, y \in \mathcal{F}$ (och $x \diamondsuit y \neq Inf$).
- Kommentar: double tillåter “gradual underflow” och $fl(0) = 0$ (se kap 1.8 för detaljer).

Flyttalsaddition i dator

Illustration av flyttalsaddition i systemet $\mathcal{F} = (10, 4, -9, 9)$.
Summan av $x = 1.234 \cdot 10^0$ och $y = 4.567 \cdot 10^{-2}$.

1. Bestäm vilken av x och y som har störst exponent. Representera båda talen med mantissor anpassad största exponenten:
$$x = 1.234 \cdot 10^0 \quad \& \quad y = 0.04567 \cdot 10^0.$$
2. Addera talen:
$$x + y = 1.27967 \cdot 10^0$$
3. Avrunda till närmaste flyttal:
$$fl(x + y) = fl(1.27967 \cdot 10^0) = 1.280 \cdot 10^0.$$

Flyttalsaddition är inte associativ

- **Exempel 1.12** Låt $\mathcal{F} = (10, 4, -9, 9)$ och betrakta flyttalen
$$a = 9.876 \cdot 10^4, \quad b = -9.880 \cdot 10^4, \quad c = 3.456 \cdot 10^1.$$
- Å ena sidan får man
$$fl(fl(a + b) + c) = fl(-4.000 \cdot 10^1 + 3.456 \cdot 10^1) = -5.440 \cdot 10^0,$$
men å andra sidan
$$\begin{aligned} fl(a + fl(b + c)) &= fl(a + fl((-9.880 + 0.003456) \cdot 10^4)) \\ &= fl((9.876 - 9.877) \cdot 10^4) \\ &= -1.000 \cdot 10^1. \end{aligned}$$
- Konklusion: Det gäller ej generellt at
$$fl(fl(a + b) + c) = fl(a + fl(b + c)).$$

Utskiftning

- Utskifning innebär att noggrannhet går förlorad när flyttal med olika storleksordningar adderas/subtraheras.
- Exempel **fullständing utskiftning**: Addition av flyttalen $a = 1.234 \cdot 10^7$ och $b = 1.678 \cdot 10^3$ i $\mathcal{F} = (10, 4, -9, 9)$ ger
$$\begin{aligned} fl(a + b) &= fl((1.234 + 0.0001678) \cdot 10^7) \\ &= fl((1.2341678) \cdot 10^7) \\ &= 1.234 \cdot 10^7 \\ &= a. \end{aligned}$$
- **Partiell utskiftning:** om en följd av flyttal $x_1, x_2, \ldots$ av olik storleksordning adderas.
- Strategi för att undvika utskiftning: Summera följden i (till beloppet) vaxande storleksordning.

Utskiftning i Matlab

```
>> eps
ans =      2.22044604925031e-16

>> 2^(1-53)
ans =      2.22044604925031e-16
%Konklusion: Matlabs variabel eps=2^(-52) = 2*mu

>> x=x+eps
x =    1.00000000000000 %   = 1+eps

>> x=1+2*eps/3
x =    1.00000000000000 % = 1+ eps

>> x=1+eps/3
x =    1
```

Kancellation

- **Kancellation** innebär nogrannhetsförlusten som uppstar vid subtraktion av två nastan lika stora tal (efter tidigare operationer med avrundingsfel).
- **Exempel 1.15** Betrakta följande beräkning

$$37654 + 25.874 - 37679 = 0.874$$

i flyttalssystemet $\mathcal{F} = (10, 5, -9, 9)$.
- Dvs

$$\begin{aligned} & fl(fl(3.7654 \cdot 10^4 + 2.5874 \cdot 10^1) - 3.7679 \cdot 10^4) \\ &= fl(fl((3.7654 + 0.0025874) \cdot 10^4) - 3.7679 \cdot 10^4) \\ &= fl(fl(3.7679874 \cdot 10^4) - 3.7679 \cdot 10^4) \\ &= fl(3.7680 \cdot 10^4 - 3.7679 \cdot 10^4) \\ &= 1.0000 \cdot 10^0 \end{aligned}$$

Framåtanalys av flyttalsalgoritm

- **Exempel 1.17:** Kvadrering: $f(x) = x^2$ för $x \in \mathbb{R} \setminus \{0\}$ approximeras i flyttalssystem av $\hat{f}(x) = fl(fl(x)^2)$. Låt oss verifiera att

$$\left| \frac{\hat{f}(x) - f(x)}{f(x)} \right| = \mathcal{O}(\mu)$$

vid framåtanalys:

1. Sedan $fl(x) = x(1 + \delta)$ för någon $\delta \in [-\mu, \mu]$ var μ är maskintalet, följer det att

$$\hat{f}(x) = x^2(1 + \delta_1)^2(1 + \delta_2), \quad \text{för några} \quad \delta_1, \delta_2 \in [-\mu, \mu].$$

2. Det ger relativa felgränsen

$$\left| \frac{\hat{f}(x) - f(x)}{f(x)} \right| = \left|(1 + \delta_1)^2(1 + \delta_2) - 1\right| = O(\mu).$$

Bakåtanalys

Repetition: den approximativa algoritmen $\hat{f} \approx f$ är stabil för $x \neq 0$ om relativa bakåtfelet är litet.

För flyttalsalgoritmer kopplar vi "litet" till flyttalssystemets maskintal $\mu = 0.5 \cdot \beta^{1-t}$. Dvs flyttalsalgoritmen är stabil om

$$\left| \frac{f^{-1}(\hat{f}(x)) - x}{x} \right| = \mathcal{O}(\mu).$$

Exempel 1.17: Låt oss verifiera att $\hat{f} = fl(fl(x)^2) \approx x^2 = f(x)$ är stabil.

1. Vi har

$$\hat{f}(x) = x^2(1+\delta_1)^2(1+\delta_2), \quad \text{för några} \quad \delta_1, \delta_2 \in [-\mu, \mu].$$

2. Inversen $f^{-1}(y) = \sqrt{y}$ och $\sqrt{1+\delta} = 1 + \delta/2 + \mathcal{O}(\delta^2)$ ger

$$f^{-1}(\hat{f}(x)) = x\left((1+\delta_1)\sqrt{(1+\delta_2)}\right) = x\left(1 + \delta_1 + \delta_2/2 + \mathcal{O}(\delta_1^2 + \delta_2^2)\right),$$

3. Flyttalskvadrering är stabil:

$$\left| \frac{f^{-1}(\hat{f}(x)) - x}{x} \right| \lesssim \left| \frac{x\,(\delta_1 + \delta_2/2)}{x} \right| = O(\mu).$$