Övningar MMG410: konditionstal, stabilitet

Larisa Beilina, e-mail: larisa.beilina@chalmers.se

L. Beilina
Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Gothenburg, Sweden, e-mail: larisa.beilina@chalmers.se

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

   Lösning:

   I vårt exempel absoluta felet ≤ 0.0005, det är en halv enhet i sista decimalen, eller en halv enhet i fjärde siffran för $x = 24.\underbrace{516}_{3\ siffror}$.
   Absoluta felet räknas: $e_{abs} = |x - \hat{x}|,\ |x - \hat{x}| \leq 0.0005$, och relativa felet räknas som
   $$e_{rel} = \frac{|x - \hat{x}|}{|x|} = \frac{0.0005}{24.516} \approx 2 \cdot 10^{-5}.$$

2. Hur känsliga är rötterna, till ekvationen $x^2 + ax + b = 0$, för ändringar i a och b? Rötterna r_1 och r_2 är funktioner av a och b: $r_1(a,b), r_2(a,b)$. Låt $r = (r_1, r_2)$ beteckna en av rötterna och låt $r + \delta r$ beteckna den störda roten när vi ändrar koefficienterna med δa respektive δb.

   Lösning:

   Vi har sambandet:
   $$x^2 + ax + b = (r + \delta r)^2 + (a + \delta a)(r + \delta r) + (b + \delta b) = 0,$$
   och vi kan skriva om den:
   $$(r^2 + ar + b) + (\delta r(2r + a) + \delta ar + \delta b) + ((\delta r)^2 + \delta a \delta r) = I_1 + I_2 + I_3 = 0,$$
   var
   $$\begin{aligned} I_1 &= (r^2 + ar + b) = 0, \\ I_2 &= (\delta r(2r + a) + \delta ar + \delta b) \approx 0, \\ I_3 &= ((\delta r)^2 + \delta a \delta r) \approx 0. \end{aligned} \tag{1}$$
   Från andra ekvation i systemet (1) får vi:
   $$\delta r \approx -\frac{(\delta a\, r + \delta b)}{2r + a}$$
   eller
   $$|\delta r| \leq \frac{(|\delta a\, r| + |\delta b|)}{|2r + a|} \tag{2}$$
   Eftersom r_1 och r_2 är rötter så gäller att:
   $$(x - r_1)(x - r_2) = x^2 - (r_1 + r_2)x + r_1 r_2 = x^2 + ax + b$$
   Vi kan jämföra koefficienterna och får

$$-(r_1+r_2)=a,\ b=r_1r_2.$$

Vi kan skriva om $r_1-r_2=2r_1+a$, och definera gapet $g:=|r_1-r_2|$.
Vi kan skriva om (2)

$$|\delta r|\leq\frac{|\delta a\, r|+|\delta b|}{|g|} \tag{3}$$

Vi ser att om g är liten eller $r_1\approx r_2$, då $|\delta r|$ är stort.
Dividera (3) med $|r|$ och förläng med $|a|$ respektive $|b|$.

$$\frac{|\delta r|}{|r|}\leq\frac{1}{|r|}\left(\frac{\frac{|a|}{|a|}|\delta a\, r|+\frac{|b|}{|b|}|\delta b|}{|g|}\right)\leq k\max\left(\frac{|\delta a|}{|a|},\frac{|\delta b|}{|b|}\right), \tag{4}$$

var uppskattning av konditionstalet är

$$k\approx\frac{|a|+|b/r|}{g}. \tag{5}$$

3. I övningen ovan härledde vi en uppskattning (5) för konditionstalet för nollställena till ett andragradspolynom. Testa uppskattningen på

$$p(x)=x^2-3x+2$$

respektive

$$p(x)=x^2-1.99x+0.99$$

Stämmer den bra?

Lösning:

Andragradspolynom $p(x)=x^2-3x+2$ har nollställena $r_1=1$ och $r_2=2$ varför uppskattningarna av konditionstalen $\kappa_{1,2}$ för polynomen på formen $p(x)=x^2+ax+b$ blir:

$$\kappa_1\approx\frac{|a|+|b/r_1|}{g}=\frac{|-3|+|2/1|}{|2-1|}=5,$$
$$\kappa_2\approx\frac{|a|+|b/r_2|}{g}=\frac{|-3|+|2/2|}{|2-1|}=4.$$

Tag $\delta a=3\cdot10^{-4},\delta b=2\cdot10^{-4}$ i uppskattningen (4). Då är $|\delta a|/|a|=|\delta b|/|b|=10^{-4}$. Vi får då

$$\begin{aligned}\frac{|\delta r_1|}{|r_1|} &\le k_1 \max\left(\frac{|\delta a|}{|a|},\frac{|\delta b|}{|b|}\right) \approx 5\max\left(10^{-4},10^{-4}\right) = 5\cdot 10^{-4},\\ \frac{|\delta r_2|}{|r_2|} &\le k_2 \max\left(\frac{|\delta a|}{|a|},\frac{|\delta b|}{|b|}\right) \approx 4\max\left(10^{-4},10^{-4}\right) = 4\cdot 10^{-4}\end{aligned} \tag{6}$$

Polynomet $p(x) = x^2 - 1.99x + 0.99$ har nollställena $r_1 = 0.99$ och $r_2 = 1$ varför uppskattningarna av konditionstalen $\kappa_{1,2}$ för polynomen på formen $p(x) = x^2 + ax + b$ blir:

$$\begin{aligned}\kappa_1 &\approx \frac{|a| + |b/r_1|}{g} = \frac{|-1.99| + |0.99/0.99|}{|0.99-1|} = 299,\\ \kappa_2 &\approx \frac{|a| + |b/r_2|}{g} = \frac{|-1.99| + |0.99/1|}{|0.99-1|} = 298.\end{aligned}$$

Tag $\delta a = 2\cdot 10^{-4}, \delta b = 10^{-4}$ i uppskattningen (4). Då är $|\delta a|/|a| = |\delta b|/|b| \approx 10^{-4}$. Vi får då

$$\begin{aligned}\frac{|\delta r_1|}{|r_1|} &\le k_1 \max\left(\frac{|\delta a|}{|a|},\frac{|\delta b|}{|b|}\right) \approx 299\max\left(10^{-4},10^{-4}\right) = 299\cdot 10^{-4},\\ \frac{|\delta r_2|}{|r_2|} &\le k_2 \max\left(\frac{|\delta a|}{|a|},\frac{|\delta b|}{|b|}\right) \approx 298\max\left(10^{-4},10^{-4}\right) = 298\cdot 10^{-4}.\end{aligned} \tag{7}$$

Detta är för stora störningar för att satsen skall fungera bra.

4. Låt $\hat{x}$ är en approximation av ett exakt värde x där $|x-\hat{x}| \le \delta$. Hur kan vi uppskatta $|f(x) - f(\hat{x})|$ givet funktionen f ?
Vi känner 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 $e_{abs} = |x-\hat{x}| \le \delta$ eller $-\delta \le x - \hat{x} \le \delta$ och då $\hat{x} - \delta \le x \le \hat{x} + \delta$. 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

$$|f(x) - f(\hat{x})| \le \delta \max_{\xi\in(\hat{x}-\delta,\hat{x}+\delta)} |f'(\xi)|.$$

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: $|f(x) - f(\hat{x})| \le 7\delta$.
2) I andra fallet får vi begränsningen $|f(x) - f(\hat{x})| \le 2\delta(|\hat{x}| + \delta)$ eftersom derivatan, $f'(x) = (x^2)'_x = 2x$, är strängt växande:

$$|f(x) - f(\hat{x})| \le \delta \max_{\xi\in(\hat{x}-\delta,\hat{x}+\delta)} |f'(\xi)| = \delta \max_{\xi\in(\hat{x}-\delta,\hat{x}+\delta)} |2\xi| = \delta\cdot 2(|\hat{x}| + \delta).$$

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 $dåx=\pi/2-\delta$, med litet $\delta>0$.

   Lösning:

   Frågan är alltså: hur ändras $f(x)$ när vi ändrar x lite? Taylors formel ger: $f(x+\delta x)=f(x)+\delta x f'(x)+....$ 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:

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

   där κ är en uppskattning av konditionstalet. Då $f(x)=\cos x$ får vi:

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

   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):

$$|(f(x)+\delta(x))' - f'(x)| = |f'(x)+\delta'(x) - f'(x)|$$
$$= |(\varepsilon\cos(\omega x)'_x| = |\omega\varepsilon(-sin\omega x)|.$$

En liten störning av f kan alltså ändra derivatan godtyckligt mycket.
2) Detta gäller inte integralen. Vi har ju:

$$\left|\int_a^b (f(x)+\delta(x))\,dx - \int_a^b f(x)\,dx\right| = \left|\int_a^b \delta(x)\,dx\right| = \left|\int_a^b \varepsilon\cos(\omega x)\,dx\right|$$
$$= \left|\frac{\varepsilon}{\omega}\sin(\omega x)|_a^b\right| = \left|\frac{\varepsilon}{\omega}(\sin(\omega b) - \sin(\omega a))\right|$$

så att

$$\lim_{\omega\to\infty}\left|\frac{\varepsilon}{\omega}(\sin(\omega b) - \sin(\omega a))\right| = 0.$$

7. 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 $|\alpha| \approx 1$.

8. 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 $|\alpha| \approx |\beta|$.

9. 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:

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

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:

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

10. 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:

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

Så trunkeringsfelet är ungefär $\frac{|x|}{2!}$ för små x.
Nu till avrundningsfelet. Vi antar att $fl(e^x) = e^x(1+\varepsilon)$ med $|\varepsilon| \le \varepsilon_{mach}$. Då gäller

$$fl\left[\frac{e^x-1}{x}\right] = \frac{e^x(1+\varepsilon_1)-1}{x}(1+\varepsilon_2)(1+\varepsilon_3) = \left[\frac{e^x-1}{x}+\varepsilon_1\frac{e^x}{x}\right](1+\varepsilon_2)(1+\varepsilon_3)$$

Avrundningsfelet är:

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

Notera att $|x|$ i nämnaren! $\varepsilon_{mach} \approx 1.11 \cdot 10^{-16}$ i dubbel precision.
b) Ge kommandot *help expm1* i MATLAB:

```
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.
```

11. Vi vill approximera $f'(x)$ med differenskvoten, $(f(x+h)-f(x))/h$. Vad är ett lämpligt värde för h?

    Lösning:

    Diskretiseringsfelet erhålles via Taylorutveckling:

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

    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 $fl(f(x)) = f(x)(1+\varepsilon_k), |\varepsilon_k| \leq \varepsilon_{mach}$ (vilket kan vara orealistiskt om f är en komplicerad funktion).

$$fl\left[\frac{f(x+h)-f(x)}{h}\right] = \frac{f(x+h)(1+\varepsilon_1)-f(x)(1+\varepsilon_2)}{h}(1+\varepsilon_3)(1+\varepsilon_4)$$
$$= ... = \frac{f(x+h)-f(x)}{h} + 3\frac{f(x+h)\varepsilon_5 - f(x)\varepsilon_6}{h}.$$

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

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

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

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

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

12. Beräkna diskretiseringsfelet för approximationen $f'(x) \approx (f(x+h)-f(x-h))/(2h)$.

    Lösning:

    Approximativt värde (Taylor's theorem):

$$(*)\ f(x+h) = f(x) + f'(x)h + \frac{f''(x)h^2}{2!} + \frac{f'''(Q)h^3}{3!},$$

$$(**)\ f(x-h) = f(x) - f'(x)h + \frac{f''(x)h^2}{2!} - \frac{f'''(Q)h^3}{3!}.$$

$(*) - (**)$:

$$f(x+h) - f(x-h) = 2f'(x)h + 2\frac{f'''(Q)}{3!}h^3,$$

$$2f'(x)h = f(x+h) - f(x-h) - 2\frac{f'''(Q)}{3!}h^3,$$

som kan skrivas om

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{2f'''(Q)h^3}{3! \cdot 2h}.$$

Trunkeringsfel ε:

$$\varepsilon = \frac{2f'''(Q)h^3}{3! \cdot 2h} = \frac{f(x+h) - f(x-h)}{2h} - f'(x).$$

Låt $M \leq |f'''(Q)|$, då trunkeringsfel, eller diskretiseringsfel, ε är begränsad med

$$\varepsilon < \frac{Mh^2}{6}.$$