Numerisk Analys, MMG410. Exercises 1. Flyttalsaritmetik.

1. Vi vet att $x = 24.516$ är ett korrekt avrundat värde. Beräkna absolutbeloppen av de maximala absoluta och relativa felen.

Lösning:

Absoluta felet ≤ 0.0005 (en halv enhet i sista decimalen (en halv enhet i fjärde siffran) för $x = 24.\underbrace{516}_{3 ciffror}$).

Absoluta felet räknas: $e_{abs} = |x - \hat{x}|$, $|x - \hat{x}| \leq 0.0005$, och

Relativa felet räknas som $e_{rel} = \frac{|x-\hat{x}|}{|x|} = \frac{0.0005}{24.516} \approx 2 \cdot 10^{-5}$.

3. $\hat{x}$ är en approximation av ett exakt värde x där $|\hat{x} - x| \leq \delta$. Hur kan vi uppskatta $|f(\hat{x}) - f(x)|$ givet funktionen f ? Vi känner $\hat{x}$ och δ men inte x. Tillämpa resonemanget på $f(x) = 7x + 3$ respektive $f(x) = x^2$. Ledning: använd Taylors formel.

Lösning:

Vi vet att $e_{abs} = |x - \hat{x}| \leq \delta$ eller $-\delta \leq x - \hat{x} \leq \delta$ och då $\hat{x} - \delta \leq x \leq \hat{x} + \delta$. Taylors formel ger
$f(x) = f(\hat{x} + x - \hat{x}) = f(\hat{x}) + (x - \hat{x})f'(\xi), \xi \in (x, \hat{x})$, så
$|f(x) - f(\hat{x})| \leq \delta \max_{\xi \in (\hat{x}-\delta, \hat{x}+\delta)} |f'(\xi)|$.
1) Om f är linjär (första fallet) så är $f'_x = (7x + 3)'_x = 7$ konstant, 7 i exemplet, varför absoluta felet begränsas av 7δ:
$|f(x) - f(\hat{x})| \leq \delta \cdot 7$.
2) I andra fallet får vi begränsningen $2\delta(|\hat{x}| + \delta)$ eftersom derivatan, $f'(x) = (x^2)'_x = 2x$, är strängt växande: $|f(x) - f(\hat{x})| \leq \delta \max_{\xi \in (\hat{x}-\delta, \hat{x}+\delta)} |f'(\xi)| = \delta \max_{\xi \in (\hat{x}-\delta, \hat{x}+\delta)} |2\xi| = \delta \cdot 2(|\hat{x}| + \delta)$.

5. Vi vill beräkna $f(x)$ givet x och den deriverbara funktionen, $f : \mathbb{R} \to \mathbb{R}$. Uppskatta konditionstalet κ för små störningar i x. Testa på $f(x) = \cos x$ då $x = \delta$ och $d\mathring{a}x = \pi/2 - \delta$, med litet $\delta > 0$.

Lösning:

Frågan är alltså: hur ändras $f(x)$ när vi ändrar x lite? Taylors formel ger: $f(x + \delta x) = f(x) + \delta x f'(x) + \ldots$. Den absoluta förändringen är: $f(x + \delta x) - f(x) \approx \delta x f'(x)$. Om $x \neq 0$ och $f(x) \neq 0$ är skilda från noll kan vi studera relativa förändringar:

$$\frac{f(x + \delta x) - f(x)}{f(x)} \approx \frac{\delta x \; f'(x)}{f(x)} \frac{x}{x},$$

$$\left| \frac{f(x + \delta x) - f(x)}{f(x)} \right| \approx \underbrace{\left| \frac{x f'(x)}{f(x)} \right|}_{\kappa} \left| \frac{\delta x}{x} \right|$$

dar ar en uppskattning av konditionstalet. Då $f(x) = \cos x$ får vi:

$$\kappa = \left| \frac{x f'(x)}{f(x)} \right| = \left| \frac{x(-\sin x)}{\cos x} \right|.$$

När $x = \delta > 0$ ($x = 0$ ger division med noll; dessutom är $\sin 0 = 0$, så för att få en uppskattning får man titta på nästa term i Taylorutvecklingen) kan vi göra följande approximation, för att lättare kunna analysera vad som händer:

$$\kappa = \left| \frac{xf'(x)}{f(x)} \right| = \left| \frac{x(-\sin x)}{\cos x} \right| \approx \delta^2$$

eftersom $\sin x \approx x, \cos x \approx 1$ för $x = \delta \approx 0$.
När $x = \pi/2 - \delta$ får vi

$$\kappa = \left| \frac{x(-\sin x)}{\cos x} \right| \approx \left| \frac{(\pi/2 - \delta)(-\sin(\pi/2 - \delta))}{\cos(\pi/2 - \delta)} \right| \approx \left| \frac{\pi/2 \cdot 1}{\delta} \right|$$

Detta κ växer som $1/\delta$ och kan bli mycket stort.

6. Antag att f är en deriverbar funktion och att δ är en deriverbar störning som är begränsad, $|\delta(x)| \leq \varepsilon$ för alla x. Diskutera hur känslig: 1) derivatan av f är för störningar i funktionen. Dvs. säg något om derivatan av $f(x) + \delta(x)$. 2) Gör motsvarande för integralen, $\int_a^b (f(x) + \delta(x))dx$.

Lösning:

1) Derivatan av en funktion behöver inte vara begränsad även om funktionen är det. Tag t.ex. $\delta(x) = \varepsilon \cos(\omega x)$. Funktionen är tydligen begränsad, men derivatan kan vara godtyckligt stor om vi väljer ett stort ω (hög frekvensen medför stor derivata):

$$\begin{aligned} |(f(x) + \delta(x))' - f'(x)| &= |f'(x) + \delta'(x) - f'(x)| \\ &= |(\varepsilon \cos(\omega x)'_x| = |\omega\varepsilon(-sin\omega x)|. \end{aligned}$$

En liten störning av f kan alltså ändra derivatan godtyckligt mycket.

2) Detta gäller inte integralen. Vi har ju:

$$\left| \int_a^b (f(x) + \delta(x))\ dx - \int_a^b f(x)\ dx \right| = \left| \int_a^b \delta(x)\ dx \right| \leq (b-a)\varepsilon.$$

8. Här följer en förberedelse för linjära ekvationssystem. Studera hur känslig lösningen, x, är för störningar i den reella parametern α, då:

$$\begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Vi kan tänka oss detta som en funktion också, nämligen den som avbildar α på x, så en funktion från $\mathbb{R}$ till $\mathbb{R}^2$. Man kan utnyttja derivator för att studera problemet, men man kan ju även lösa ut x som funktion av $\alpha : x = x(\alpha)$. För vilka α är x känslig för förändringar i α ?

Lösning:

Vi kan beräkna x explicit genom $x = A^{-1}b$ förutsatt att inversen existerar, dvs. då $|\alpha| \neq 1$. Beräkning av A^{-1}:

$$A^{-1} = \frac{1}{\det A}[C_{ij}^T] = \frac{1}{1-\alpha^2}\begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix}$$

$$x = \frac{1}{1-\alpha^2}\begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{1-\alpha^2}\cdot\begin{bmatrix} 1 \\ -\alpha \end{bmatrix}.$$

Om $|\alpha| \approx 1$ kommer små variationer i α att ge upphov till stora förändringar i x.

Låt, t.ex. $\alpha = 1 - \varepsilon$ då blir x

$$x = \frac{1}{1-\alpha^2} \cdot \begin{bmatrix} 1 \\ -\alpha \end{bmatrix} = \frac{1}{1-(1-\varepsilon)^2} \cdot \begin{bmatrix} 1 \\ -(1-\varepsilon) \end{bmatrix} \approx \frac{1}{2\varepsilon} \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Små variationer i ε ger upphov till stora variationer i x. Beräkningen av x är således illakonditionerad då $|\alpha| \approx 1$.

9. Upprepa ovanstående då vi har två parametrar, α och β:

$$\begin{bmatrix} \alpha & \beta \\ \beta & \alpha \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Lösning:
Vi kan beräkna x explicit genom $x = A^{-1}b$ förutsatt att inversen existerar, dvs. då $|\alpha| \neq |\beta|$.

$$x(\alpha, \beta) = \frac{1}{\alpha^2 - \beta^2} \begin{bmatrix} \alpha & -\beta \\ -\beta & \alpha \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\alpha^2 - \beta^2} \cdot \begin{bmatrix} \alpha \\ -\beta \end{bmatrix}$$

som är känslig för störningar då $|\alpha| \approx |\beta|$.

14. Antag att vi arbetar med fyrsiffrig decimal aritmetik. Beräkna följande summor samt de absoluta och relativa felen:

- 6.278 + 4.039
- 6.278e10 + 4.039e10
- 6.278e-10 + 4.039e-10

Lösning:
Vi kan klara av alla tre fallen på en gång genom att låta p, nedan, anta värdena $p = 0, 10, -10$. Det exakta värdet blir
$x = 6.278 \cdot 10^p + 4.039 \cdot 10^p = 1.0317 \cdot 10^{p+1}$ och det lagras, korrekt avrundat och i normaliserad form, som $\hat{x} = 1.032 \cdot 10^{p+1}$. Det absoluta felet blir:

- Det absoluta felet blir:
  $|x - \hat{x}| = |1.0317 \cdot 10^{p+1} - 1.032 \cdot 10^{p+1}| = 3 \cdot \cdot 10^{-4} \cdot 10^{p+1} = 3 \cdot 10^{p-3}$.
- Det relativa felet blir: $\frac{|x-\hat{x}|}{|x|} = 3 \cdot 10^{p-3} / 1.0317 \cdot 10^{p+1} \approx 3 \cdot 10^{-4}$.

Notera att de tre absoluta felen blir: $3 \cdot 10^{-3}$, $3 \cdot 10^{7}, 3 \cdot 10^{-13}$. Det relativa felet blir, i alla tre fallen, $\approx 3 \cdot 10^{-4}$.

15. Visa att addition enligt IEEE är en stabil algoritm.
Lösning:
Vi vet att

$$fl(a+b) = (a+b)(1+\epsilon), |\epsilon| \leq \epsilon_{mach}.$$

Alltså gäller att

$$\begin{aligned} fl(a+b) &= \tilde{a} + \tilde{b}, \\ \tilde{a} &= a(1+\epsilon),\ \tilde{b} = b(1+\epsilon), \\ |\tilde{a} - a| &\leq \epsilon_{mach}|a|,\ |\tilde{b} - b| \leq \epsilon_{mach}|b| \end{aligned}$$

fl(a + b) är således den exakta summan av två något störda tal.

17. Är skalärproduktsberäkning med IEEE stabil? Dvs. är det stabilt att bilda

$$\sum_{k=1}^{n} x_k y_k$$

Lösning:

Låt oss betrakta specialfallet då $n = 4$. För att få mindre oläsliga formler inför vi $\sigma_k = 1 + \delta_k$, $\pi_k = 1 + \epsilon_k$ (σ för summa och π för produkt) där alla $|\epsilon_k|$ och $|\delta_k|$ är mindre är ϵ_{mach}. Vi har:

- $fl(x_1 y_1) = x_1 y_1 \pi_1$ (vi använder inte σ_1)
- $fl(x_1 y_1 + x_2 y_2) = [x_1 y_1 \pi_1 + x_2 y_2 \pi_2]\sigma_2$
- $f(x_1 y_1 + x_2 y_2 + x_3 y_3) = ([x_1 y_1 \pi_1 + x_2 y_2 \pi_2]\sigma_2 + x_3 y_3 \pi_3)\sigma_3$.
- $fl(x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4) =$
  $[([x_1 y_1 \pi_1 + x_2 y_2 \pi_2]\sigma_2 + x_3 y_3 \pi_3)\sigma_3 + x_4 y_4 \pi_4]\sigma_4$.

Allmänt gäller tydligen, om vi inför $\sigma_{j:k} = \sigma_j \sigma_{j+1} \sigma_k$, att:

$$fl(x_1 y_1 + x_2 y_2 + ... + x_n y_n) = x_1 y_1 \pi_1 \sigma_{2:n} + x_2 y_2 \pi_2 \sigma_{2:n} + x_3 y_3 \pi_3 \sigma_{3:n} + ... + x_k y_k \pi_k \sigma_{k:n} + ... + x_n y_n \pi_n \sigma_n.$$

Med våra antaganden om flyttalsaritmetik kan vi skriva detta

$$fl(\sum_{k=1}^{n} x_k y_k) = x_1 y_1(1 + n\epsilon_1') + \sum_{k=2}^{n} x_k y_k(1 + (n - k + 2)\epsilon_k')$$

där alla $|\epsilon_k'| \leq \epsilon_{mach}$. Om vi t.ex. sätter $\tilde{x}_k = x_k\sqrt{1 + (n - k + 2)\epsilon_k'}$ och $\tilde{y}_k = y_k\sqrt{1 + (n - k + 2)\epsilon_k'}$ så gäller tydligen att:

$$fl(\sum_{k=1}^{n} x_k y_k) = \sum_{k=1}^{n} \tilde{x}_k \tilde{y}_k$$

Om då $n\epsilon_{mach}$ är tillräckligt litet så är den beräknade skalärprodukten en exakt skalärprodukt av näraliggande värden och algoritmen är stabil.

18. Vi har ett flyttalsystem med basen $\beta = 10$, precision $t = 4$, och exponentomfång $L = -10$ och $U = 10$. Vilket är det största respektive minsta positiva talet i detta system?
a) Om systemet är normaliserat?
b) Om vi tillåter denormaliserade tal?
Lösning:
Flyttal (tal med flytande decimalpunkt):

$$x = \pm \left(d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + ... + \frac{d_{t-1}}{\beta^{t-1}} \right) \beta^e,$$

var

$$0 \leq d_k \leq \beta - 1, L \leq e \leq U,$$

Största talet i båda fallen är: $\underbrace{9.999}_{4 ciffror} \cdot 10^{10}$.

Minsta normaliserade talet är $\underbrace{1.000}_{4 ciffror} \cdot 10^{-10}$ och det minsta

denormaliserade talet är $\underbrace{0.001}_{4 ciffror} \cdot 10^{-10}$.

Denormaliserade flyttal används bara (i IEEE-standarden) kring nollan och inte kring största/minsta tal. Det är därför det största talet i b)-uppgiften är samma som i a)-uppgiften. De största talet är i båda deluppgifterna normaliserat.

20. Vi har ett flyttalsystem med basen 10, precision t och exponentomfång $[L, U]$. a) Vilka är de minsta värdena på t, U och det största på L så att både 2365.27 och 0.0000512 kan representeras exakt i normaliserad form? b) Om vi tillåter denormaliserade tal (denormaliserade eller subnormala tal: offra "decimaler" för att få större exponentomfång) ?

Lösning:

Flyttal (tal med flytande decimalpunkt):

$$x = \pm \left(d_0 + \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + ... + \frac{d_{t-1}}{\beta^{t-1}} \right) \beta^e,$$

var

$$0 \leq d_k \leq \beta - 1, L \leq e \leq U,$$

a) 2365.27 normaliserat blir $\underbrace{2.36527}_{6 ciffror} \cdot 10^3$ med $t = 6$ och $U = 3$.

$0.0000512 = 5.12 \cdot 10^{-5}$ så $L = -5$.

b) Med $t = 6$ får vi $0.0000512 = \underbrace{0.00512}_{6 ciffror} \cdot 10^{-2}$ så $L = -2$.

23.
Vi vill lösa ekvationen $x^2 + ax + b = 0$ då vi vet att a och b båda är positiva och där a är mycket större än b, $a >> b$. Den matematiska formeln inte fungerar tillfredsställande när vi räknar med avrundningsfel. Visa att rötterna är välkonditionerade genom att uppskatta konditionstalen med formeln som vi härledde på föreläsning 1 (det finns en stor rot (mycket negativ) och en liten (nära noll)). Visa att den stora roten går bra att beräkna med standardformeln, men att det blir problem med den lilla. Försök att hitta en bra algoritm för den lilla roten. Taylorutveckling är, som oftast, ett användbart redskap i detta sammanhang.
Lösning:
Låt oss kalla den stora (negativa) roten R och den lilla, nära noll, r. Standardformeln och Taylorutveckling ger:

$$R = -\frac{a}{2} - \sqrt{\frac{a^2}{4} - b} = -\frac{a}{2}\left[1 + \sqrt{1 - \frac{4b}{a^2}}\right] = -\frac{a}{2}\left[2 - \frac{2b}{a^2} - \frac{2b^2}{a^4} - ...\right] \approx a,$$

$$r = -\frac{a}{2} + \sqrt{\frac{a^2}{4} - b} = \frac{a}{2}\left[-1 + \sqrt{1 - \frac{4b}{a^2}}\right] = \frac{a}{2}\left[-\frac{2b}{a^2} - \frac{2b^2}{a^4} - ...\right] \approx -\frac{b}{a}$$

Vi kan uppskatta konditionstalen enligt formeln som vi härledde på föreläsning 1 ($a >> b$):

$$k_R = \frac{|a| + |b/R|}{|R - r|} \approx \frac{a + b/a}{a} \approx 1,$$
$$k_r = \frac{|a| + |b/r|}{|R - r|} \approx \frac{a + b/(b/a)}{a} \approx 2.$$

När vi beräknar $r = -\frac{a}{2} + \sqrt{\frac{a^2}{4} - b}$ kommer att få utskiftning av b. I det mest extrema fallet kommer inte b alls med och approximationen blir noll. Hur skall vi beräkna r? Ett sätt är att använda utvecklingen ovan:

$$r = -\frac{b}{a} - \frac{b^2}{a^3} - \frac{2b^3}{a^5} \ldots$$

Ett standardtrick är att förlänga med konjugatet,

$$r = \frac{\left(-\frac{a}{2} + \sqrt{\frac{a^2}{4} - b}\right)\left(-\frac{a}{2} - \sqrt{\frac{a^2}{4} - b}\right)}{-\frac{a}{2} - \sqrt{\frac{a^2}{4} - b}} = \frac{b}{-\frac{a}{2} - \sqrt{(\frac{a}{2})^2 - b}}.$$

Ytterligare ett sätt, är att göra en transformation så att r blir en dominant rot i det transformerade problemet. Sätt $y = 1/x$ (så att $r \to 1/r$). Ekvationen $x^2 + ax + b = 0$ övergår då till $y^2 + (a/b)y + 1/b = 0$. Om vi använder standardformeln får vi för den sökta roten:

$$\frac{1}{r} = -\frac{a}{2b} - \sqrt{\frac{a^2}{4b^2} - \frac{1}{b}}.$$

25. Låt $f(x) = (e^x - 1)/x$. Vi vet att $f(x) \to 1$ då $x \to 0$.
a) Bevisa detta genom att beräkna $f(10^{-k}), k = 1, ..., 16$.
b) Ge kommandot i MATLAB: *help expm1*

Lösning:

Låt oss studera trunkeringsfelet. Dvs. vad är skillnaden mellan gränsvärdet 1 och $f(x) = (e^x - 1)/x$ för $x \approx 0$. Taylorutveckling $F(x) = F(x_0) + F'(x_0)(x - x_0) + F''(x_0)(x - x_0)^2/2 + ...$ för $F(x) = e^x$ och $x_0 = 0$ ger:

$$\frac{e^x - 1}{x} = \frac{1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ... - 1}{x} = 1 + \frac{x}{2!} + \frac{x^2}{3!} + ...$$

Sa trunkeringsfelet ar ungefar $\frac{|x|}{2!}$ för små x.
Nu till avrundningsfelet. Vi antar att $fl(e^x) = e^x(1 + \epsilon)$ med $|\epsilon| \leq \epsilon_{mach}$.
Då gäller

$$fl\left[\frac{e^x - 1}{x}\right] = \frac{e^x(1 + \epsilon_1) - 1}{x}(1 + \epsilon_2)(1 + \epsilon_3) = \left[\frac{e^x - 1}{x} + \epsilon_1 \frac{e^x}{x}\right](1 + \epsilon_2)(1 + \epsilon_3)$$

Avrundningsfelet är:

$$\left| fl\left[\frac{e^x - 1}{x}\right] - \frac{e^x - 1}{x} \right| \leq \left[2 + \frac{1}{|x|}\right] \epsilon_{mach}.$$

Notera att $|x|$ i nämnaren! $\epsilon_{mach} \approx 1.11 \cdot 10^{-16}$ i dubbel precision.

b) Ge kommandot i MATLAB: *help expm1*
help expm1
expm1 Compute EXP(X)-1 accurately. expm1(X) computes EXP(X)-1, compensating for the roundoff in EXP(X).
For small real X, expm1(X) should be approximately X, whereas the computed value of EXP(X)-1 can be zero or have high relative error.

26. Vi vill approximera $f'(x)$ med differenskvoten, $(f(x+h)-f(x))/h$. Vad är ett lämpligt värde för h? Vad gäller om vi använder approximationen $f'(x) \approx (f(x+h)-f(x-h))/(2h)$?
Lösning:
Diskretiseringsfelet erhålles via Taylorutveckling:

$$\frac{f(x+h)-f(x)}{h} = \frac{f(x)+hf'(x)+h^2f''(x)/2+\ldots-f(x)}{h}$$
$$= f'(x)+\frac{h}{2}f''(x)+\ldots$$

Om vi antar att $|f''(x)| < M$ då trunkeringsfel är begränsad med $\frac{Mh}{2}$.
När vi uppskattar avrundningsfelet antar vi (för att förenkla) att $x+h$ beräknas exakt (se dock nedan). Dessutom antar vi att $fl(f(x)) = f(x)(1+\epsilon_k), |\epsilon_k| \leq \epsilon_{mach}$ (vilket kan vara orealistiskt om f är en komplicerad funktion).

$$fl\left[\frac{f(x+h)-f(x)}{h}\right] = \frac{f(x+h)(1+\epsilon_1)-f(x)(1+\epsilon_2)}{h}(1+\epsilon_3)(1+\epsilon_4)$$
$$= ... = \frac{f(x+h)-f(x)}{h} + 3\frac{f(x+h)\epsilon_5 - f(x)\epsilon_6}{h}.$$

Antar vi dessutom att $f(x) \approx f(x+h)$ får vi uppskattningen:

$$\left| fl\left[\frac{f(x+h)-f(x)}{h}\right] - \frac{f(x+h)-f(x)}{h} \right| \leq \left|\frac{f(x)}{h}\right| 6\epsilon_{mach}.$$

Det totala felet e_{total} får vi om vi adderar de två felen:

$$e_{total} \leq \underbrace{\left|\frac{f(x)}{h}\right| 6\epsilon_{mach}}_{avrundningsfelet} + \underbrace{\frac{Mh}{2}}_{diskretiseringsfelet}.$$

Tar vi $|h|$ för litet dominerar avrundningsfelet och om vi tar ett för stort $|h|$ dominerar diskretiseringsfelet.

Övning:
Beräkna diskretiseringsfelet och avrundningsfelet för approximationen $f'(x) \approx (f(x+h) - f(x-h))/(2h)$.
Använd föreläsning 2.