Övningar MMG410

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 Övningar: flyttalsaritmetik

1. 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: $\underbrace{9.999}_{4\ siffror} \cdot 10^{10}$.

   Minsta normaliserade talet är $\underbrace{1.000}_{4\ siffror} \cdot 10^{-10}$ och det minsta denormaliserade talet är $\underbrace{0.001}_{4\ siffror} \cdot 10^{-10}$.

   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.
2. 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) För 2365.27 normaliserat blir $\underbrace{2.36527}_{6\ siffror} \cdot 10^3$ med $t = 6$ och $U = 3$. För $0.0000512 = 5.12 \cdot 10^{-5}$ så $L = -5$.
   b) Med $t = 6$ får vi $0.0000512 = \underbrace{0.00512}_{6\ siffror} \cdot 10^{-2}$ så $L = -2$.

3. Skriv talet 6 i binär (bas 2) som flyttal i enkel precision i dator.

   Lösning:

$$6: \ [1.5]\cdot 2^2 = \underbrace{[1+0.5]}_{mantissa}\cdot 2^2$$

   Exponenten $e = 2$ lagras som: $2+127 = 129 = 2^7 + 2^0$. Mantissa: 1 kodas inte,

$$0.5 = \frac{1}{2}x_1 + \frac{1}{4}x_2 + \frac{1}{8}x_3 + \frac{1}{16}x_4 + \ldots = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 0 + \ldots$$

   Resten i mantissa ska vara 0.
   Vi får följande binär (bas 2) representation för 6 i enkel precision:

0	10000001	10 0
tecken	exponent 8 bitar	mantissa 23 bitar

4. Skriv talet -3.25 i binär (bas 2) och i hexadecimalt (bas 16) form, som flyttal i dubbel precision i dator.

   Lösning:

$$-3.25: \ -[1.625]\cdot 2^1 = -\underbrace{[1+0.625]}_{mantissa}\cdot 2^1$$

   Exponenten $e = 1$ lagras som: $1+1023 = 1024 = 2^{10}$. Mantissa: 1 kodas inte,

$$0.625 = \frac{1}{2}x_1 + \frac{1}{4}x_2 + \frac{1}{8}x_3 + \frac{1}{16}x_4 + \ldots = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 0 + \frac{1}{8}\cdot 1 + \ldots$$

   $\frac{1}{2} = 0.5 < 0.625$; $\quad \frac{1}{4} = 0.25 : 0.625 - 0.5 = 0.125; 0.25 > 0.125$, darför $\frac{1}{4}\cdot 0$,
   $\frac{1}{8} = 0.125 = 0.125$ och darför $\frac{1}{8}\cdot 1$, resten i mantissa ska vara 0.
   Vi får följande binär (bas 2) representation för -3.25 :

1	10000000000	1010 0
tecken	exponent 11 bitar	mantissa 52 bitar

   Bas 16:

0 1 2 3 4 5 6 7 8 9	10	11	12	13	14	15
	a	b	c	d	e	f

   Nu grupperar vi om binär form för -3.25 i 4 bitar:

1100	0000	0000	1010	0000	....	0000

   och kodar första fyra bitar:

$\boxed{1100}$ = c
eftersom
$1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 0 \cdot 2^0 = 12 = c$
0000 koderas som 0, och sedan

$\boxed{1010}$ = a
eftersom

$1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0 = 10 = a.$
Slutsats: -3.25 i hexadecimalt (bas 16) form lagras som:

c00a000000000000

5. Skriv talet -9.28 i binär (bas 2) form som flyttal i dubbel precision i dator.

Lösning:

$-9.28 := -[1.16] \cdot 2^3 = -[1 + 0.16] \cdot 2^3$
Exponenten 3 lagras som:

$$3 + 1023 = 1026 = 1024 + 2 = 1 \cdot 2^{10} + 1 \cdot 2^1 + 0 \cdot 2^0$$

I mantissan 1 kodas inte, vi kodar bara 0.16:

$$0.16 = \frac{1}{2}x_1 + \frac{1}{4}x_2 + \frac{1}{8}x_3 + \frac{1}{16}x_4 + \frac{1}{32}x_5 + \ldots =$$
$$= \frac{1}{2} \cdot 0 + \frac{1}{4} \cdot 0 + \frac{1}{8} \cdot 1 + \frac{1}{16} \cdot 0 + \ldots$$

Förklaringen hur kodas mantissa:
$\frac{1}{2} = 0.5 > 0.16$, darför $\frac{1}{2} \cdot 0$, $\frac{1}{4} = 0.25 : 0.25 > 0.16$, darför $\frac{1}{4} \cdot 0$, $\frac{1}{8} = 0.125 <$ 0.16 och darför $\frac{1}{8} \cdot 1$, $\frac{1}{16} = 0.0625 : 0.16 - 0.125 = 0.035, 0.0625 > 0.035$, och darför $\frac{1}{16} \cdot 0$, $\frac{1}{32} = 0.0312 : 0.0312 < 0.035$, och darför $\frac{1}{32} \cdot 1$, och så vidare ...
Vi observerar att det är inte tillräkligt 52 bitar i mantissan för att koda exakt 0.16. I andra ord, 0.16 kan inte presenteras exakt i binär form, darför binär form introducerar fel i lagringen av flyttal.

1	10000000010	00101 ....
tecken	exponent 11 bitar	mantissa 52 bitar

Kollar i Matlab:

```
q = quantizer('double');

y = num2bin(q,-9.28)

y =
```

```
1100000000010001010001111010111000010100011110101110000101000111
```

6. Skriv talet 1 i binär (bas 2) form som flyttal i dubbel precision i dator.

   Lösning:

   $1 := [1] \cdot 2^0 = [1+0.0] \cdot 2^0$
   Mantissa är 0 här.
   Exponenten är också 0 och lagras som:
   $0+1023 = 1024-1 = 1 \cdot 2^{10} - 1 \cdot 2^0 = \underbrace{10000000000}_{11\ bitar} - \underbrace{00000000001}_{11\ bitar} = \underbrace{01111111111}_{11\ bitar}$.

0	01111111111	0000 ....
tecken	exponent 11 bitar	mantissa 52 bitar

7. Skriv talet 0.5 i binär (bas 2) form som flyttal i dubbel precision i dator.

   Lösning:

   $0.5 := [1] \cdot 2^{-1} = [1+0.0] \cdot 2^{-1}$
   Mantissa är 0 här.
   Exponenten -1 lagras som:
   $-1+1023 = 1024-2 = 1 \cdot 2^{10} - 1 \cdot 2^1 = \underbrace{10000000000}_{11\ bitar} - \underbrace{00000000010}_{11\ bitar} = \underbrace{01111111110}_{11\ bitar}$.

0	01111111110	0000 ....
tecken	exponent 11 bitar	mantissa 52 bitar

8. Skriv talet 0.1 i binär (bas 2) form som flyttal i dubbel precision i dator.

   Lösning:

   $0.1 := [1.6] \cdot 2^{-4} = [1+0.6] \cdot 2^{-4}$
   Mantissa är 0.6 här. Det är inte tillräkligt 52 bitar i mantissan för att koda exakt 0.6, se nedan Matlab's presentation. Det betyder att 0.6 presenteras inte exakt i binär form.
   Exponenten -4 lagras som:
   $-4+1023 = 1019 = 1024-5 = 1 \cdot 2^{10} - (1 \cdot 2^2 + 1 \cdot 2^0) = \underbrace{10000000000}_{11\ bitar} - \underbrace{00000000101}_{11\ bitar} =$

   $\underbrace{01111111011}_{11\ bitar}$.

0	01111111011	100...
tecken	exponent 11 bitar	mantissa 52 bitar

   Kollar i Matlab:

```
q = quantizer('double');
```

```
y = num2bin(q,0.1)


  y =

  0011111110111001100110011001100110011001100110011001100110011010
```

9. 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: $|x - \hat{x}| = |1.0317 \cdot 10^{p+1} - 1.032 \cdot 10^{p+1}| = 3 \cdot 10^{-4} \cdot 10^{p+1} = 3 \cdot 10^{p-3}$.
   - 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, $\approx 3 \cdot 10^{-4}$.

10. Visa att addition enligt IEEE är en stabil algoritm för maskinnoggrannheten ε_{mach}.

    Lösning:

    Vi vet att

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

    Alltså gäller att

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

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

11. 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 $\sigma_k = 1+\delta_k$, $\pi_k = 1+\varepsilon_k$ (σ för summa och π för produkt) där alla $|\varepsilon_k| < \varepsilon_{mach}$ och $|\delta_k| < \varepsilon_{mach}$. Vi har:

- $fl(x_1y_1) = x_1y_1\pi_1$ (vi använder inte σ_1)
- $fl(x_1y_1 + x_2y_2) = [x_1y_1\pi_1 + x_2y_2\pi_2]\sigma_2$
- $f(x_1y_1 + x_2y_2 + x_3y_3) = ([x_1y_1\pi_1 + x_2y_2\pi_2]\sigma_2 + x_3y_3\pi_3)\sigma_3$.
- $fl(x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4) = [([x_1y_1\pi_1 + x_2y_2\pi_2]\sigma_2 + x_3y_3\pi_3)\sigma_3 + x_4y_4\pi_4]\sigma_4$.

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

$$fl(x_1y_1 + x_2y_2 + ... + x_ny_n) = x_1y_1\pi_1\sigma_{2:n} + x_2y_2\pi_2\sigma_{2:n} + x_3y_3\pi_3\sigma_{3:n} + ... + x_ky_k\pi_k\sigma_{k:n} + ... + x_ny_n\pi_n\sigma_n.$$

Med våra antaganden om flyttalsaritmetik kan vi skriva detta

$$fl(\sum_{k=1}^{n} x_ky_k) = x_1y_1(1+n\varepsilon_1') + \sum_{k=2}^{n} x_ky_k(1+(n-k+2)\varepsilon_k')$$

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

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

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

2 Övningar: konditionstal, stabilitet

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: $e_{abs} = |x-\hat{x}|$, $|x-\hat{x}| \le 0.0005$, 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:

$$(x-r_1)(x-r_2)=x^2-(r_1+r_2)x+r_1r_2=x^2+ax+b$$

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

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

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:

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

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

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

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

4. Låt $\hat{x}$ är en approximation av ett exakt värde x där $|x - \hat{x}| \leq \delta$. Hur kan vi uppskatta $|f(x) - f(\hat{x})|$ 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: $|f(x) - f(\hat{x})| \leq 7\delta$.
   2) I andra fallet får vi begränsningen $|f(x) - f(\hat{x})| \leq 2\delta(|\hat{x}| + \delta)$ 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 *då* $x = \pi/2 - \delta$, med litet $\delta > 0$.

   Lösning:

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

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

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) Beräkna trunkeringsfel och avrundningsfel för $|f(x) - 1|$ för $x \to 0$.
b) Ge kommandot i MATLAB:

```
help expm1
```

Lösning:

Låt oss studera trunkeringsfelet för $|f(x) - 1|$. 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:

$$\frac{e^x - 1}{x} = \frac{1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ... - 1}{x} = 1 + \frac{x}{2!} + \frac{x^2}{3!} + ... \tag{8}$$

Från (8) följer att trunkeringsfelet är ungefär

$$\left| \frac{e^x - 1}{x} - 1 \right| \approx \frac{|x|}{2!}$$

för små x.
Nu till avrundningsfelet. Vi antar att $fl(e^x) = e^x(1 + \varepsilon)$ med $|\varepsilon| \leq \varepsilon_{mach}$. Då gäller

$$\begin{aligned} fl\left[\frac{e^x - 1}{x}\right] &= \frac{e^x(1 + \varepsilon_1) - 1}{x}(1 + \varepsilon_2)(1 + \varepsilon_3) = \left[\frac{e^x - 1}{x} + \varepsilon_1 \frac{e^x}{x}\right](1 + \varepsilon_2)(1 + \varepsilon_3) \\ &\leq \left[\frac{e^x - 1}{x} + \varepsilon_{mach} \frac{e^x}{x}\right](1 + 2\varepsilon_{mach} + \varepsilon_{mach}^2) \approx \left[\frac{e^x - 1}{x} + \varepsilon_{mach} \frac{e^x}{x}\right](1 + 2\varepsilon_{mach}) \\ &= \left[\frac{e^x - 1}{x} + \varepsilon_{mach} \frac{e^x}{x}\right] + 2\varepsilon_{mach}\left[\frac{e^x - 1}{x} + \varepsilon_{mach} \frac{e^x}{x}\right] \\ &\approx \frac{e^x - 1}{x} + \varepsilon_{mach} \frac{e^x}{x} + 2\varepsilon_{mach} \frac{e^x - 1}{x}. \end{aligned}$$

Avrundningsfelet för $x \to 0$ är:

$$\left| fl\left[\frac{e^x - 1}{x}\right] - \frac{e^x - 1}{x} \right| \leq \left[3 + \frac{2}{|x|}\right] \varepsilon_{mach}$$

Notera att $|x|$ i nämnaren! $\varepsilon_{mach} \approx 1.11 \cdot 10^{-16}$ i dubbel precision.
Man kan analysera i Matlab trunkeringsfel, avrundningsfel och gemensamt trunkeringsfel + avrundningsfel:

```
>>x = 10.^-(1:10)'
```

```
>>limit = (exp(x) - 1) ./ x

>>sprintf('%10.5e %10.5e  %10.5e %10.5e %10.5e\n',  ...

 [sort(x), limit, x/2, (3 + 2 ./ x) * eps, x/2 + (3 + 2 ./ x) * eps]')

ans =

     1.00000e-10 1.05171e+00  5.00000e-02 5.10703e-15 5.00000e-02
     1.00000e-09 1.00502e+00  5.00000e-03 4.50751e-14 5.00000e-03
     1.00000e-08 1.00050e+00  5.00000e-04 4.44755e-13 5.00000e-04
     1.00000e-07 1.00005e+00  5.00000e-05 4.44156e-12 5.00000e-05
     1.00000e-06 1.00001e+00  5.00000e-06 4.44096e-11 5.00004e-06
     1.00000e-05 1.00000e+00  5.00000e-07 4.44090e-10 5.00444e-07
     1.00000e-04 1.00000e+00  5.00000e-08 4.44089e-09 5.44409e-08
     1.00000e-03 1.00000e+00  5.00000e-09 4.44089e-08 4.94089e-08
     1.00000e-02 1.00000e+00  5.00000e-10 4.44089e-07 4.44589e-07
     1.00000e-01 1.00000e+00  5.00000e-11 4.44089e-06 4.44094e-06
```

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:

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

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 $fl(f(x)) = f(x)(1+\varepsilon_k), |\varepsilon_k| \leq \varepsilon_{mach}$ (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

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

3 Övningar: linjära ekvationssystem

1. a) Visa att matrisen A är singulär.

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 2 \end{bmatrix}$$

b) Hur många lösningar har systemet $Ax = [2,4,6]^T$?

Lösning: a) $A[1,-1,1]^T = \mathbf{0} \rightarrow \det(A) = 0$.

b) In Matlab:

```
>> b/A
Warning: Matrix is singular to working precision.
ans =

   NaN   Inf  -Inf
```

2. Beräkna A^{-1} då

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 1 & -2 & 1 \end{bmatrix}$$

Lösning:

Inversen beräknas normalt med LU-faktorisering. Beteckna inversen med X, s.a. $AX = I$. Kolonnvis får vi $A\mathbf{x}_k = \mathbf{e}_k$, där $\mathbf{x}_k$ och $\mathbf{e}_k$ är kolonn k i X resp. I. Vi har n linjära ekvationssystem att lösa. A är triangulär vilket förenklar lösningsprocessen. Vi kan i detta specialfall beräkna inversen med tre framåt substitutioner.
Problemet kan även lösas via ansats. En triangulär matris har en triangulär invers (om den existerar) s.a. $(A^{-1})_{k,k} = 1/a_{k,k}$. Ansatsen ger

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ \alpha & -1 & 0 \\ \beta & \gamma & 1 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 1 & -2 & 1 \end{bmatrix}}_{A}$$

vilket ger systemet (visa !)

$$\begin{cases} \alpha - 1 = 0 \\ \beta + \gamma + 1 = 0 \\ -\gamma - 2 = 0 \end{cases}$$

vilket ger $\alpha = \beta = 1$ och $\gamma = -2$.

3. A är kvadratisk med $A^2 = 0$ Visa att A är singulär.

   Lösning:

$$0 = \det(A^2) = (\det A)^2 \Rightarrow \det A = 0.$$

4. Antag att $A, B \in \mathbb{R}^{n \times n}$ Visa att: a) $(AB)^T = B^T A^T$ samt b) $(AB)^{-1} = B^{-1}A^{-1}$ (när A och B är ickesingulära).

   Lösning:

   a) Vi visar istället $(A^T B)^T = B^T A$. Detta är ekvivalent med det som efterfrågas om vi tar $C = A^T$. Partitionera matriserna kolonnvis $A = [\mathbf{a_1}, ..., \mathbf{a_n}]$ och $B = [\mathbf{b_1}, ..., \mathbf{b_n}]$. Vi får:

$$A^T B = \begin{bmatrix} \mathbf{a_1}^T \\ \mathbf{a_2}^T \\ \vdots \\ \mathbf{a_n}^T \end{bmatrix} \begin{bmatrix} \mathbf{b_1} & \mathbf{b_2} & \cdots & \mathbf{b_n} \end{bmatrix} = \begin{bmatrix} \mathbf{a_1}^T\mathbf{b_1} & \mathbf{a_1}^T\mathbf{b_2} & \cdots & \mathbf{a_1}^T\mathbf{b_n} \\ \mathbf{a_2}^T\mathbf{b_1} & \mathbf{a_2}^T\mathbf{b_2} & \cdots & \mathbf{a_2}^T\mathbf{b_n} \\ \vdots & \vdots & & \vdots \\ \mathbf{a_n}^T\mathbf{b_1} & \mathbf{a_n}^T\mathbf{b_2} & \cdots & \mathbf{a_n}^T\mathbf{b_n} \end{bmatrix}$$

   Transponatet av ovanstående blir

$$\begin{bmatrix} \mathbf{a_1}^T\mathbf{b_1} & \mathbf{a_2}^T\mathbf{b_1} & \cdots & \mathbf{a_n}^T\mathbf{b_1} \\ \mathbf{a_1}^T\mathbf{b_2} & \mathbf{a_2}^T\mathbf{b_2} & \cdots & \mathbf{a_2}^T\mathbf{b_2} \\ \vdots & \vdots & & \vdots \\ \mathbf{a_1}^T\mathbf{b_n} & \mathbf{a_2}^T\mathbf{b_n} & \cdots & \mathbf{a_n}^T\mathbf{b_n} \end{bmatrix} = \begin{bmatrix} \mathbf{b_1}^T\mathbf{a_1} & \mathbf{b_1}^T\mathbf{a_2} & \cdots & \mathbf{b_1}^T\mathbf{a_n} \\ \mathbf{b_2}^T\mathbf{a_1} & \mathbf{b_2}^T\mathbf{a_2} & \cdots & \mathbf{b_2}^T\mathbf{a_n} \\ \vdots & \vdots & & \vdots \\ \mathbf{b_n}^T\mathbf{a_1} & \mathbf{b_n}^T\mathbf{a_2} & \cdots & \mathbf{b_n}^T\mathbf{a_n} \end{bmatrix}$$

   vilket är lika med $B^T A$. Likheten ovan följer av att $\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x}$.
   b) Det enklaste sättet är att multiplicera ihop matriserna och se att vi får enhetsmatrisen:

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AIA^{-1} = AA^{-1} = I$$

   Multiplikationen från andra hållet följer analogt.

5. A är ickesingulär. Visa att $(A^{-1})^T = (A^T)^{-1}$ Vi skriver därför A^{-T}.

   Lösning:

   Vi visar först att inversen är entydig. Om C är ickesingulär och $CX = I$, $CY = I$ följer $C(X - Y) = 0$, men C är ickesingulär så $X - Y = 0$.

Vidare är $A^T(A^T)^{-1} = I$ men det gäller även att $I = A^{-1}A = (A^{-1}A)^T = A^T(A^{-1})^T$. Det följer att $(A^{-1})^T = (A^T)^{-1}$.

6. Beskriv, i punktform, hur man löser systemet

$$\begin{bmatrix} L_1 & 0 \\ B & L_2 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix},$$

L_1 och L_2 är undertriangulära ickesingulära matriser. Vektorerna har partitionerats så att de passar ihop med blocken i matrisen.

Lösning:

Systemet är ekvivalent med

$$\begin{aligned} L_1\mathbf{x} &= \mathbf{b} \\ B\mathbf{x} + L_2\mathbf{y} &= \mathbf{c} \end{aligned}$$

Algoritm: Lös $L_1\mathbf{x} = \mathbf{b}$, bilda $\mathbf{t} = \mathbf{c} - B\mathbf{x}$ och lös slutligen $L_2\mathbf{y} = \mathbf{t}$.

7. a) Beräkna LU-faktoriseringen av matrisen nedan. b) När är matrisen singulär?

$$\begin{bmatrix} 1 & a \\ c & b \end{bmatrix}$$

Lösning:

$$A = \begin{bmatrix} 1 & a \\ c & b \end{bmatrix} = \underbrace{\begin{bmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{bmatrix}}_{L} \cdot \underbrace{\begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}}_{U} = \begin{bmatrix} u_{11}\ell_{11} & \ell_{11}u_{12} \\ \ell_{21}u_{11} & \ell_{21}\cdot u_{12} + \ell_{22}\cdot u_{22} \end{bmatrix}.$$

$$\ell_{11}u_{11} = 1, \ell_{11}u_{12} = a, \ell_{21}u_{11} = c, \ell_{21}\cdot u_{12} + \ell_{22}\cdot u_{22} = b, \ell_{11} = \ell_{22} = 1.$$

$$A = \begin{bmatrix} 1 & 0 \\ c & 1 \end{bmatrix} \begin{bmatrix} 1 & a \\ 0 & b - ca \end{bmatrix}$$

Singulär om $b = ca$.

8. Visa att en symmetrisk och positivt definit matris A har: a) positiva diagonalelement; b) ”stor diagonal”, $a_{j,j} + a_{k,k} > 2|a_{j,k}|$; c) det till beloppet största elementet på diagonalen; d) har positiva diagonalelement, i D, i LDL^T faktoriseringen.

Lösning:

Definition: $\mathbf{x}^T A\mathbf{x} > 0, \forall x \neq 0$.

a) Tag $\mathbf{x} = \mathbf{e}_k, \mathbf{e}_k^T A \mathbf{e}_k = a_{k,k}$.
b) Med $\sigma = -\text{sign}(a_{j,k})$ och $\mathbf{x} = \mathbf{e}_j + \sigma \mathbf{e}_k$ fås

$$\mathbf{x}^T A \mathbf{x} = a_{j,j} + 2\sigma a_{j,k} + a_{k,k} > 0 \Rightarrow \frac{a_{j,j} + a_{k,k}}{2} > |a_{j,k}|$$

$$\text{c) } |a_{j,k}| < \frac{a_{j,j} + a_{k,k}}{2} \leq \max(a_{j,j}, a_{k,k})$$

d) Eftersom A är positivt definit kan man visa att LDL^T-faktoriseringen alltid existerar (dvs. inget pivotelement kan bli noll). L är alltså ickesingulär och vi kan ta $\mathbf{x} = L^{-T}\mathbf{e}_k$ ($\mathbf{e}_k$ är kolonn k i I). x kan inte vara noll (varför?) och vi får

$$0 < \mathbf{x}^T A \mathbf{x} = [L^{-T}\mathbf{e}_k]^T L D L^T [L^{-T}\mathbf{e}_k] = \mathbf{e}_k^T D \mathbf{e}_k = d_{k,k}$$

9. Visa att matrisen nedan saknar LU-faktoriseringen:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Lösning:

Gör ansatsen

$$\begin{bmatrix} l_1 & 0 \\ l_2 & l_3 \end{bmatrix} \begin{bmatrix} u_1 & u_2 \\ 0 & u_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \Rightarrow \begin{cases} l_1 u_1 = 0 \\ l_1 u_2 = 1, \\ l_2 u_1 = 1 \\ l_2 u_2 + l_3 u_3 = 0 \end{cases}$$

$l_1 u_1 = 0 \Rightarrow l_1 = 0$ eller $u_1 = 0$, men då kan inte $l_1 u_2 = 1$ och $l_2 u_1 = 1$.

10. Använd Choleskyfaktorisering för att avgöra för vilka α följande matris är positivt definit.

$$A = \begin{bmatrix} \alpha & 1 \\ 1 & 2 \end{bmatrix}$$

Lösning:

Vi behöver antaga $\alpha \neq 0$ för första steget i Gausseleminationen:

$$\begin{bmatrix} \alpha & 1 \\ 1 & 2 \end{bmatrix} = \underbrace{\begin{bmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{bmatrix}}_{L} \cdot \underbrace{\begin{bmatrix} \ell_{11} & \ell_{21} \\ 0 & \ell_{22} \end{bmatrix}}_{L^T} = \begin{bmatrix} \ell_{11}^2; & \ell_{11} \cdot \ell_{21} \\ \ell_{21} \cdot \ell_{11}; & \ell_{21}^2 + \ell_{22}^2 \end{bmatrix}$$

$$\Rightarrow A = LL^T,$$

$$\text{L} = \begin{bmatrix} \sqrt{\alpha} & 0 \\ 1/\sqrt{\alpha} & \sqrt{2 - 1/\alpha} \end{bmatrix}$$

För att kunna dra roten ur diagonalen måste $\alpha > 0$ och $2 - 1/\alpha > 0$. Alltså $\alpha > 1/2$.

11. $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. När existerar $(I - \mathbf{u}\mathbf{v}^T)^{-1}$? Bestäm inversen när så är fallet (Den har nästan samma form som matrisen själv).

Lösning:

Om matrisen är singulär existerar $\mathbf{x} \neq 0$ så att $(I - \mathbf{u}\mathbf{v}^T)\mathbf{x} = \mathbf{0}$ dvs. $\mathbf{x} = \mathbf{u}(\mathbf{v}^T\mathbf{x})$, dvs. $\mathbf{x}$ måste vara parallell med $\mathbf{u}$. Tag $\mathbf{x} = \mathbf{u}$. Detta ger $\mathbf{u} = \mathbf{u}(\mathbf{v}^T\mathbf{u})$ dvs. $\mathbf{v}^T\mathbf{u} = 1$. Om $\mathbf{v}^T\mathbf{u} \neq 1$ är matrisen ickesingulär.

$$(I - \mathbf{u}\mathbf{v}^T)(I - \sigma\mathbf{u}\mathbf{v}^T) = I - \sigma\mathbf{u}\mathbf{v}^T - \mathbf{u}\mathbf{v}^T + \mathbf{u}\mathbf{v}^T\sigma\mathbf{u}\mathbf{v}^T = I - \mathbf{u}\mathbf{v}^T(1 + \sigma - \sigma\mathbf{v}^T\mathbf{u})$$

Detta är enhetsmatrisen om $\sigma = 1/(\mathbf{v}^T\mathbf{u} - 1)$ och $\mathbf{v}^T\mathbf{u} \neq 1$.

12. Visa att $\|\cdot\|_p,\ p = 1, 2, \infty$ verkligen är vektornormer.

Lösning:

$p = 1$:
1) $\|\mathbf{x}\|_1 = \sum_{k=1}^n |x_k| > 0$ om $\mathbf{x} \neq \mathbf{0}$.
2) $\|\gamma\mathbf{x}\|_1 = \sum_1^n |\gamma x_k| = |\gamma| \sum_1^n |x_k| = |\gamma| \|\mathbf{x}\|_1$
3) $\|\mathbf{x} + \mathbf{y}\|_1 = \sum_1^n |x_k + y_k| \leq \sum_1^n (|x_k| + |y_k|) = \sum_1^n |x_k| + \sum_1^n |y_k| = \|\mathbf{x}\|_1 + \|\mathbf{y}\|_1$
$p = 2$:
1) och 2) enkla, visar tredje villkoret.

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|_2^2 &= (\mathbf{x} + \mathbf{y})^T(\mathbf{x} + \mathbf{y}) = \mathbf{x}^T\mathbf{x} + 2\mathbf{x}^T\mathbf{y} + \mathbf{y}^T\mathbf{y} \\ &\leq \|\mathbf{x}\|_2^2 + 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2 = (\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)^2 \end{aligned}$$

$p = \infty$:
1) och 2) enkla, visar tredje villkoret.

$$\begin{aligned} &\|\mathbf{x} + \mathbf{y}\|_\infty = \max_k |x_k + y_k| \leq \max_k (|x_k| + |y_k|) \leq \max_k |x_k| + \max_k |y_k| \\ &= \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty \end{aligned}$$

13. Visa att $\|\cdot\|_p,\ p = 1, \infty$ verkligen är matrisnormer.

Lösning:

$p = 1$:
1) $\|A\|_1 = \max_j \sum_i |a_{i,j}|$ så om $A \neq 0$ finns något $a_{i,j} \neq 0$ varför $\|A\|_1 > 0$.
2) $\|\gamma A\|_1 = \max_j \sum_i |\gamma a_{i,j}| = \max_j \sum_i |\gamma| |a_{i,j}| = |\gamma| \max_j \sum_i |a_{i,j}| = |\gamma| \|A\|_1$
3) $\|A + B\|_1 = \max_j \sum_i |a_{i,j} + b_{i,j}| \leq \max_j \sum_i (|a_{i,j}| + |b_{i,j}|) \leq \max_j \sum_i |a_{i,j}| + \max_j \sum_i |b_{i,j}| = \|A\|_1 + \|B\|_1$
Nu till submultiplikativiteten. Vi visar $\|A\mathbf{x}\|_1 \leq \|A\|_1 \|\mathbf{x}\|_1$ först. Det följer från definitionen av normen

$$\|A\|_1 = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1}$$

att $\|A\|_1 \geq \|A\mathbf{x}\|_1/\|\mathbf{x}\|_1$. Nu till $\|AB\|_1$. Antag att max antas för kolonn k i B:

$$\|AB\|_1 = \|A\mathbf{b}_k\|_1 \leq \|A\|_1\|\mathbf{b}_k\|_1 \leq \|A\|_1\|B\|_1$$

$p = \infty$ kan visas analogt. Ett trick är att $\|A^T\|_1 = \|A\|_\infty$.

14. Visa att $\|\mathbf{x}\|_A = (\mathbf{x}^T A\mathbf{x})^{1/2}$ definierar en vektornorm (elliptisk norm), då A är symmetrisk och positivt definit.

Lösning:

Låt $A = CC^T$ vara Choleskyfaktoriseringen av A. Då är

$$\|\mathbf{x}\|_A = (\mathbf{x}^T A\mathbf{x})^{1/2} = (\mathbf{x}^T CC^T\mathbf{x})^{1/2} = ((C^T x)^T(C^T x))^{1/2} = \|C^T x\|_2$$

Vi kan alltså återinföra $\|\mathbf{x}\|_A$ på tvånormen. Eftersom A är positivt definit och därmed ickesingulär är även C ickesingulär, varför $C^T\mathbf{x} = \mathbf{0} \Leftrightarrow \mathbf{x} = \mathbf{0}$ Vi testar nu de tre normvillkoren:
1) $\|\mathbf{x}\|_A > 0, \mathbf{x} \neq \mathbf{0}$ ty $\|C^T\mathbf{x}\|_2 > 0$ om $\mathbf{x} \neq \mathbf{0}$ ty $\|\cdot\|_2$ är en norm.
2) $\|\alpha\mathbf{x}\|_A = \|\alpha C^T\mathbf{x}\|_2 = |\alpha|\|C^T\mathbf{x}\|_2 = |\alpha|\|\mathbf{x}\|_A$
3) $\|\mathbf{x}+\mathbf{y}\|_A = \|C^T(\mathbf{x}+\mathbf{y})\|_2 \leq \|C^T\mathbf{x}\|_2 + \|C^T\mathbf{y}\|_2 = \|\mathbf{x}\|_A + \|\mathbf{y}\|_A$

15. a) Visa att $\|A\|_{\max} = \max_{i,j}|a_{i,j}|$ definierar en matrisnorm, men att den ej är submultiplikativ. b) Visa att $\|A\|_F = (\sum_{i,j}|a_{i,j}|^2)^{1/2}$ är en matrisnorm (Frobeniusnormen).

Lösning:

a)
1) $\|A\|_{\max} = \max_{j,k}|a_{j,k}| > 0$ om något $a_{j,k} \neq 0$.
2) $\|\gamma A\|_{\max} = \max_{j,k}|\gamma a_{j,k}| = |\gamma|\max_{j,k}|a_{j,k}| = |\gamma|\|A\|_{\max}$
3) $\|A+B\|_{\max} = \max_{j,k}|a_{j,k}+b_{j,k}| \leq \max_{j,k}(|a_{j,k}|+|b_{j,k}|) \leq \max_{j,k}|a_{j,k}| + \max_{j,k}|b_{j,k}| = \|A\|_{\max} + \|B\|_{\max}$

Notera att denna norm inte är submultiplikativ. Tag `A = ones(2)`. Då är $\|AA\|_{\max} = 2$, men $\|A\|_{\max} = 1$.
b) Låt $\mathrm{vec}(A)$ vara den vektor som fås om man staplar alla A:s kolonner på varandra. Vi ser att $\|A\|_F = \|\mathrm{vec}(A)\|_2$.

1) $\|A\|_F = \|\mathrm{vec}(A)\|_2 > 0$ om något $a_{i,j} \neq 0$.
2) $\|\gamma A\|_F = \|\gamma\mathrm{vec}(A)\|_2 = |\gamma|\|\mathrm{vec}(A)\|_2 = |\gamma|\|A\|_F$
3) $\|A+B\|_F = \|\mathrm{vec}(A+B)\|_2 = \|\mathrm{vec}(A)+\mathrm{vec}(B)\|_2 \leq \|\mathrm{vec}(A)\|_2 + \|\mathrm{vec}(B)\|_2 = \|A\|_F + \|B\|_F$

16. Låt $D = \mathrm{diag}(d_1, ..., d_n)$ med alla $d_i \neq 0$. Beräkna $\kappa(D)$ (för de tre normer vi använder).

Lösning:

$D^{-1} = \mathrm{diag}(1/d_1, ..., 1/d_n)$. För en diagonalmatris gäller $\|D\| = \max_k |d_k|$ (för de tre normer vi använder). Så $\kappa(D) = \max|d_k| \max|1/d_k| = \max|d_k| / \min|d_k|$.

17. Beränka $\kappa_1(A)$ som funktion av α då

$$\begin{bmatrix} 1 & \alpha \\ 1 & 1 \end{bmatrix}$$

Lösning:

Om $\alpha = 1$ så är matrisen singulär och vi säger att $\kappa_1(A) = \infty$. I annat fall gäller

$$\kappa_1(A) = \underbrace{\left\| \begin{bmatrix} 1 & \alpha \\ 1 & 1 \end{bmatrix} \right\|_1}_{\|A\|_1} \underbrace{\left\| \frac{1}{1-\alpha} \begin{bmatrix} 1 & -\alpha \\ -1 & 1 \end{bmatrix} \right\|_1}_{\|A^{-1}\|_1}$$
$$= \underbrace{\max(1+|\alpha|, 2)}_{\|A\|_1} \cdot \underbrace{\max(1+|\alpha|, 2)/|1-\alpha|}_{\|A^{-1}\|_1}$$

Med andra ord, $\kappa_1(A) = 4/|1-\alpha|$ om $|\alpha| < 1$ och $(1+|\alpha|)^2/|1-\alpha|$ annars.

18. Visa att en positivt definit matris A är: a) ickesingulär och att b) inversen är positivt definit.

Lösning:

a) Om A är singulär existerar $\mathbf{x} \neq \mathbf{0}$ s.a. $A\mathbf{x} = \mathbf{0}$. Medför att $\mathbf{x}^T A\mathbf{x} = 0$, motsägelse!

b) Vi kräver inte att A är symmetrisk utan vet bara att $\mathbf{x}^T A\mathbf{x} > 0$ om $\mathbf{x} \neq \mathbf{0}$. Tag $\mathbf{x} = A^{-1}\mathbf{y}$ (notera att $\mathbf{x} = \mathbf{0} \Leftrightarrow \mathbf{y} = \mathbf{0}$). Vi får $0 < \mathbf{x}^T A\mathbf{x} = (A^{-1}\mathbf{y})^T A(A^{-1}\mathbf{y}) = \mathbf{y}^T A^{-T}\mathbf{y}$ som är en skalär så att vi kan skriva (eftersom (skalär)^T är det samma skalär) $(\mathbf{y}^T A^{-T}\mathbf{y})^T = \mathbf{y}^T A^{-T}\mathbf{y}$. Men $(\mathbf{y}^T A^{-T}\mathbf{y})^T = \mathbf{y}^T A^{-1}\mathbf{y}$. Alltså är $0 < \mathbf{y}^T A^{-1}\mathbf{y}$.

19. Antag att $A = BB^T$ där B är ickesingulär. Visa att A är symmetrisk och positivt definit.

Lösning:

Symmetrisk ty $A^T = (BB^T)^T = (B^T)^T B^T = BB^T = A$.

Positivt definit ty $\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T BB^T\mathbf{x} = (B^T\mathbf{x})^T B^T\mathbf{x} = \|B^T\mathbf{x}\|_2^2 > 0$ om $\mathbf{x} \neq 0$.

20. Antag att B nedan, av ordning $n+1$, är symmetrisk och positivt definit, α är en skalär, $\mathbf{a}$ en kolonnvektor om n element, och A en kvadratisk matris av ordning n.

$$B = \begin{bmatrix} \alpha & \mathbf{a}^T \\ \mathbf{a} & A \end{bmatrix}$$

a) Visa att $\alpha > 0$ och att A är positivt definit.
b) Beräkna Choleskyfaktoriseringen av B i termen av α, $\mathbf{a}$ och A.

Lösning:

a) $\mathbf{x}^T B\mathbf{x} > 0$ om $\mathbf{x} \neq \mathbf{0}$. Tag $x = e_1 \neq \mathbf{0}$. Ger $0 < e_1^T B e_1 = e_1^T [\alpha, \mathbf{a}^T]^T = \alpha$. Tag nu $\mathbf{x}^T = [0, \mathbf{y}]$ med godtyckligt $\mathbf{y} \neq \mathbf{0}$. Detta medför att

$$0 < \mathbf{x}^T B\mathbf{x} = [0, \mathbf{y}^T] \underbrace{\begin{bmatrix} \alpha & \mathbf{a}^T \\ \mathbf{a} & A \end{bmatrix} \begin{bmatrix} 0 \\ \mathbf{y} \end{bmatrix}}_{= \begin{bmatrix} \mathbf{a}^T\mathbf{y} \\ A\mathbf{y} \end{bmatrix}} = \mathbf{y}^T A\mathbf{y}.$$

b) Gör ansatsen (L matris, $\mathbf{z}$ vektor, och λ skalär):

$$\begin{bmatrix} \alpha & \mathbf{a}^T \\ \mathbf{a} & A \end{bmatrix} = \underbrace{\begin{bmatrix} \lambda & \mathbf{0} \\ \mathbf{z} & L \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} \lambda & \mathbf{z}^T \\ \mathbf{0} & L^T \end{bmatrix}}_{L^T} = \begin{bmatrix} \lambda^2 & \lambda\mathbf{z}^T \\ \lambda\mathbf{z} & \mathbf{z}\mathbf{z}^T + LL^T \end{bmatrix}$$

vilket medför $\lambda = \sqrt{\alpha}$, $\mathbf{z} = \mathbf{a}/\sqrt{\alpha}$ och $LL^T = A - \mathbf{z}\mathbf{z}^T = A - \mathbf{a}\mathbf{a}^T/\alpha$.

21. Antag att $A \in \mathbb{R}^{m \times n}$ har rang n. Visa att $A^T A$ är positivt definit.

Lösning:

Vi kollar:

$$x^T (A^T A) x = (Ax)^T (Ax) = ||Ax||_2^2$$

Det sista uttrycket kan inte vara negativt (det är ju en norm). Så frågan är om det kan vara noll även om $x \neq 0$.

$$||Ax||_2 = 0 \rightarrow Ax = 0$$

vilket inte kan inträffa eftersom A har full kolonnrang enligt förutsättningarna.

22. Antag att $A \in \mathbb{R}^{n \times n}$ är både ortogonal och triangulär.

- a) Visa att A är diagonal.
- b) Vilka diagonalelement har A?

Lösning:

- a) Antag att A är reell och undertriangulär. Eftersom A dessutom ar ortogonal gäller att $AA^T = I$. Innerprodukterna mellan första raden i A och kolonnerna i A^T blir

  $$a_{1,1} \cdot a_{1,1}, a_{1,1} \cdot a_{2,1}, a_{1,1} \cdot a_{3,1}, \ldots, a_{1,1} \cdot a_{n,1}$$

  eftersom första raden i A endast innehåller ett element skilt från noll (nämligen $a_{1,1}$). Men eftersom första raden i enhetsmatrisen har nollor utom i första positionen måste

  $$a_{2,1} = a_{3,1} = \ldots a_{n,1} = 0.$$

Vi kan nu utnyttja induktion och upprepa resonemanget för andra kolonnen i A osv.

- b) Det måste gälla att $a_{k,k}^2 = 1$ så att $a_{k,k} = \pm 1$.

För $n = 2$ har vi:

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} \\ 0 & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11}^2 & a_{11}a_{21} \\ a_{21}a_{11} & a_{21}^2 + a_{22}^2 \end{bmatrix} = \begin{bmatrix} a_{11}^2 & 0 \\ 0 & a_{22}^2 \end{bmatrix} = \begin{bmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{bmatrix}$$

23. Antag att den partitionerade matrisen nedan är ortogonal (A och C är kvadratiska). Visa att A och C måste vara ortogonala och att $B = 0$.

$$\begin{bmatrix} A & B \\ 0 & C \end{bmatrix}$$

Lösning:

Detta är en form av generalisering av föregående övning. Vi kollar:

$$\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}^T \begin{bmatrix} A & B \\ 0 & C \end{bmatrix} = \begin{bmatrix} A^T & 0 \\ B^T & C^T \end{bmatrix} \begin{bmatrix} A & B \\ 0 & C \end{bmatrix} = \begin{bmatrix} A^TA & A^TB \\ B^TA & B^TB + C^TC \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

Detta innebär att $A^TA = I$, så att A är ortogonal. $A^TB = 0$ skall vara nollmatrisen vilket, eftersom A är ortogonal och därmed ickesingulär, medför att $B = 0$. Detta medför slutligen att C är ortogonal, ty $B^TB + C^TC = I$.

4 Övningar: minstakvadratproblem

1. Vi vill lösa minstakvadratproblemet

$$\min_x ||Ax - b||_2^2$$

då A har ortogonala kolumner ($a_j^T a_k = 0$ då $j \neq k$). Hur förenklar denna egenskap hos A lösandet av problemet?

Lösning:

Om A har ortogonala kolonner så är $A^TA = D$, där $D = diag(a_1^Ta_1, a_2^Ta_2, ..., a_n^Ta_n)$. Antag att ingen kolonn har längden noll, då är matrisen D ickesingulär med invers $D^{-1} = diag(1/||a_1||_2^2, 1/||a_2||_2^2, ..., 1/||a_n||_2^2)$. Lösningen till minstakvadratproblemet kan skrivas (normalekvationerna) $x = (A^TA)^{-1}A^Tb$ (ty om A har ortogonala kolonner och inga nollkolonner så måste den ha full kolonnrang. Varför?) Så, $x = D^{-1}A^Tb$ eller $x_k = a_k^Tb/||a_k||_2^2$.

2. Vi vill anpassa mätpunkter (t_k, N_k) (alla $N_k > 0$) till funktionen

$$N(t) = N_0 \exp^{-\lambda t}.$$

Gör en lämplig omskriving av problemet så att parametrarna, N_0 och λ, i modellen kan bestämmas med hjälp av ett (linjärt) minstakvadratproblem.

Lösning:

Problemet är att N_0 och λ ej ingår linjärt i modellen varför vi inte kan använda $\min_x ||Ax-b||_2^2$ direkt. Idé: logaritmera

$$\log N(t) = \log N_0 - \lambda t.$$

$x_1 := \log N_0, x_2 := \lambda, b := \log N(t)$, då

$$b = x_1 - x_2 t.$$

Då får vi

$$\min_x \left\| \begin{bmatrix} 1 & -t_1 \\ 1 & -t_2 \\ \vdots & \vdots \\ 1 & -t_m \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \log N(t_1) \\ \log N(t_2) \\ \vdots \\ \log N(t_m) \end{bmatrix} \right\|_2^2$$

När x är beräknad sätter vi $N_0 = e^{x_1}$ och $\lambda = x_2$.

3. Givet mätpunkterna

$$(t_k, b_k) = (-n,0),(-(n-1),0),...,(-1,0),(0,1),(1,0),(2,0),...,(n,0)$$

anpassa en rät linje till punkterna i tvånorm. (Alla b_k-värden är 0 förutom då $t_k = 0$ när $b_k = 1$.)

Lösning:

En rät linje har ekvation: $f(t) = x_1 + x_2 t$. A och b har $2n+1$ rader och ser ut som:

$$A = \begin{bmatrix} 1 & -n \\ 1 & -(n-1) \\ \vdots & \vdots \\ 1 & -1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & n \end{bmatrix}, \; b = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}.$$

På rad $n+1$ står 1 i vektor b. Vi noterar att vi kan utnyttja föregående uppgift eftersom A har ortogonala kolonner.

Lösningen blir alltså

$$x = (A^T A)^{-1} A^T b = \begin{bmatrix} 2n+1 & 0 \\ 0 & 2n(n+1)(2n+1)/6 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2n+1} \\ 0 \end{bmatrix}.$$

Den räta linjen ges alltså av $f(t) = x_1 + x_2 t = 1/(2n+1) + 0 \cdot t$ eller $f(t) = 1/(2n+1)$ (konstant oberoende av t).

4. Antag att x och y löser problemen

$$\min_x \|Ax - b\|_2^2 \text{ resp. } \min_y \|(A+E)y - b\|_2^2$$

y är alltså lösningen till ett stört problem. Studera minstakvadratproblemet $\min_x \|Ax - b\|_2^2$ ur störningssynpunkt då

$$\mathrm{A} = \begin{bmatrix} 1 & 0 \\ 0 & \delta \\ 0 & 0 \end{bmatrix}, 0 < \delta << 1, b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}, b_3 \neq 0$$

och då vi stör A med E

$$E = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & \varepsilon \end{bmatrix}, 0 < \varepsilon << \delta << 1.$$

Lösning:

Vi sätter upp normalekvationerna $A^T Ax = A^T b$ och får

$$\mathrm{A}^T Ax = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & \delta^2 \end{bmatrix}}_{A^T A} x = \underbrace{\begin{bmatrix} b_1 \\ \delta b_2 \end{bmatrix}}_{A^T b} \rightarrow x = (A^T A)^{-1} A^T b = \begin{bmatrix} b_1 \\ b_2/\delta \end{bmatrix}$$

Vi sätter nu upp normalekvationerna för det störda problemet, dvs. $(A+E)^T (A+E)y = (A+E)^T b$.

$$(A+E)^T (A+E)y = \begin{bmatrix} 1 & 0 \\ 0 & \delta^2 + \varepsilon^2 \end{bmatrix} y = (A+E)^T b = \begin{bmatrix} b_1 \\ \delta b_2 + \varepsilon b_3 \end{bmatrix} \rightarrow$$

$$y = (A+E)^T (A+E)^{-1} (A+E)^T b = \begin{bmatrix} b_1 \\ \frac{\delta b_2 + \varepsilon b_3}{\delta^2 + \varepsilon^2} \end{bmatrix} \approx \begin{bmatrix} b_1 \\ \frac{\delta b_2 + \varepsilon b_3}{\delta^2} \end{bmatrix}$$

Så

$$y-x \approx \frac{\varepsilon}{\delta^2}\begin{bmatrix}0\\ b_3\end{bmatrix} = \|E\|_2 k_2^2(A)\begin{bmatrix}0\\ b_3\end{bmatrix}$$

med $k_2(A) = \|A^+\| \cdot \|A\| = \frac{1}{\delta}$ eftersom

$$\|A^+\| = \|(A^TA)^{-1}A^T\| = \left\| \begin{bmatrix}1 & 0 & 0\\ 0 & 1/\delta & 0\end{bmatrix} \right\| = 1/\delta; \|A\| = 1.$$

5. Vi har modellen $e^{\alpha t} \approx b$ och vill bestämma α. Formulera ett ickelinjärt och ett linjärt minstakvadratproblem. Lös problemen exakt då antalet observationer är $m = 2$ och där $t_1 = 1, t_2 = 2$ samt $b_2 = 1/2$. (Den optimala lösningen beror alltså av b_1.)

   Lösning:

   Ickelinjärt minstakvadratproblem:

$$\min_{\alpha} \|e^{\alpha t} - b\|_2^2$$

   Linjärt minstakvadratproblem:

$$\min_{\alpha} \|\alpha t - \log(b)\|_2^2$$

   För antalet observationer $m = 2$: Ickelinjärt minstakvadratproblem:

$$\min_{\alpha} (e^{\alpha t_1} - b_1)^2 + (e^{\alpha t_2} - b_2)^2$$

   Linjärt minstakvadratproblem:

$$\min_{\alpha} (\alpha t_1 - \log(b_1))^2 + (\alpha t_2 - \log(b_2))^2$$

   Först för antalet observationer $m = 2$ vi l¨oser ickelinjärt minstakvadratproblem. Vi vill lösa:

$$((e^{\alpha t_1} - b_1)^2 + (e^{\alpha t_2} - b_2)^2)'_{\alpha} = 0$$

   och deriverar ickelinjärt minstakvadratproblem:

$$((e^{\alpha t_1} - b_1)^2 + (e^{\alpha t_2} - b_2)^2)'_{\alpha} = 2t_1 e^{\alpha t_1}(e^{\alpha t_1} - b_1) + 2t_2 e^{\alpha t_2}(e^{\alpha t_2} - b_2).$$

   Då $t_1 = 1, t_2 = 2$ och $b_2 = 1/2$ vi får ekvationen:

$$0 = e^{\alpha}(e^{\alpha} - b_1) + 2e^{\alpha 2}(e^{\alpha 2} - 1/2) = e^{2\alpha} - b_1 e^{\alpha} + 2e^{4\alpha} - e^{2\alpha} = 2e^{4\alpha} - b_1 e^{\alpha}.$$

   och $2e^{4\alpha} = b_1 e^{\alpha}$, logaritmerar med $\log_e$:

$$\log(2) + 4\alpha = \log(b_1) + \alpha,$$

$$4\alpha - \alpha = \log(b_1) - \log(2),$$
$$\alpha = \frac{\log(b_1) - \log(2)}{3}$$

Nu för antalet observationer $m = 2$ vi löser linjärt minstakvadratproblem. Vi vill lösa:

$$((\alpha t_1 - \log(b_1))^2 + (\alpha t_2 - \log(b_2))^2)'_\alpha = 0$$

och deriverar linjärt minstakvadratproblem:

$$((\alpha t_1 - \log(b_1))^2 + (\alpha t_2 - \log(b_2))^2)'_\alpha = 2t_1(\alpha t_1 - \log(b_1)) + 2t_2(\alpha t_2 - \log(b_2))$$

Då $t_1 = 1, t_2 = 2$ och $b_2 = 1/2$: vi får ekvationen:

$$0 = \alpha - \log(b_1) + 2(2\alpha - \log(b_2)) = 5\alpha - (log(b_1) + 2\log(b_2))$$

och $5\alpha = log(b_1) + 2\log(b_2)$,

$$\alpha = \frac{log(b_1) + 2\log(b_2)}{5} = \frac{log(b_1) + 2\log(b_2)}{5}.$$

5 Övningar: ickelinjära ekvationer

1. Man kan härleda Newtons metod med hjälp av Taylors formel. Vi står i punkten x_k och söker en korrektion, h, så att $f(x_k + h) = 0$. Gör en Taylorutveckling och ta bara med upp till första ordningens termer i h.

   Lösning:

   Taylors formel:

   $$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)(x - x_0)^2/2 + ...$$

   $$0 = f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

   eller i iterativt form för $x = x_k + h$,

   $$0 = f(x_k + h) \approx f(x_k) + hf'(x_k)$$

   så

   $$h \approx -f(x_k)/f'(x_k)$$

   vilket ger nästa punkt

   $$x_{k+1} = x_k - f(x_k)/f'(x_k),$$

   eller Newtons metod.

2. Uppskatta $|x^* - \hat{x}|$ då $f(x) = x^3 - 2x - 5$ och $\hat{x} = 2.1$ för problemet $f(x) = 0$.

   Lösning:

   Residualen är $f(\hat{x}) = \hat{x}^3 - 2\hat{x} - 5 = 0.061 > 0$. Eftersom $f(2) = 2^3 - 2 \cdot 2 - 5 = -1 < 0$ finns minst en rot i intervallet $(2, 2.1)$. Vi noterar att $f'(x) = 3x^2 - 2 > 0$ om $x > \sqrt{2/3} \approx 0.82$. Detta innebär att funktionen är strängt växande vid (det enda) nollstället och att $x^* < \hat{x}$. Vi har visat att $|x^* - \hat{x}| \leq |f(\hat{x}|/M$ där M är en undre begränsning av $|f'(x)|$ i intervallet $x(x^*, \hat{x})$. Eftersom derivatan är strängt växande och positiv för $x \in [2, 2.1]$ och eftersom $2 < x^* < \hat{x}$ så gäller att $|x^* - \hat{x}| \leq 0.061/f'(2.1) = 0.0054319$. För att sammanfatta: $2.1 - 0.0054319 = 2.0946 \leq x^* < 2.1$ Taylors formel är:

$$0 = f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

   För exakt x^* och $x_0 = \hat{x}$ Taylors formel är:

$$0 = f(x^*) \approx f(\hat{x}) + f'(\hat{x})(x^* - \hat{x})$$

   eller för $\hat{x} = 2.1$:

$$|x^* - \hat{x}| \leq |f(\hat{x})/f'(\hat{x})| = 0.061/f'(2.1) = 0.0054319.$$

3. Vi vill uppskatta $|x^* - \hat{x}|$, då $f(x) = x^4 - 6x^2 + 9$ med $\hat{x} = 1.7$. Notera att $\sqrt{3}$ är en dubbelrot.

   Lösning:

   Residualen är $f(\hat{x}) = 0.0121 > 0$. Det är inte möjligt att få någon användbar begränsning av derivatan, $f'(x) = 4x^3 - 12x = 4x \cdot (x^2 - 3)$, ty $f'(\sqrt{3}) = 0$. Vi misstänker starkt att $\sqrt{3}$ är en dubbelrot (dvs. $f(x^*) = f'(x^*) = 0$). För att få en gräns på $|x^* - \hat{x}|$ utnyttjar vi en term till i Taylorutvecklingen och får:

$$f(\hat{x}) = f(x^* + \underbrace{\hat{x} - x^*}_{h}) = \underbrace{f(x^*)}_{0} + (\hat{x} - x^*)\underbrace{f'(x^*)}_{0} + \frac{(\hat{x} - x^*)^2}{2} f''(\xi), \quad \xi \in (x^*, \hat{x}).$$

   Om $M \leq |f''(\xi)|, \xi \in (\hat{x}, x^*)$ gäller att

$$|\hat{x} - x^*| \leq \sqrt{\frac{2|f(\hat{x})|}{M}}$$

   Vi vill uppskatta $|x^* - \hat{x}|$, då $f(x) = x^4 - 6x^2 + 9$ med $\hat{x} = 1.7$.
   Vi ser att $f''(x) = (x^4 - 6x^2 + 9)''_x = 12(x^2 - 1)$ som är växande i en omgivning kring roten $\sqrt{3}$. Vi noterar att $f(x)$ uppför sig som en parabel i en omgivning av roten. Eftersom $f'(\hat{x}) = f'(1.7) = -0.748 < 0$ ligger $\hat{x}$ till vänster om x^*. Alltså kan vi ta $M = f''(\hat{x})$ och får:

$$|\hat{x}-x^*| \leq \sqrt{\frac{2|f(\hat{x})|}{f''(\hat{x})}} < 3.3 \cdot 10^{-2}$$

För att sammanfatta:

$$1.7 \leq x^* \leq 1.7 + 3.3 \cdot 10^{-2}.$$

4. Sätt upp Newtons metod för problemet $x^2 = 1$ och visa att metoden aldrig konvergerar om $x^0 = \alpha i, \alpha \neq 0 \in \mathbb{R}$. (Vi studerar komplexa x_k med andra ord). Visa slutligen att metoden konvergerar för alla reella $x_0 \neq 0$ (svårare).

   Lösning:

   Newtons metod är:

$$x_{k+1} = x_k - f(x_k)/f'(x_k).$$

   Vi har: $f(x) = x^2 - 1 = 0, f(x_k) = x_k^2 - 1, f'(x_k) = 2x_k$, Newtons metod:

$$x_{k+1} = x_k - \frac{x_k^2 - 1}{2x_k} = \frac{x_k^2 + 1}{2x_k} = (x_k + 1/x_k)/2.$$

   Om x_k är rent imaginärt $x_k = \alpha i$ så gäller

$$\begin{aligned} x_{k+1} &= (x_k + 1/x_k)/2 = (\alpha i + 1/\alpha i)/2 = (\alpha i + i/\alpha i \cdot i)/2 \\ &= \frac{i}{2}\left[\alpha - \frac{1}{\alpha}\right] \end{aligned}$$

   som också är rent imaginärt. Så om något x_k är rent imaginärt så kommer också alla efterföljande värden att vara det (enda undantaget är om något $x_k = 0$ i vilket fall x_{k+1} inte existerar). Fixpunkterna: $x^* = f(x^*)$:

$$\begin{aligned} x^* &= (x^* + 1/x^*)/2, \\ 2x^* &= x^* + 1/x^*;\ 2x^* = \frac{(x^*)^2 + 1}{x^*}, \\ 2(x^*)^2 - (x^*)^2 - 1 &= 0;\ (x^*)^2 - 1 = 0; x^* = \pm 1. \end{aligned}$$

   Eftersom fixpunkterna är ± 1 och reella (och inte är rent imaginära) så får vi ingen konvergens.

5. Sätt upp Newtons metod för följande problem:
   a) $x^3 - 2x - 5 = 0$.
   b) $e^{-x} = x$.
   c) $x \sin x = 1$.

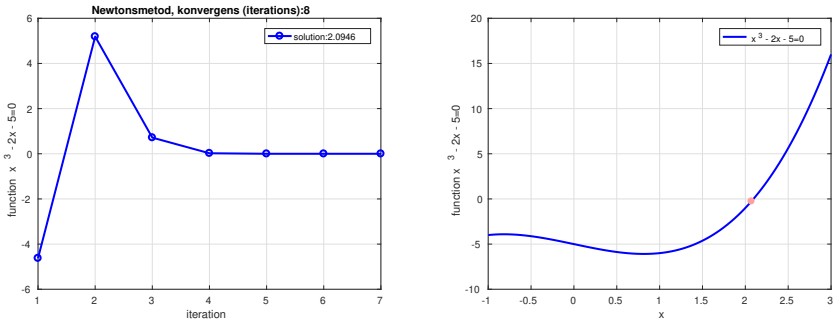
   Lösning:

   Newtons metod är för att lösa $f(x) = 0$:

$$x_{k+1} = x_k - f(x_k)/f'(x_k).$$

a) $f(x) = x^3 - 2x - 5, f'(x) = 3x^2 - 2$. Newtons metod:

$$x_{k+1} = x_k - \frac{x_k^3 - 2x_k - 5}{3x_k^2 - 2}.$$



a) Konvergens b) Exakt funktion $f(x) = x^3 - 2x - 5 = 0$

Fig. 1 *Newtons metod för* $f(x) = x^3 - 2x - 5 = 0$ *med* $x_0 = 1.5, tol = 10^{-15}$.

b) $e^{-x} = x$.
Newtons metod är för att lösa $f(x) = 0$:

$$x_{k+1} = x_k - f(x_k)/f'(x_k).$$

$f(x) = e^{-x} - x, f'(x) = -e^{-x} - 1$. Newtons metod:

$$x_{k+1} = x_k - \frac{e^{-x_k} - x_k}{-e^{-x_k} - 1}.$$

c) $x \sin x = 1$.
Newtons metod är för att lösa $f(x) = 0$:

$$x_{k+1} = x_k - f(x_k)/f'(x_k).$$

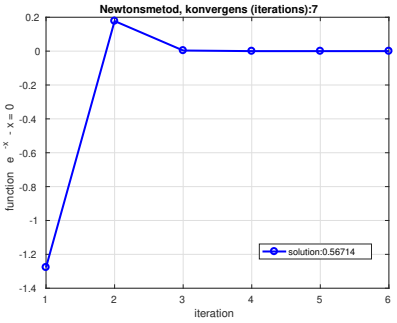
$f(x) = x \sin x - 1, f'(x) = \sin x + x \cos x$. Newtons metod:

$$x_{k+1} = x_k - \frac{x_k \sin x_k - 1}{\sin x_k + x_k \cos x_k}.$$

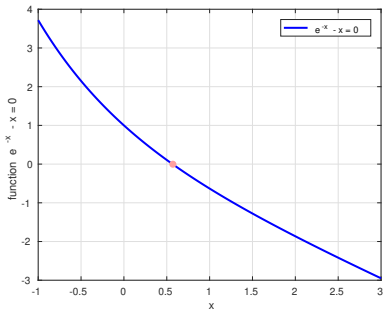
6. Newtons metod används ibland för att implementera kvadratrotsfunktionen. Vi vill beräkna $\sqrt{y}$, sätt upp Newtons metod för problemet.

Lösning:

Newtons metod är för att lösa $f(x) = 0$:

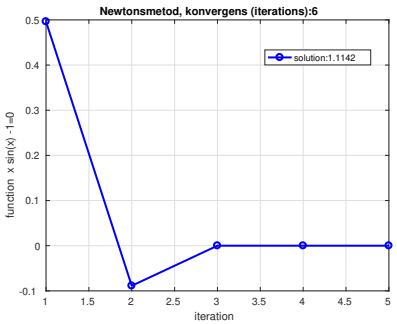


a) Konvergens

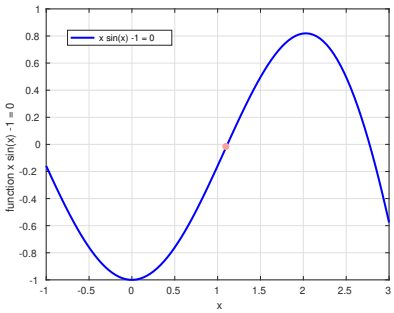


b) Exakt funktion $f(x) = e^{-x} - x = 0$

Fig. 2 *Newtons metod för* $f(x) = e^{-x} - x = 0$ *med* $x_0 = 1.5, tol = 10^{-15}$.



a) Konvergens



b) Funktion $f(x) = x\sin x - 1 = 0$

Fig. 3 *Newtons metod för* $f(x) = x\sin x - 1 = 0$ *med* $x_0 = 1.5, tol = 10^{-15}$.

$$x_{k+1} = x_k - f(x_k)/f'(x_k).$$

Vi har: $x = \sqrt{y}$, här x är okänt, då $x^2 = y$ och vi kan ta $f(x) = x^2 - y = 0$. Newtons metod för att lösa $f(x) = x^2 - y = 0$:

$$x_{k+1} = x_k - \frac{x_k^2 - y}{2x_k}.$$

7. Även division, $1/y$, kan implementeras med hjälp av Newtons metod. Formulera en lämplig ekvation och sätt sedan upp Newtons metod (som givetvis inte får innehålla någon division) för ekvationen.

   Lösning:

   Newtons metod är för att lösa $f(x) = 0$:

$$x_{k+1} = x_k - f(x_k)/f'(x_k).$$

Vi har: $x = 1/y$, här x är okänt, då $1/x = y$ och vi kan ta $f(x) = 1/x - y = 0$. Newtons metod för att lösa $f(x) = 1/x - y = 0$:

$$\begin{aligned} x_{k+1} &= x_k - \frac{1/x_k - y}{(-x_k^{-2})} = x_k + (1/x_k - y)(x_k)^2 \\ &= x_k + ((1 - yx_k)/x_k)(x_k)^2 = x_k + (1 - yx_k)x_k = x_k + x_k - yx_k^2 \\ &= 2x_k - yx_k^2 = x_k(2 - yx_k). \end{aligned}$$

8. Vi vill lösa ekvationen $x^2 - y = 0$ givet y och studerar därför fixpunktsiterationer, $x_{k+1} = g(x_k)$.
a) Är $g_1(x) = y + x - x^2$ respektive $g_2(x) = 1 + x - x^2/y$ lokalt konvergenta metoder om $y = 3$?
b) Hur ser den $g(x)$ ut som svarar mot Newtons metod?

Lösning:

a) Om vill lösa ekvationen $x^2 - y = 0$, då $x^2 = y$ och $x = \sqrt{y}$.
När $g_1(x) = y + x - x^2$ har vi: $g_1'(x) = 1 - 2x$ eller för $x = \sqrt{y}$ har vi: $g_1'(\sqrt{y}) = 1 - 2\sqrt{y}$. Om $y = 3$ då $g_1'(\sqrt{3}) = 1 - 2\sqrt{3}, |g_1'(\sqrt{3})| = |1 - 2\sqrt{3}| \approx 2.5 > 1$ och metoden $x_{k+1} = g_1(x_k)$ är ej konvergent.
För funktionen $g_2(x) = 1 + x - x^2/y$ har vi $g_2'(x) = 1 - 2x/y$ eller för $x = \sqrt{y}$ har vi: $g_2'(\sqrt{y}) = 1 - 2\sqrt{y}/y = 1 - 2/\sqrt{y}$. Om $y = 3$ då $g_2'(\sqrt{3}) = 1 - 2/\sqrt{3} \approx 0.15 < 1$ och metoden $x_{k+1} = g_2(x_k)$ är konvergent. Fixpunkt: för exakt x^* vi ska lösa $g_2(x) : x^* = g_2(x^*)$ eller $g_2(\sqrt{y}) = 1 + \sqrt{y} - y/y = \sqrt{y}$. Vi kan skriva om den ekvation som $x^* = g_2(x^*)$ för $x^* = \sqrt{y}$, alltså $x^* = \sqrt{y}$ är fixpunkt.
b) Hur ser den $g(x)$ ut som svarar mot Newtons metod ?
Newtons metod är:

$$x_{k+1} = x_k - f(x_k)/f'(x_k).$$

Nu har vi: $f(x) = x^2 - y = 0, f'(x) = 2x$. Newtons metod är:

$$x_{k+1} = x_k - \frac{{x_k}^2 - y}{2x_k} = \underbrace{\frac{x_k}{2} + \frac{y}{2x_k}}_{g(x_k)}.$$

Så $g(x) = \frac{x}{2} + \frac{y}{2x}$ och $g'(x) = 1/2 - \frac{y}{2x^2}$. För $x = \sqrt{y}$ har vi: $g_2'(\sqrt{y}) = 1/2 - \frac{y}{2y} = 0 < 1$ och metoden $x_{k+1} = x_k - \frac{{x_k}^2 - y}{2x_k}$ är konvegrent.
Fixpunkt: för exakt x^* vi ska lösa $g_2(x) : x^* = g(x^*)$ eller

$$x^* = \frac{x^*}{2} + \frac{y}{2x^*}.$$

För $x^* = \sqrt{y}$ har vi:

$$\sqrt{y} = \frac{\sqrt{y}}{2} + \frac{y}{2\sqrt{y}} = \frac{\sqrt{y}}{2} + \frac{\sqrt{y}}{2} = \sqrt{y}$$

och $x^* = \sqrt{y}$ är fixpunkt.

9. Formulera Newtons metod för följande två problem:

a)
$$\begin{cases} x_1^2 + x_2^2 - 1 = 0, \\ x_1^2 - x_2 = 0. \end{cases}$$

b)
$$\begin{cases} x_1^2 + x_1 x_2^3 - 9 = 0, \\ 3x_1^2 x_2 - x_2^3 - 4 = 0. \end{cases}$$

Lösning:

Newtons metod för system är:

$$x_{k+1} = x_k - J(f(x_k)^{-1}) f(x_k), \tag{9}$$

var $J(f(x_k)^{-1})$ är inversa Jacobianen.

a) För problem
$$\begin{cases} x_1^2 + x_2^2 - 1 = 0, \\ x_1^2 - x_2 = 0. \end{cases}$$
har vi för $f = (f_1, f_2)^T$
$$\begin{cases} f_1(x_1, x_2) = x_1^2 + x_2^2 - 1 = 0, \\ f_2(x_1, x_2) = x_1^2 - x_2 = 0. \end{cases}$$
Newtons metod (9) blir då

$$\begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \end{bmatrix} = \begin{bmatrix} x_1^k \\ x_2^k \end{bmatrix} - \begin{bmatrix} 2x_1^k & 2x_2^k \\ 2x_1^k & -1 \end{bmatrix}^{-1} \begin{bmatrix} (x_1^k)^2 + (x_2^k)^2 - 1 \\ (x_1^k)^2 - (x_2^k) \end{bmatrix}$$

b) För problem
$$\begin{cases} x_1^2 + x_1 x_2^3 - 9 = 0, \\ 3x_1^2 x_2 - x_2^3 - 4 = 0. \end{cases}$$
har vi för $f = (f_1, f_2)^T$
$$\begin{cases} f_1(x_1, x_2) = x_1^2 + x_1 x_2^3 - 9 = 0, \\ f_2(x_1, x_2) = 3x_1^2 x_2 - x_2^3 - 4 = 0. \end{cases}$$
Newtons metod (9) för $x = (x_1, x_2)^T$ blir då

$$\begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \end{bmatrix} = \begin{bmatrix} x_1^k \\ x_2^k \end{bmatrix} - \begin{bmatrix} 2x_1^k + (x_2^k)^3 & 3x_1^k (x_2^k)^2 \\ 3x_2^k \cdot 2x_1^k & 3(x_1^k)^2 - 3(x_2^k)^2 \end{bmatrix}^{-1} \begin{bmatrix} (x_1^k)^2 + x_1^k (x_2^k)^3 - 9 \\ 3(x_1^k)^2 x_2^k - (x_2^k)^3 - 4 \end{bmatrix}$$

10. Tag ett steg av Newtons metod för problemet:
$$\begin{cases} x_1^2 - x_2^2 = 0, \\ 2x_1x_2 = 1. \end{cases}$$

Lösning:

För system $\begin{cases} x_1^2 - x_2^2 = 0, \\ 2x_1x_2 = 1. \end{cases}$ har vi för $f = (f_1, f_2)^T$
$$\begin{cases} f_1(x_1,x_2) = x_1^2 - x_2^2 = 0, \\ f_2(x_1,x_2) = 2x_1x_2 - 1 = 0. \end{cases}$$
Newtons metod (9) för $x = (x_1,x_2)^T$ blir då

$$\begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \end{bmatrix} = \begin{bmatrix} x_1^k \\ x_2^k \end{bmatrix} - \begin{bmatrix} 2x_1^k & -2x_2^k \\ 2x_2^k & 2x_1^k \end{bmatrix}^{-1} \begin{bmatrix} (x_1^k)^2 - (x_2^k)^2 \\ 2x_1^k x_2^k - 1 \end{bmatrix}$$

För $x^0 = (0,1)^T$ Newtons metod är:

$$\begin{bmatrix} x_1^1 \\ x_2^1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix}^{-1} \begin{bmatrix} -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} - \begin{bmatrix} -1/2 \\ -1/2 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}.$$

11. Givet en lokalt konvergent fixpunktsiteration, $x_{k+1} = g(x_k)$. Ge en bevis-skiss för att vi får linjär konvergens om $g'(x^*) \neq 0$.

Lösning:

Från Taylors formel för $\Theta_k \in (x^*, x_k)$

$$\begin{aligned} x_{k+1} - x^* &= \left(\underbrace{g(x^*)}_{x^*} + g'(\Theta_k)(x_k - x^*) \right) - x^* \\ &= \underbrace{g'(\Theta_k)}_{\neq 0}(x_k - x^*), \quad \Theta_k \in (x^*, x_k) \end{aligned}$$

så att

$$\frac{|x_{k+1} - x^*|}{|x_k - x^*|} = |g'(\Theta_k|.$$

Om g är tillräckligt snäll kommer $g'(\Theta_k) < C < \infty$ då $k \to \infty$ vi har minst linjär konvergens som konvergerar mot $g'(x^*) \neq 0$.

12. Givet en lokalt konvergent fixpunktsiteration, $x_{k+1} = g(x_k)$. Ge en bevis-skiss för att vi får kvadratisk konvergens om $g'(x^*) = 0$.

Lösning:

Från Taylors formel för $\Theta_k \in (x^*, x_k)$

$$x_{k+1} - x^* = \left(\underbrace{g(x^*)}_{x^*} + g'(x^*)(x_k - x^*) + \frac{1}{2} g''(\Theta_k)(x_k - x^*)^2 \right) - x^*$$
$$= \underbrace{g'(x^*)}_{0}(x_k - x^*) + \frac{1}{2} g''(\Theta_k)(x_k - x^*)^2$$
$$= \frac{1}{2} g''(\Theta_k)(x_k - x^*)^2, \;\; \Theta_k \in (x^*, x_k)$$

så att

$$\frac{|x_{k+1} - x^*|}{|x_k - x^*|^2} = \frac{|g''(\Theta_k|)}{2}$$

Om g är tillräckligt snäll kommer $g''(\Theta_k) < C < \infty$ då $k \to \infty$ vi har minst kvadratisk konvergens.

13. Vi studerar Newtons med fix riktning (modifierad Newton):

$$x_{k+1} = x_k - f(x_k)/d.$$

a) Vad måste d uppfylla för att metoden skall vara lokalt konvergent? b) Vad blir, i allmänhet, konvergensordningen? c) Finns det något värde på d så att vi fortfarande får kvadratisk konvergens?

Lösning:

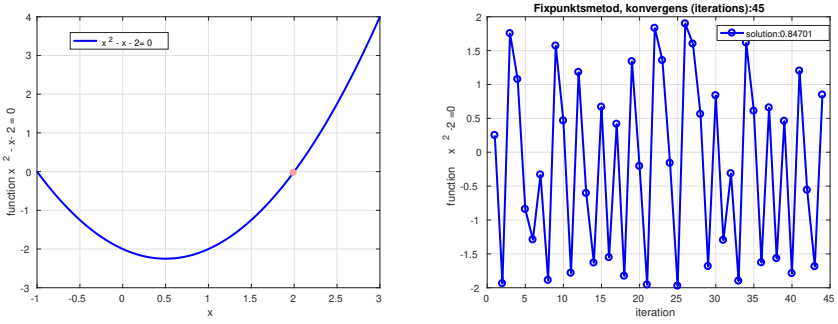
a) $g(x) = x - f(x)/d, x_{k+1} = g(x_k)$. x^* är en fixpunkt, ty $g(x^*) = x^* - f(x^*)/d = x^*$ och $f(x^*) = 0$.
$g'(x) = 1 - f'(x)/d$,
För konvergensen vi ska ha $|g'(x^*)| = |1 - f'(x^*)/d| < 1$. Ett sätt att skriva detta är $|d - f'(x^*)|/|d| < 1$, d måste alltså likna $f'(x^*)$ i denna relativa mening.
b) Vi kommer normalt att få linjär konvergens. Se föregående övning.
c) Om $d = f'(x^*)$ har vi minst kvadratisk konvergens, ty $g'(x^*) = 1 - f'(x^*)/d = 0$. Se föregående övning.

14. Vi vill lösa $x^2 - x - 2 = 0$ och studerar följande fixpunktsiterationer:
a) $g_1(x) = x^2 - 2$,
b) $g_2(x) = \sqrt{x+2}$,
c) $g_3(x) = 1 + 2/x$,
d) $g_4(x) = (x^2+2)/(2x-1)$.
Analysera konvergensen mot $x = 2$.

Lösning:

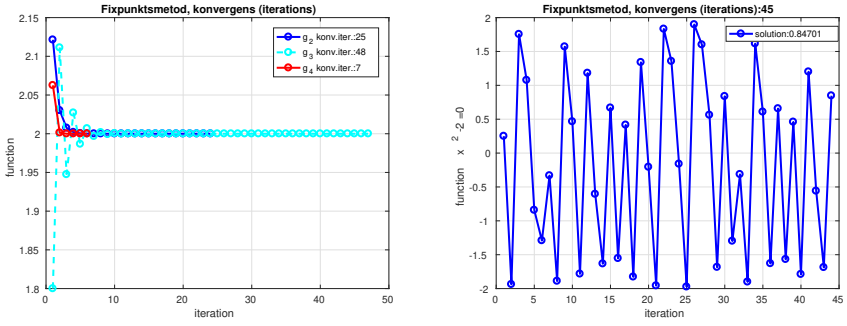
Alla funktionerna har $(2, -1)$ som fixpunkter. Räknar fixpunkter:
a) $x^* = g_1(x^*) \to x^* = (x^*)^2 - 2 \to (x^*)^2 - x^* - 2 = 0 \to x^*_{1,2} = 2; -1;$

b) $x^* = g_2(x^*) = \sqrt{x^*+2} \to (x^*)^2 - x^* - 2 = 0 \to x^*_{1,2} = 2; -1;$
c) $x^* = g_3(x^*) = 1 + 2/x^* \to (x^*)^2 - x^* - 2 = 0 \to x^*_{1,2} = (2; -1).$
Vi ska analysera konvergensen mot $x = 2$. Det återstår att undersöka derivatorna.
a) $g_1' = 2x, |g_1'(2)| = 2^2 = 4 > 1$ -ingen konv.
b) $g_2' = 1/(2\sqrt{x+2}), |g_2'(2)| = (1/2)/\sqrt{2+2} = 1/4 < 1$. Konvergent.
c) $g_3' = -2/x^2, |g_3'(2)| = 1/2$. Konvergent.
d) $|g_4'(x)| = (x^2+2)'(2x-1) - (x^2+2)(2x-1)'/(2x-1)^2 = (2x(2x-1) - 2(x^2+2))/(2x-1)^2$, så $|g_4'(2)| = 0$. Konvergent.



a) exakt $f(x) = x^2 - x - 2 = 0$ b) g_1

Fig. 4 *Övning 14. Exakt* $f(x) = x^2 - x - 2 = 0$ *och konvergenshastigheterna för* g_1 *med* $x_0 = 1.5, tol = 10^{-15}$. *Vi ser att* g_1 *är divergent.*



a) konvergens för g_2, g_3, g_4 b) divergense för g_1

Fig. 5 Övning 14. Konvergenshastigheterna för $g_i, i = 2,3,4$ med $x_0 = 2.5, tol = 10^{-15}$. Vi ser att g_1 är divergent, g_2 och g_3 konvergerar linjärt, g_4 slutligen är kvadratiskt konvergent.

15. Försök att hitta så många rötter som möjligt till systemet med hjälp av Newtons metod:

$$\begin{cases} \sin(x)+y^2+\log(z)=3, \\ 3x+2^y-z^3=0, \\ x^2+y^2+z^3=6. \end{cases}$$

Lösning:

Newtons metod blir:

$$\begin{bmatrix} x^{k+1} \\ y^{k+1} \\ z^{k+1} \end{bmatrix} = \begin{bmatrix} x^k \\ y^k \\ z^k \end{bmatrix} - \begin{bmatrix} \cos x^k & 2y^k & \frac{1}{z^k} \\ 3 & 2^{y^k}\ln 2 & -3(z^k)^2 \\ 2x^k & 2y^k & 3(z^k)^2 \end{bmatrix}^{-1} \cdot f^k,$$

var

$$f^k = \begin{bmatrix} \sin x^k + (y^k)^2 + \log z^k - 3 \\ 3x^k + 2^{y^k} - (z^k)^3 \\ (x^k)^2 + (y^k)^2 + (z^k)^3 - 6 \end{bmatrix}.$$

Om man väljer olika startpunkter (x^0, y^0, z^0) - kan man få flera olika rötter, testa i MATLAB !

6 Övningar: interpolation

1. Givet de tre punkterna $(-1,2)$, $(0,3)$ och $(1,6)$, bestäm interpolationspolynomet av grad 2:
a) med basfunktioner $t_i^j,\ i=1,2,3, j=0,1,2$ (Vandermontes form).
b) på Lagranges form,
c) på Newtons form.
Visa slutligen att vi får samma polynom i de tre fallen.

Lösning:

a) Med basfunktioner $t_i^j,\ i=1,2,3, j=0,1,2$ konstruerar vi Vandermontes matrisen. Ansätt $p(t)=x_1+x_2t+x_3t^2$

$$\begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 6 \end{bmatrix} \Rightarrow x = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

Så $p(t)=3+2t+t^2$.

b) Polynom på Langranges form i 3 punkter är:

$$p(t) = y_1 \frac{(t-t_2)(t-t_3)}{(t_1-t_2)(t_1-t_3)} + y_2 \frac{(t-t_1)(t-t_3)}{(t_2-t_1)(t_2-t_3)} + y_3 \frac{(t-t_1)(t-t_2)}{(t_3-t_1)(t_3-t_2)}$$

I vårt fall har vi:

$$p(t) = 2\frac{(t-0)(t-1)}{(-1-0)(-1-1)} + 3\frac{(t-(-1))(t-1)}{(0-(-1))(0-1)} + 6\frac{(t-(-1))(t-0)}{(1-(-1))(1-0)}$$

Förenklar vi detta uttryck får vi $p(t) = 3+2t+t^2$.
c) Polynom på Newtons form i 3 punkter är:

$$p(t) = x_1 + x_2(t-t_1) + x_3(t-t_1)(t-t_2)$$

och i vårt fall:

$$p(t) = x_1 + x_2(t-(-1)) + x_3(t-(-1))(t-0)$$

Observera:

$$\begin{aligned}
y_1 &= p(t_1) = x_1 + x_2(t_1-t_1) + x_3(t_1-t_1)(t_1-t_2) = x_1,\\
y_2 &= p(t_2) = x_1 + x_2(t_2-t_1) + x_3(t_2-t_1)(t_2-t_2) = x_1 + x_2(t_2-t_1),\\
y_3 &= p(t_3) = x_1 + x_2(t_3-t_1) + x_3(t_3-t_1)(t_3-t_2).
\end{aligned}$$

Vi får det undertriangulära systemet:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & t_2-t_1 & 0 \\ 1 & t_3-t_1 & (t_3-t_1)(t_3-t_2) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 6 \end{bmatrix}$$

Vi får: $x = [2,1,1]^T$ och interpolationspolynom på Newtons form är

$$p(t) = x_1 + x_2(t+1)) + x_3(t+1)t = 2 + (t+1)) + (t+1)t = 3+2t+t^2.$$

2. Finn $p(t)$ i Newtons form som interpolerar funktionen $f(t) = t^3$ på $1 \le t \le 4$, i punkter: $t_1 = 1, t_2 = 2, t_3 = 3, t_4 = 4$.

   Lösning:

   Polynom i Newtons form för 4 punkter är:

$$p(t) = x_1 + x_2(t-t_1) + x_3(t-t_1)(t-t_2) + x_4(t-t_1)(t-t_2)(t-t_3).$$

   Polynom som interpolerar funktionen $f(t) = t^3$ på $1 \le t \le 4$, i punkter: $t_1 = 1, t_2 = 2, t_3 = 3, t_4 = 4$ är:

$$p(t) = x_1 + x_2(t-1) + x_3(t-1)(t-2) + x_4(t-1)(t-2)(t-3).$$

Observera:

$$
\begin{aligned}
1 &= t_1^3 = p(t_1) = x_1 + x_2(t_1 - t_1) + x_3(t_1 - t_1)(t_1 - t_2) = x_1,\\
8 &= t_2^3 = p(t_2) = x_1 + x_2(t_2 - t_1) + x_3(t_2 - t_1)(t_2 - t_2) = x_1 + x_2(t_2 - t_1),\\
27 &= t_3^3 = p(t_3) = x_1 + x_2(t_3 - t_1) + x_3(t_3 - t_1)(t_3 - t_2),\\
64 &= t_4^3 = p(t_4) = x_1 + x_2(t_4 - t_1) + x_3(t_4 - t_1)(t_4 - t_2) + x_4(t_4 - t_1)(t_4 - t_2)(t_4 - t_3).
\end{aligned}
$$

Vi får det undertriangulära systemet:

$$
\begin{bmatrix}
1 & 0 & 0 & 0\\
1 & t_2 - t_1 & 0 & 0\\
1 & t_3 - t_1 & (t_3 - t_1)(t_3 - t_2) & 0\\
1 & (t_4 - t_1) & (t_4 - t_1)(t_4 - t_2) & (t_4 - t_1)(t_4 - t_2)(t_4 - t_3)
\end{bmatrix}
\begin{bmatrix} x_1\\ x_2\\ x_3\\ x_4 \end{bmatrix}
=
\begin{bmatrix}
1 & 0 & 0 & 0\\
1 & 1 & 0 & 0\\
1 & 2 & 2 & 0\\
1 & 3 & 6 & 6
\end{bmatrix}
\begin{bmatrix} x_1\\ x_2\\ x_3\\ x_4 \end{bmatrix}
=
\begin{bmatrix} 1\\ 8\\ 27\\ 64 \end{bmatrix}.
$$

Vi får: $x = [1,7,6,1]^T$ och interpolationspolynom på Newtons form är

$$
\begin{aligned}
p(t) = x_1 + x_2(t-1) + x_3(t-1)(t-2) + x_4(t-1)(t-2)(t-3) =\\
1 + 7(t-1) + 6(t-1)(t-2) + 1(t-1)(t-2)(t-3).
\end{aligned}
$$

3. Hur beräknar vi $p(t) = 5t^3 - 3t^2 + 7t - 2$ med hjälp av Horners metod?

   Lösning:

   Horner's method för $p(t) = x_1 + x_2 t + x_3 t^2 + x_4 t^3$ är:

   $$p(t) = x_1 + x_2 t + x_3 t^2 + x_4 t^3 = x_1 + t(x_2 + t(x_3 + t x_4)).$$

   I vårt fall Horners metod är:

   $$p(t) = -2 + 7t - 3t^2 + 5t^3 = -2 + t(7 + t(-3 + 5t)).$$

4. Vi vill interpolera $(t_k, y_k), k = 1, ..., n$ med $n-1$ styckvis kvadratiska polynom sådana att knutpunkterna sammanfaller med (t_k, y_k). Hur många kontinuerliga derivator kan vi rimligtvis kräva av interpolanten?

   Lösning:

   Delpolynom är andragradspolynom, och vi kan kräva 1 kontinuerligt derivata (förstaderivatan är kontinuerlig).

5. Transformera Chebyshevpunkterna

   $$t_k = -\cos\left[\frac{(2k-1)\pi}{2n}\right],\ k = 1,2,...,n, t_k \in [-1,1]$$

   från intervalet $[-1,1]$ till intervalet $[\alpha, \beta]$.

Lösning:

När t ligger i ett annat intervall, $[\alpha,\beta]$ får vi göra en linjär avbilding $kt+b$ av Chebyshevpunkterna $[-1,1]$ till detta intervall $[\alpha,\beta]$:

$$\begin{aligned} k\cdot(-1)+b &= \alpha,\\ k\cdot 1+b &= \beta, \end{aligned}$$

då

$$\begin{aligned} b &= \alpha+k,\\ k\cdot 1+\alpha+k &= \beta, \end{aligned}$$

och från andra ekvation i systemet ovan har vi

$$\begin{aligned} k &= \frac{\beta-\alpha}{2},\\ b &= \alpha+k=\alpha+\frac{\beta-\alpha}{2}=\frac{\alpha+\beta}{2}, \end{aligned}$$

och linjär avbilding av Chebyshevpunkterna till intervall $[\alpha,\beta]$ är:

$$kt_k+b=\frac{\beta-\alpha}{2}\underbrace{t_k}_{[-1,1]}+\frac{\alpha+\beta}{2} \tag{10}$$

så de transformerade Chebyshevpunkterna

$$t_k=-\cos\left[\frac{(2k-1)\pi}{2n}\right],\ k=1,2,...,n$$

blir

$$-\frac{\beta-\alpha}{2}\cos\left[\frac{(2k-1)\pi}{2n}\right]+\frac{\alpha+\beta}{2}$$

6. Vi bestämmer interpolationspolynomet, p_n, på $[0,1]$ som interpolerar e^t i punkterna $0=t_1<t_2<...<t_n=1$. Visa att oavsett hur vi väljer t_k-punkterna (i övrigt) så gäller:
$$\lim_{n\to\infty}\max_{0\le t\le 1}|e^t-p_n(t)|=0.$$
Visa att om vi väljer Chebyshevpunkterna så gäller att:
$$\max_{0\le t\le 1}|e^t-p_n(t)|\le\frac{e}{n!2^{2n-1}}.$$

Lösning:

Vi vet att

$$\underbrace{p_n(t)}_{beräknad} - \underbrace{f(t)}_{exakt} = \frac{f^{(n)}(\theta)}{n!}(t-t_1)(t-t_2)...(t-t_n)$$

där $\theta \in (t,t_1,t_2,...,t_n)$. Vi vet att $(e^t)^{(n)} = e^t$ och $|t-t_k| \le 1$ då

$$\underbrace{p_n(t)}_{beräknad} - \underbrace{e^t}_{exakt} = \frac{e^t(\theta)}{n!}(t-t_1)(t-t_2)...(t-t_n) \le \frac{e}{n!}(t-t_1)(t-t_2)...(t-t_n) \le \frac{e}{n!}. \tag{11}$$

eftersom $|t-t_k| \le 1$.
Observera att det ger oss ett konvergensresultat för varje funktion vars alla derivator är begränsade på $[0,1]$, så $|f^{(n)}(t)| \le M, 0 \le t \le 1$:

$$\lim_{n\to\infty} \max_{0\le t\le 1} |e^t - p_n(t)| = \lim_{n\to\infty} \frac{e}{n!} = 0.$$

Vi redan vet att Chebyshevpunkterna minimerar $\prod_{k=1}^{n} |t-t_k|$ när $|t| \le 1$ och maximala värdet på $|(t-t_1)(t-t_2)...(t-t_n)|$ är då $1/2^{n-1}$. Nu har vi intervallet $[0,1]$ och vi får transformera punkterna c_k med hjälp av (10) så att

$$\begin{aligned} t_k = kc_k + b &= \frac{\beta-\alpha}{2} \underbrace{c_k}_{[-1,1]} + \frac{\alpha+\beta}{2} \\ &= \frac{1-0}{2}c_k + \frac{0+1}{2} = \frac{c_k+1}{2}. \end{aligned} \tag{12}$$

Från (12) vet vi att $c_k = 2t_k - 1$ då $t_k = \frac{c_k+1}{2}$ och

$$\begin{aligned} \max_{0\le t\le 1} \prod_{k=1}^{n} |t-t_k| &= \max_{0\le t\le 1} \prod_{k=1}^{n} \left| t - \frac{c_k+1}{2} \right| \\ &= \max_{0\le t\le 1} \prod_{k=1}^{n} \left| \frac{\underbrace{2t-1}_{c} - c_k}{2} \right| = \frac{1}{2^n} \max_{-1\le c\le 1} \prod_{k=1}^{n} |c-c_k| = \frac{1}{2^{2n-1}}. \end{aligned} \tag{13}$$

Nu använder vi (13) i (11) för att få

$$\lim_{n\to\infty} \max_{0\le t\le 1} | \underbrace{p_n(t)}_{beräknad} - \underbrace{e^t}_{exakt} | = \lim_{n\to\infty} \max_{0\le t\le 1} \frac{e}{n!2^{2n-1}} = 0.$$

7. En kubisk spline kan skrivas $p_k(t) = a_k t^3 + b_k t^2 + c_k t + d_k$ på intervallet $[t_k, t_{k-1}]$. Antag att vi har n stycken t-värden. Detta ger $n-1$ intervall (lika många polynom), så antalet obestämda koefficienter är $4(n-1)$. Hur många villkor har vi?

Lösning:

Interpolationskravet ger $2(n-1)$ villkor (ty varje polynom måste interpolera 2 knutpunkter). Detta ger oss kontinuiteten. Kontinuerlig förstaderivata ger $n-2$ villkor (inre punkter) och lika många för andraderivatan. Så summa för villkor blir:

$$2(n-1)+n-2+n-2=4n-6.$$

Eftersom

$$\underbrace{4(n-1)}_{obestämda\ koeff.} \neq \underbrace{4n-6}_{villkor}$$

då det innebär att vi saknar två villkor som måste bestämmas på något sätt. Här är några vanliga tilläggsvillkor (s är splinefunktionen):

- $s''(t_1)=s''(t_n)=0$ s.k. naturliga splines (minimerar $\int_{t_1}^{t_n}(s''(t))^2dt$)
- $s'(t_1)=f'(t_1)$ och $s'(t_n)=f'(t_n)$ komplett spline
- $s'(t_1)=s'(t_n)$ samt $s''(t_1)=s''(t_n)$ periodisk första- och andraderivata (kanske rimligt med $y_1=y_n$ i detta fall)
- $p_1(t)=p_2(t),\ t\in[t_1,t_3]$ och $p_{n-2}(t)=p_{n-1}(t),\ t\in[t_{n-2},t_n]$, not-a-knot; medför att s''' kontinuerlig i $t=t_2$ och $t=t_{n-1}$. Det är alltså ett tredjegradspolynom i $[t_1,t_3]$ (och ett (annat) i $[t_{n-2},t_n]$).

8. Skriv kubisk spline för 3 punkter t_1,t_2,t_3.

Lösning:

En kubisk spline för 3 punkter t_1,t_2,t_3 kan skrivas som:

$$p_1(t)=\alpha_1+\alpha_2t+\alpha_3t^2+\alpha_4t^3, \tag{14}$$

$$p_2(t)=\beta_1+\beta_2t+\beta_3t^2+\beta_4t^3. \tag{15}$$

Koefficienterna $\alpha_i,\beta_i, i=1,2,3,4$ (8 koefficienter) ska bestämmas.
Interpolationskravet ger 4 villkor (1)-(4) (ty varje polynom måste interpolera 2 knutpunkter), som ger oss kontinuiteten:

- (1)
$$p_1(t_1)=y_1=\alpha_1+\alpha_2t_1+\alpha_3t_1^2+\alpha_4t_1^3$$
- (2)
$$p_1(t_2)=y_2=\alpha_1+\alpha_2t_2+\alpha_3t_2^2+\alpha_4t_2^3$$
- (3)
$$p_2(t_2)=y_2=\beta_1+\beta_2t_2+\beta_3t_2^2+\beta_4t_2^3,$$
- (4)
$$p_2(t_3)=y_3=\beta_1+\beta_2t_3+\beta_3t_3^2+\beta_4t_3^3$$

Kontinuerlig förstaderivata $p'_1(t), p'_2(t)$ ger 1 villkor (inre punkt)

$$p'_1(t_2) \in C \Longrightarrow p'_1(t_2) = p'_2(t_2)$$

och lika många för andraderivatan:

$$p''_1(t_2) \in C \Longrightarrow p''_1(t_2) = p''_2(t_2).$$

- (5) $p'_1(t_2) \in C \Longrightarrow p'_1(t_2) = p'_2(t_2)$

$$p'_1(t) = \alpha_2 + 2\alpha_3 t + 3\alpha_4 t^2$$
$$p'_2(t) = \beta_2 + 2\beta_3 t + 3\beta_4 t^2$$

$p'_1(t_2) = p'_2(t_2)$:

$$p'_1(t_2) = \alpha_2 + 2\alpha_3 t_2 + 3\alpha_4 t_2^2 =$$
$$= \beta_2 + 2\beta_3 t_2 + 3\beta_4 t_2^2 = p'_2(t_2)$$

- (6) $p''_1(t_2) \in C \Longrightarrow p''_1(t_2) = p''_2(t_2)$

$$p''_2(t) = 2\beta_3 + 6\beta_4 t$$
$$p''_1(t) = 2\alpha_3 + 6\alpha_4 t$$
$$p''_2(t_2) = 2\beta_3 + 6\beta_4 t_2 =$$
$$= 2\alpha_3 + 6\alpha_4 t_2 = p''_1(t_2)$$

Så vi har $4+2=6$ villkor (1)-(4), (5)-(6), behöver 2 till (vi har 8 koefficienter, som ska bestämmas). Vi väljer följande 2 tillägsvillkor: $p''_1(t_1) = 0; p''_2(t_3) = 0$:

$$2\alpha_3 + 6\alpha_4 t_1 = 0,$$
$$2\beta_3 + 6\beta_4 t_3 = 0.$$

9. Definera splinefunktion av grad 1 som interpolerar $(t_1, y_1), (t_2, y_2), (t_3, y_3)$ och som består av styckvisa polynom av grad 1 på intervallen $[t_1, t_2], [t_2, t_3]$.

Lösning:

Splinefunktion av grad 1 för 3 punkter $(t_1, y_1), (t_2, y_2), (t_3, y_3)$ kan skrivas som:

$$p_1(t) = \alpha_1 + \alpha_2 t, \tag{16}$$
$$p_2(t) = \beta_1 + \beta_2 t. \tag{17}$$

Fyra koefficienterna $\alpha_i, \beta_i, i = 1, 2$ ska bestämmas.
Interpolationskravet ger $2(n-1)$ villkor (ty varje polynom måste interpolera 2 knutpunkter). Detta ger oss kontinuiteten. Vi ska inte ha villkor för derivator eftersom vi ska definera splinefunktion av grad 1. Så vi har för 3 punkter: $2(n-$

$1) = 2(3-1) = 4$ villkor och 4 koefficienterna $\alpha_i, \beta_i, i = 1,2$ ska bestämmas från systemet:

$$p_1(t_1) = y_1 = \alpha_1 + \alpha_2 t_1$$
$$p_1(t_2) = y_2 = \alpha_1 + \alpha_2 t_2$$
$$p_2(t_1) = y_1 = \beta_1 + \beta_2 t_1$$
$$p_2(t_2) = y_2 = \beta_1 + \beta_2 t_2.$$

7 Övningar: kvadratur

1. Använd trapetsmetoden för att beräkna $\int_0^1 x^2 dx$.

   Lösning:

   Trapetsmetoden:

   $$\int_a^b f(x)dx \approx \frac{h}{2}(f(a)+f(b)),\ h = b-a$$

   Trapetsmetoden för $\int_0^1 f(x)dx$ är:

   $$\int_0^1 f(x)dx \approx \frac{1}{2}(f(1)+f(0)) \cdot (1-0).$$

   I vårt fall vi har $f(x) = x^2$, då trapetsmetoden för $\int_0^1 x^2 dx$ ger oss:

   $$\int_0^1 x^2 dx \approx \frac{1}{2}(1^2+0^2) = \frac{1}{2}.$$

2. Använd mittpunktsmetoden (rektangelmetoden) för att beräkna integralen $\int_0^1 4x^3 dx$.

   Lösning:

   Rektangelmetoden för $\int_a^b f(x)dx$ är:

   $$\int_a^b f(x)dx \approx (b-a) f\left(\frac{a+b}{2}\right).$$

   I vårt fall vi har $f(x) = 4x^3$, då rektangelmetoden för $\int_0^1 4x^3 dx$ ger oss:

   $$\int_0^1 4x^3 dx \approx (1-0) f\left(\frac{1+0}{2}\right) = f(1/2) = 4 \cdot (1/2)^3 = 1/2.$$

3. Använd Simpsons metod för att beräkna $\int_0^1 x^2 dx$.

   Lösning:

   Simpsons metod :

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]$$

   Vi har: $a=0, b=1, f(x)=x^2, f(a)=a^2, f(0)=0, f(b)=f(1)=1^2=1, f\left(\frac{a+b}{2}\right)=f((0+1)/2)=f(1/2)=(1/2)^2=1/4$.

$$\int_0^1 x^2 dx \approx \frac{1-0}{6}[0+4\cdot 1/4+1]=1/3.$$

4. Vi har följande kvadraturformel:

$$\int_0^1 f(x)dx \approx \sum_{k=1}^{n} \omega_k f(x_k)$$

   där vi vet att vi approximerar $f(x)$ med polynom $p(x)$ som har polynomiella gradtalet minst ett. Visa att $\sum_{k=1}^{n} \omega_k = 1$.

   Lösning:

   Metoden är exakt för polynom av åtminstone grad noll (konstant polynom $p(x) = 1$). Då gäller:

$$1 = \int_0^1 1dx = \sum_{i=1}^{n} w_i p(x_i) = \sum_{i=1}^{n} w_i.$$

5. Vi har en kvadraturformel $\int_{-1}^{1} f(x)\,dx \approx \omega_1 f(x_1)+\omega_2 f(x_2)$. Hur ser motsvarande kvadraturformel ut på intervalet $[7,10]$?

   Lösning:

   Om vi ska approximera integral

$$\int_a^b f(t)dt,$$

   t ligger i ett intervall $[a,b]$, och x ligger på $[-1,1]$, får vi göra en linjär avbilding till detta intervall:

$$t = \frac{b-a}{2}x+\frac{a+b}{2}.$$

   Vi gör ett variabelbytte oh sätter $t = 1.5x+8.5, dt = 1.5dx$, då integral $\int_7^{10} f(t)dt$ beräknas som

$$\int_7^{10} f(t)dt = 1.5\int_{-1}^{1} f(1.5x+8.5)\ dx$$
$$\approx 1.5\omega_1 f(1.5x_1+8.5)+1.5\omega_2 f(1.5x_2+8.5).$$

6. Hitta ω och x_k så att följande kvadraturformel får så högt polynomiellt gradtal som möjligt. Vad är detta gradtal?

$$\int_{-1}^{1} f(x)dx \approx \sum_{k=1}^{3} \omega f(x_k)$$

Lösning:

Vi approximerar $f(x)$ med polynom $p(x) = x^k$. Formeln skall vara exakt för polynom $x^k, k = 0,1,...,m$ för maximalt m så att

$$\int_{-1}^{1} x^k dx = \sum_{j=1}^{n} w_j p(x_j) = w\sum_{j=1}^{n} p(x_j). \tag{18}$$

Vi beräknar först

$$\int_{-1}^{1} x^k dx = \frac{x^{k+1}}{k+1}\Big|_{-1}^{1} = \frac{1-(-1)^{k+1}}{k+1}. \tag{19}$$

Vi använder (18) och (19) för $n = 3$ för att få:

$$2 = w(1+1+1), k = 0,$$
$$0 = w(x_1+x_2+x_3), k = 1,$$
$$2/3 = w(x_1^2+x_2^2+x_3^2), k = 2,$$
$$0 = w(x_1^3+x_2^3+x_3^3), k = 3.$$

Från första ekvation får vi $w = 2/3$.Vi vet inte helt säkert hur många ekvationer, n, vi skall ställa upp. Tar vi för litet n kommer x_j att bero på parametrar och om vi tar för stort n blir systemet inte lösbart. Fyra ekvationer verkar dock lämpligt eftersom vi har fyra obekanta. Vi borde kunna anta att x_j uppvisar vissa symmetriegenskaper eftersom integrationsintervallet är symmetriskt kring nollan och det är rimligt att anta att $x_1 = -x_3, x_2 = 0$. Lösningen blir: $x_1 = -1/\sqrt{2}, x_2 = 0, x_3 = 1/\sqrt{2}$. Kunde vi ha tagit $m = 4$? Vi använder (18) och (19) för $n = 4$ för att få:

$$2/5 \neq w(x_1^4+x_2^4+x_3^4) = 1/3.$$

Från andra sidan, med $x_1 = -1/\sqrt{2}, x_2 = 0, x_3 = 1/\sqrt{2}, w = 2/3$ får vi $w(x_1^4 + x_2^4+x_3^4) = 1/3$, och då $2/5 \neq 1/3$. Det betyder att metoden är exakt för polynom upp till och med grad tre.

7. Välj w_1, w_2, x_1, x_2, i kvadraturformeln nedan, så att den får så högt polynomiellt gradtal m som möjligt. Vad blir detta gradtal ?

$$\int_0^1 x^k dx = w_1 x_1^k + w_2 x_2^k, \;\; k = 0, 1, ..., m.$$

Lösning:

Formeln skall vara exakt för polynom $x^k, k = 0, 1, ..., m$ för maximalt m. Vi beräknar först

$$\int_0^1 x^k dx = \frac{x^{k+1}}{k+1}\Big|_0^1 = 1/(k+1). \tag{20}$$

Vi använder (20) för att få:

$$\begin{aligned} 1 &= w_1 + w_2, k = 0, \\ 1/2 &= w_1 x_1 + w_2 x_2, k = 1, \\ 1/3 &= w_1 x_1^2 + w_2 x_2^2, k = 2, \\ 1/4 &= w_1 x_1^3 + w_2 x_2^3, k = 3, \\ 1/5 &= w_1 x_1^4 + w_2 x_2^4, k = 4. \end{aligned}$$

Första ekvationen ger $w_{1,2} = 1/2$. Lös ut för $k = 1, 2$ ekvationen $2x_2^2 - 2x_2 + \frac{1}{3} = 0$ (vi noterar, att $x_1 < x_2$) för att få $x_1 = \frac{1-1/\sqrt{3}}{2}, x_2 = \frac{1+1/\sqrt{3}}{2}$. Vi kollar nu fall $k = 3$. Utnyttjar vi binomialsatsen ser vi att $(1+c)^3 + (1-c)^3 = 2(1+3c^2)$ så att $w_1 x_1^3 + w_2 x_2^3 = (1/2^4) \cdot 2(1+3/3) = 1/4$, vilket är lika med det exakta värdet. Stämmer det för k = 4? Inte. Så, det polynomiella gradtalet är 3.

8. Vi vill beräkna

$$\int_0^3 f(x)dx = \int_0^3 e^{-x^2} dx$$

med hjälp av Gausskvadratur med 3 vikter.

Lösning:

Metoden (Gausskvadratur med 3 vikter) är:

$$\int_{-1}^1 f(x)dx \approx \frac{5}{9} f\left(-\sqrt{3/5}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{3/5}\right).$$

Vi transformerar interval $[0, 3]$ för x, till $[-1, 1]$ för t, med hjälp av följande linjär transformation:

$$x = \frac{b-a}{2} t + \frac{a+b}{2} = \frac{3-0}{2} t + \frac{3+0}{2}$$

och integral $\int_0^3 e^{-x^2} dx$ för $f(x) = e^{-x^2}$ kan beräknas som

$$\int_0^3 e^{-x^2}dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}t+\frac{a+b}{2}\right)\,dt$$
$$\approx \frac{b-a}{2}\sum_{i=1}^{3}\omega_i f\left(\frac{b-a}{2}t_i+\frac{a+b}{2}\right)$$
$$= \frac{3-0}{2}\cdot\Big[\frac{5}{9}\cdot f\left(\frac{3-0}{2}t_1+\frac{3+0}{2}\right)$$
$$+\frac{8}{9}\cdot f\left(\frac{3-0}{2}t_2+\frac{3+0}{2}\right)+\frac{5}{9}\cdot f\left(\frac{3-0}{2}t_3+\frac{3+0}{2}\right)\Big]$$

i Gausspunkter

$$t_1 = -\sqrt{3/5}; t_2 = 0; t_3 = \sqrt{3/5}$$

med vikter

$$\omega_1 = 5/9; \omega_2 = 8/9; \omega_3 = 5/9.$$

9. Använd Taylorutveckling för att härleda första ordningen noggrannhet för

$$f'(x) \approx \frac{f(x+h)-f(x)}{h}. \tag{21}$$

Lösning:

Approximativt värde (Taylor's theorem):

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

$Q \in [x, x+h]$

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

Dividera med h:

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

Trunkeringsfel:

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

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

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

Vi fick för $f'(x)$ första ordningen noggrannhet för approximation (21).

10. Använd Taylorutveckling för att härleda andra ordning noggrannhet för

$$f'(x) \approx \frac{f(x+h)-f(x-h)}{2h} \tag{22}$$

Lösning:

Approximativt värde (Taylor's theorem):

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

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

$(*)-(**)$:

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

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

Eller

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

Trunkeringsfel:

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

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

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

Vi fick andra ordningen noggrannhet för approximation (22).

11. Använd Taylorutveckling för att härleda andra ordning noggrannhet för

$$f''(x) \approx \frac{f(x+h)-2f(x)+f(x-h)}{h^2} \tag{23}$$

Lösning:

Approximativt värde (Taylor's theorem):

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

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

$(*)+(**)$:

$$f(x+h) + f(x-h) = 2f(x) + \frac{2f''(x)}{2!}h^2 + O(h^4)$$

$$f''(x) = \frac{f(x+h) + f(x-h) - 2f(x) - O(h^4)}{h^2}$$

$O(h^4) = \frac{2f^{(4)}(x)}{24}h^4$; $\frac{f^{(4)}(x)}{12}\frac{h^4}{h^2} = \frac{f^{(4)}(x)}{12}h^2 \longrightarrow$ Låt $M \leq |f^{(4)}(Q)|$, då trunkeringsfel ε är begränsad med $\varepsilon < \frac{Mh^2}{12}$, därför approximation (23) har andra ordningen noggrannhet.

8 Övningar: ordinära differentialekvationer

1. Sätt upp Eulers metod för problemet $y'(t) = t + 2y, y(0) = 1$ och beräkna $y_k, k = 0,1,2,3$ med $\tau = 0.1$.

   Lösning: Explicit Eulers metod är:

   $y_{k+1} = y_k + \tau f(t_k, y_k), y_0 = y(t_0)$. I vårt fall $f(t,y) = t + 2y, t_0 = 0, y(t_0) = 1$ och vi får följande approximationer:

   $$\begin{aligned} &y_0 = 1, \\ &y_1 = y_0 + \tau f(t_0, y_0) = 1 + 0.1(0 + 2\cdot 1) = 1.2, \\ &y_2 = y_1 + \tau f(t_1, y_1) = 1.2 + 0.1(0.1 + 2\cdot 1.2) = 1.45, \\ &y_3 = y_2 + \tau f(t_2, y_2) = 1.45 + 0.1(0.2 + 2\cdot 1.45) = 1.76. \end{aligned}$$

2. Tag två steg med framåt, eller explicit, Eulers metod för systemet:
   $$\begin{cases} y_1'(t) = y_2, \\ y_2'(t) = t + y_1 + y_2 \\ y_1(0) = 1, \\ y_2(0) = 2. \end{cases}$$
   med $\tau = 0.1$.

   Lösning:

   Framåt, eller explicit, Eulers metod är:
   $y_{k+1} = y_k + \tau f(t_k, y_k), y_0 = y(t_0)$.
   I vårt fall: $t_0 = 0, y(t_0) = [y_1(t_0), y_2(t_0)]^T = [1,2]^T$, $f(t,y) = [y_2, t + y_1 + y_2]^T$.
   Vi får följande approximationer:

$y_0 = [1,2]^T$,
$y_1 = y_0 + \tau f(t_0, y_0) = [1,2]^T + 0.1 \cdot [2, 0+1+2]^T = [1.2, 2.3]^T$,
$y_2 = y_1 + \tau f(t_1, y_1) = [1.2,2.3]^T + 0.1 \cdot [2.3, 0.1+1.2+2.3]^T = [1.43, 2.66]^T$

3. Skriv om följande system ekvationer som ett första ordningens system:

$$\begin{cases} u'' = 2u'v' + v^2 + t, \\ v''' = u + v + v''u, \\ u(0) = 1,\ u'(0) = -1, \\ v(0) = 2,\ v'(0) = 3,\ v''(0) = -4. \end{cases}$$

Lösning:

Inför $y_1 = u$, $y_2 = u' = y_1'$, $y_3 = v$, $y_4 = v' = y_3'$ och $y_5 = v'' = y_4'$. Systemet blir

$$\begin{cases} y_1' = y_2, \\ y_2' = 2y_2y_4 + y_3^2 + t, \\ y_3' = y_4, \\ y_4' = y_5, \\ y_5' = y_1 + y_3 + y_5y_1, \\ y_1(0) = 1, \\ y_2(0) = -1, \\ y_3(0) = 2, \\ y_4(0) = 3, \\ y_5(0) = -4. \end{cases}$$

4. Skriv om följande ekvationer som första ordningens system:

- a) $y'' = t + y + y', y(0) = 1, y'(0) = -1$
- b) $y''' = y'' + ty, y(0) = 1, y'(0) = -1, y''(0) = 3,$
- c) $y''' = y'' - 2y' + y - t + 1, y(0) = 1, y'(0) = -1, y''(0) = 3.$

Lösning:

- a) $y'' = t + y + y', y(0) = 1, y'(0) = -1$:
  Sätt $u_1 = y, u_2 = u_1' = y'$. Vi får systemet:
  $$\begin{cases} u_1' = u_2, \\ u_2' = t + u_1 + u_2, \\ u_1(0) = 1, u_2(0) = -1. \end{cases}$$

- b) $y''' = y'' + ty, y(0) = 1, y'(0) = -1, y''(0) = 3$:
  Sätt $y = u_1, y' = u_1' = u_2, y'' = u_2' = u_3$. Vi får systemet:
  $$\begin{cases} u_1' = u_2, \\ u_2' = u_3, \\ u_3' = u_3 + tu_1 \\ u_1(0) = 1, u_2(0) = -1, u_3(0) = 3. \end{cases}$$
- c) $y''' = y'' - 2y' + y - t + 1, y(0) = 1, y'(0) = -1, y''(0) = 3$:
  Sätt $y = u_1, y' = u_1' = u_2, y'' = u_2' = u_3$. Vi får systemet:
  $$\begin{cases} u_1' = u_2, \\ u_2' = u_3, \\ u_3' = u_3 - 2u_2 + u_1 - t + 1, \\ u_1(0) = 1, u_2(0) = -1, u_3(0) = 3. \end{cases}$$

5. Skriv om följande problem på standardform och sedan som första ordningens system:

$$\begin{cases} t^2 v''(t) & = t^3 + v(t)v'(t) + z'(t)z(t) + (w(t))^3, \\ \frac{z''(t)}{v(t)} & = z(t) + \frac{v'(t)+t}{v(t)} - w(t), \\ w'(t) & = 5v(t)z'(t) + w(t) + t, \\ v(-1) & = -0.1, \\ v'(-1) & = -0.1, \\ z(-1) & = -0.1, \\ z'(-1) & = -0.2, \\ w(-1) & = 0.5. \end{cases}$$

Lösning:

Först skriver vi om systemet på standardform:

$$\begin{cases} v''(t) & = t + \frac{v(t)v'(t)+z'(t)z(t)+(w(t))^3}{t^2}, \\ z''(t) & = z(t)v(t) + v'(t) + t - w(t)v(t), \\ w'(t) & = 5v(t)z'(t) + w(t) + t, \\ v(-1) & = -0.1, \\ v'(-1) & = -0.1, \\ z(-1) & = -0.1, \\ z'(-1) & = -0.2, \\ w(-1) & = 0.5. \end{cases}$$

Sätt

$$x_1(t) = v(t),$$
$$x_2(t) = v'(t),$$
$$x_3(t) = z(t),$$
$$x_4(t) = z'(t),$$
$$x_5(t) = w(t).$$

Vi får systemet:

$$\begin{cases} x_1'(t) & = x_2(t), \\ x_2'(t) & = t + \frac{x_1(t)x_2(t)+x_3(t)x_4(t)+(x_5(t))^3}{t^2}, \\ x_3'(t) & = x_4(t), \\ x_4'(t) & = x_1(t)x_3(t) + x_2(t) + t - x_5(t)x_1(t), \\ x_5'(t) & = 5x_1(t)x_4(t) + x_5(t) + t, \\ x_1(-1) & = -0.1, \\ x_2(-1) & = -0.1, \\ x_3(-1) & = -0.1, \\ x_4(-1) & = -0.2, \\ x_5(-1) & = 0.5. \end{cases}$$

6. Sätt upp bakåt-Euler för problemet

$$y' = -y^2, y(0) = 1.$$

Formulera den ickelinjära ekvation som uppkommer för att beräkna y_{k+1} samt ställ upp Newtons metod för denna ekvation.

Lösning:

Bakåt Euler:

$$\frac{y^{k+1} - y^k}{h} = -(y^{k+1})^2$$

eller

$$y^{k+1} + h(y^{k+1})^2 = y^k.$$

För att lösa den ekvation vi använder Newtons metod: vi inför ny variabel $z = y^{k+1}$ och skriver om bakåt Eulers metod som:

$$z + hz^2 = y^k.$$

Newtons' metod för $f(z) = z + hz^2 - y^k$ blir:

$$z^{j+1} = z^j - \frac{f(z^j)}{f'(z^j)} = z^j - \frac{h(z^j)^2 + z^j - y^k}{2hz^j + 1}.$$

Här, j är iteration i Newtons metod.

7. Vilka lösningar har följande problem?

$$y' = 3/2y^{1/3}, y(0) = 0$$

Lösning:

Ekvationen är separabel. Löser vi på den på ett av de vanliga sätten, får vi:

$$\int \frac{dy}{y^{1/3}} = 3/2 \int dt$$

och

$$3/2y^{2/3} = 3/2t + const,$$

eller $y(t) = (t + 2/3 \cdot const)^{3/2}$.
Begynnelsevärdet ger $const = 0$ och då $y(t) = t^{3/2}$. Lösningen är inte entydig eftersom även $y(t) = 0$ är en lösning.

8. Eulers metod kan härledas på följande sätt:

$$y(t+h) = y(t) + hy'(t) + \frac{h^2}{2}y''(t) + ... \approx y(t) + hy'(t) = y(t) + hf(t, y(t)).$$

vilket ger framåt Eulers metoden

$$y_{k+1} = y_k + hf(t_k, y_k).$$

Härled en högre ordningens metod genom att ta med nästa term i Taylorutvecklingen.

Lösning:

Approximera

$$y''(t) \approx \frac{y'(t) - y'(t-h)}{h}$$

så att

$$y(t+h) \approx y(t) + hy'(t) + \frac{h^2}{2}y''(t) \tag{24}$$

$$= y(t) + hy'(t) + \frac{h^2}{2}\frac{y'(t) - y'(t-h)}{h} \tag{25}$$

$$= y(t) + hf(t,y) + \frac{h}{2}(f(t,y) - f(t-h, y-h)) \tag{26}$$

$$= y(t) + \frac{h}{2}(3f(t,y) - f(t-h, y-h)). \tag{27}$$

Detta leder till metoden:

$$y_{k+1} = y_k + \frac{h}{2}(3f(t_k, y_k) - f(t_{k-1}, y_{k-1})),$$

som är andra ordningens flerstegsmetod.
En annan tänkbar approximation är t.ex.

$$y''(t) \approx \frac{y'(t+h) - y'(t)}{h}$$

så att

$$y(t+h) \approx y(t) + hy'(t) + \frac{h^2}{2}y''(t) \tag{28}$$

$$= y(t) + hy'(t) + \frac{h^2}{2}\frac{y'(t+h) - y'(t)}{h} \tag{29}$$

$$= y(t) + hf(t,y) + \frac{h}{2}(f(t+h, y+h) - f(t,y)) \tag{30}$$

$$= y(t) + \frac{h}{2}(f(t+h, y+h) + f(t,y)). \tag{31}$$

Detta leder till implicita flerstegsmetoden:

$$y_{k+1} = y_k + \frac{h}{2}(f(t_{k+1}, y_{k+1}) + f(t_k, y_k)).$$

9. Sätt upp implicit Eulers, eller bakåt-Eulers, metod och första iteration i den för problemet

$$\begin{cases} x'(t) = 5x(t) - 2y(t) + 2t, \\ y'(t) = x(t) + y(t) + t + 1, \\ x(5) = 0. \\ y(5) = 0. \end{cases} \tag{32}$$

Lösning:

Implicit, eller bakåt-Eulers metod är:
$v_{k+1} = v_k + \tau f(t_{k+1}, y_{k+1})$ för diskretiseringen $v'(t) \approx \frac{v_{k+1} - v_k}{\tau}$ var $v_k = v(t_k), t_{k+1} = t_k + \tau$.
Bakåt-Eulers metod för vårt problem är:

$$\begin{cases} \frac{x_{k+1} - x_k}{\tau} & = 5x_{k+1} - 2y_{k+1} + 2t_{k+1}, \\ \frac{y_{k+1} - y_k}{\tau} & = x_{k+1} + y_{k+1} + t_{k+1} + 1, \end{cases}$$

som kan skrivas om :

$$\begin{cases} x_{k+1}-x_k &= 5\tau x_{k+1}-2\tau y_{k+1}+2\tau(t_k+\tau), \\ y_{k+1}-y_k &= \tau x_{k+1}+\tau y_{k+1}+\tau(t_k+\tau)+\tau, \end{cases}$$

eller

$$\begin{cases} x_{k+1}-5\tau x_{k+1}+2\tau y_{k+1} &= x_k+2\tau(t_k+\tau), \\ -\tau x_{k+1}+y_{k+1}-\tau y_{k+1} &= y_k+\tau(t_k+\tau)+\tau. \end{cases}$$

För att hitta x_{k+1}, y_{k+1} konstruerar vi systemet av ekvationer $Av = b$ med okänt vektorn $v = [x_{k+1}, y_{k+1}]^T$, känt vektor $b = [x_k + 2\tau(t_k+\tau), y_k + \tau(t_k+\tau)+\tau]^T$ och matrisen

$$A = \begin{bmatrix} 1-5\tau & 2\tau \\ -\tau & 1-\tau \end{bmatrix}.$$

För $k = 0$ har vi : $[x_0, y_0]^T = [x(t_0), y(t_0)]^T = [x(5), y(5)]^T = [0,0]^T$. Första iteration i Bakåt-Eulers metod ska vara:

$$[x_1, y_1]^T = A^{-1}[x_0 + 2\tau(t_k+\tau), y_0 + \tau(t_k+\tau)+\tau]^T = A^{-1}[2\tau(5+\tau), \tau(5+\tau)+\tau]^T.$$

10. Sätt upp explicit Eulers eller Framåt-Eulers metod och första iteration i den för problemet

$$\begin{cases} y'(t) = \sin(x(t)) + 2t, \\ x'(t) = \cos(y(t)) - 2tx(t), \\ y(0) = 0, \\ x(0) = 0. \end{cases} \tag{33}$$

Lösning:

Explicit, eller Framåt-Eulers metod är:
$v_{k+1} = v_k + \tau f(t_k, v_k)$ för diskretiseringen $v'(t) \approx \frac{v_{k+1}-v_k}{\tau}$.
Framåt-Eulers metod för vårt problem är:

$$\frac{y_{k+1}-y_k}{\tau} = \sin(x_k) + 2t_k;$$
$$\frac{x_{k+1}-x_k}{\tau} = \cos(y_k) - 2t_k x_k.$$

eller

$$y_{k+1} = y_k + \tau(\sin(x_k) + 2t_k),$$
$$x_{k+1} = x_k + \tau(\cos(y_k) - 2t_k x_k).$$

Första iteration i den för $k = 0, t_0 = 0, y_0 = y(t_0) = y(0) = 0, x_0 = x(t_0) = 0$ ska vara:

$$y_1 = y_0 + \tau(\sin(x_0) + 2t_0) = 0 + \tau(0 + 2 \cdot 0) = 0;$$
$$x_1 = x_0 + \tau(\cos(y_0) - 2t_0 x_0) = 0 + \tau(1 - 2 \cdot 0) = \tau.$$

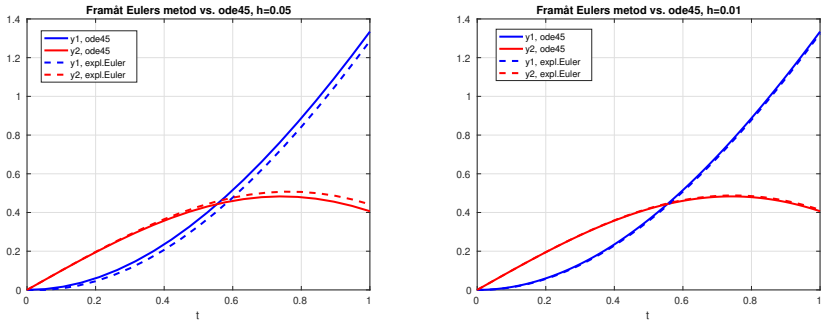


Fig. 6 *Framåt-Eulers metod versus ode45 för lösning av system* (33) *a) med* $h = 0.05$*; b) med* $h = 0.01$.