LANA 2016 Demo 11-1

NA 1.6 Undersök felfortplantning av det relativa felet i x vid evaluering av $\sin(x)$.

(a) Ge en gräns för det absoluta felet.

Lösning Använd Taylors formel: $\sin(x+\delta x) = \sin(x) + \cos(y)\,\delta x$ för något y mellan x och $x+\delta x$. Alltså

$$\underbrace{|\sin(x) - \sin(x+\delta x)|}_{\text{absolut fel i utdata}} = |\cos(y)\,\delta x| = \underbrace{|\cos(y)|}_{\leq 1}|\delta x| \leq \underbrace{|\delta x|}_{\substack{\text{absolut fel}\\ \text{i indata}}} .$$

(b) Ge en gräns för det relativa felet.

Lösning Enligt ovan:

$$\underbrace{\frac{|\sin(x) - \sin(x+\delta x)|}{|\sin(x)|}}_{\text{relativt fel i utdata}} = \frac{|\cos(y)|}{|\sin(x)|}|\delta x| = \frac{|\cos(y)||x|}{|\sin(x)|}\underbrace{\frac{|\delta x|}{|x|}}_{\text{relativt fel i indata}} \approx \frac{|\cos(x)||x|}{|\sin(x)|}\frac{|\delta x|}{|x|}$$

(c) Ge en gräns för konditionstalet.

Lösning Enligt definitionen av konditionstal K (relativt fel ut / relativt fel in) får vi $K = \dfrac{|\cos(y)||x|}{|\sin(x)|} \approx \dfrac{|\cos(x)||x|}{|\sin(x)|}$.

(d) För vilka x är evalueringen illakonditionerad, dvs. $K \gg 1$?

Lösning K är stort om $|x|/|\sin(x)|$ är stort, alltså $\sin(x) \approx 0$ och $|x| \gg 0$, alltså $x = n\pi$, $n = \pm1, \pm2, \pm3, \ldots$

NA 1.7 (a) Bestäm framåt- och bakåtfel för approximationen $\sin(x) \approx x$ med $x = 0.1$, 0.5 resp. 1.0.

Lösning Framåtfel := |"exakt utdata" - "approximation"| = $|\sin(x) - x|$.

Bakåtfel := $|x - \hat{x}|$ där $\hat{x}$ är sådan indata som exakt reproducerar approximationen, alltså här

$$\sin(\hat{x}) = x \implies \hat{x} = \arcsin(x).$$

Vi får

x	0.1	0.5	1.0
$\lvert\sin(x) - x\rvert$	0.000167	0.0206	0.159
$\lvert x - \arcsin(x)\rvert$	0.000167	0.0236	0.571

(b) Samma som (a), men approximation $\sin(x) \approx x - \frac{x^3}{3!}$.

Lösning Här får vi istället $\hat{x} = \arcsin\left(x - \frac{x^3}{3!}\right)$.

x	0.1	0.5	1.0
$\lvert\sin(x) - x + \frac{x^3}{3!}\rvert$	0.0000000833	0.000259	0.00814
$\lvert x - \arcsin(x + \frac{x^3}{3!})\rvert$	0.0000000837	0.000295	0.0149

LANA 2016 Demo 11-2

NA 1.16 Visa att det relativa felet vid flyttalsberäkning av $x-y$ begränsas av $\mu + 2\mu(1+\mu)\frac{\max\{|x|,|y|\}}{|x-y|}$, där μ är avrundningsenheten. Relatera felet till kancellation.

Lösning Avrundningsenheten uppfyller $fl(x*y)=(x*y)(1+\delta_*)$, $fl(z)=z(1+\delta_z)$ med $|\delta_*|, |\delta_z| \le \mu$, där $fl(\cdot)$ representerar flyttalsapproximation, x, y, z är tal, $*$ är någon räkneoperation. Vi får alltså

$$\text{relativt fel} = \frac{|(x-y) - fl(fl(x)-fl(y))|}{|x-y|} = \frac{|(x-y)-(fl(x)-fl(y))(1+\delta_-)|}{|x-y|}$$

$$= \frac{|(x-y) - (x(1+\delta_x) - y(1+\delta_y))(1+\delta_-)|}{|x-y|}$$

$$= \frac{|(x-y)\delta_- + x\delta_x(1+\delta_-) - y\delta_y(1+\delta_-)|}{|x-y|} \le \text{triangelolikheten}$$

$$\le |\delta_-| + \frac{|x||\delta_x|(1+|\delta_-|)}{|x-y|} + \frac{|y||\delta_y|(1+|\delta_-|)}{|x-y|}$$

$$\le \mu + \mu(1+\mu)\frac{|x|+|y|}{|x-y|} \le \mu + 2\mu(1+\mu)\frac{\max\{|x|,|y|\}}{|x-y|}.$$

Kancellation uppstår genom skadlig interaktion mellan avrundning av tal och approximation av räkneapproximationer. För att avgöra vilken del av felet som uppstår genom kancellation jämför vi uttrycket ovan med det som erhålls då x och y är exakt lagrade flyttal, dvs. $fl(x)=x$, $fl(y)=y$. Då får vi

$$\text{relativt fel} = \frac{|(x-y) - fl(x-y)|}{|x-y|} = \frac{|(x-y)-(x-y)(1+\delta_-)|}{|x-y|}$$

$$= \frac{|(x-y)\delta_-|}{|x-y|} = |\delta_-| \le \mu.$$

Vi sluter oss till att den andra termen i den första uppskattningen (som blir stor då $x-y \approx 0$ men $|x|,|y| \gg 0$) kommer från kancellation.

LANA 2016 Demo 11-3.

NA1.18 Beskriv hur $\|x\|_2 := \sqrt{\sum_{i=1}^n x_i^2}$ kan beräknas utan risk för "overflow" eller "underflow".

Lösning Risken med "overflow" (för stora tal sätts till ∞ i flyttalssystem) är att ett x_i kvadreras och avrundas till ∞, även om $\sqrt{x_i^2}$ kunde lagras i flyttalssystemet.

Risken med "underflow" (för små tal avrundas till 0 i flyttalssystem) är att ett x_i kvadreras och avrundas till 0, även om $\sqrt{x_i^2}$ inte är försumbart och kunde lagras i flyttalssystemet.

Båda problemen kan avrundas genom att normera beräkningen: Låt $m := \max_i |x_i|$ och beräkna $\|x\|_2 = m\sqrt{\sum_{i=1}^n \left(\frac{x_i}{m}\right)^2}$. Eftersom $|x_i/m| \le 1$ undviks "overflow", och termer $\left(\frac{x_i}{m}\right)^2$ försummas bara om de faktiskt är försumbara i förhållande till det största elementet x_i med $|x_i| = m$.

NA1.22 Låt $x, y \ge 0$ vara intilliggande flyttal i IEEE-DP (dubbel presicion).

(a) Vilket är det minsta möjliga avståndet mellan x och y?

Lösning IEEE-DP är flyttalssystemet $(\beta, t, L, U) = (2, 53, -1022, 1023)$, alltså tal på formen $(a_0 \cdot 2^0 + a_1 \cdot 2^{-1} + \dots + a_{52} \cdot 2^{-52}) \cdot 2^e$ med $-1022 \le e \le 1023$ och $a_i \in \{0, 1\}$, $i = 0, \dots, 52$, med $a_0 = 1$ om $e > -1022$. Om $e = -1022$ tillåts $a_0 = 0$ så att mindre tal kan representeras (gradvis underspill, "gradual underflow").

Talen i ett flyttalssystem är som tätast kring 0, så minsta avståndet är det mellan det minsta representerbara talet och 0 (eller näst minsta utan gradvis underspill), här $(0 \cdot 2^0 + 0 \cdot 2^{-1} + \dots + 0 \cdot 2^{-51} + 1 \cdot 2^{-52}) \cdot 2^{-1022} - (0 \cdot 2^0 + \dots + 0 \cdot 2^{-52}) \cdot 2^{-1022} = 2^{-1022-52} = 2^{-1074}$.

(b) Vilket är det största möjliga avståndet mellan x och y.

Lösning Talen blir glesare ju större de är, så största avståndet är det mellan det största och det näst största talet. Alltså $(1 \cdot 2^0 + \dots + 1 \cdot 2^{-52}) \cdot 2^{1023} - (1 \cdot 2^0 + \dots + 1 \cdot 2^{-51} + 0 \cdot 2^{-52}) = 2^{1023-52} = 2^{971}$.

LANA 2016 Demo 11-4

NA1.25 Betrakta flyttalssystemet $(\beta, t, L, U) = (10, 3, -98, 98)$.

(a) Bestäm UFL i detta system.

Lösning UFL ("under flow limit", underspillsgränsen) = "minsta tal som kan representeras" $= (1\cdot 10^0 + 0\cdot 10^{-1} + 0\cdot 10^{-2})\cdot 10^{-98} = 10^{-98}$.
Notera att vi inte antar gradvis underspill, dvs. första siffran är alltid nollskild.

(b) Beräkna $x-y$ i detta system då $x = 6.87\cdot 10^{-97}$ och $y = 6.81\cdot 10^{-97}$.

Lösning $x - y = (6.87 - 6.81)\cdot 10^{-97} = 0.06\cdot 10^{-97} < \text{UFL}$

$$\Rightarrow fl(x-y) = 0.$$

(c) Vad ger beräkningen i (b) om gradvis underspill tillåts?

Lösning $x-y = 0.06\cdot 10^{-97} = 0.60\cdot 10^{-98}$ kan nu representeras i flyttalssystemet, så $fl(x-y) = 0.60\cdot 10^{-98}$.